3-2008

Peer Review In An Undergraduate Biology Curriculum: Effects On Students' Scientific Reasoning, Writing and Attitudes

Briana Eileen Timmerman
University of South Carolina - Columbia, timmerman@sc.edu

Publication Info
2008, pages 1-235.

Science and Mathematics Education Centre

Peer Review In An Undergraduate Biology Curriculum: Effects On Students' Scientific Reasoning, Writing and Attitudes

Briana Eileen Timmerman

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology

March 2008

DECLARATION

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature:  .

Date: 24 February 2008

ACKNOWLEDGEMENTS

I would like to express my thanks to all the people without whose assistance, I could not have completed this work. Dr. David Treagust has been invaluable as my supervisor. I am also grateful to Dr. Barry Fraser as the chair of my committee and the Director of the Science and Math Education Centre and Dr. Robert Johnson at the University of South Carolina as my external committee member.

At the University of South Carolina, I am especially grateful to Sue Carstensen, Dr. Laurel Hester, Dr. Michelle Vieyra and Dr. Kirk Stowe whose support and willingness made peer review a reality in the Introductory Biology courses for so many semesters despite server shutdowns, time warps and other anomalies. They all made invaluable suggestions and adjustments to the rubric, the peer review process in CPR and otherwise worked tirelessly to improve and refine the educational value of the experience for the students. I would also like to express my appreciation for Dr. Sally Woodin's catalytic encouragement and support, both personally and professionally as well as other members of the Biology Faculty who contributed their time and thoughts to the development of the Rubric. John Payne provided a great deal of assistance in the generalizability analysis for the rubric, Katie Dahlke and Larry Powell were powerhouses of help for the peer review survey and Shobana Kubendran helped a great deal with demographics and database management. Denise Strickland requires special mention as her intellectual support and meticulous nature improved multiple portions of the research and her friendship and wit often made the work that much more enjoyable. Last, but most certainly not least, the multitude of Biology department graduate teaching assistants without whom there would be no biology labs, inquiry-based curriculum or lab reports. Graduate teaching assistants deserve thanks and applause for the thousands of students they tirelessly educate and mentor through the biology courses. Many also volunteered to review and test the Universal Rubric and provided much valuable feedback and I thank the seventeen in particular who participated in the Rubric reliability study. This work has only been possible with the efforts, help and support of many people.

This research was partially supported by National Science Foundation Award 0410992 with myself as Principle Investigator as well as by resources provided by the University of South Carolina, Department of Biological Sciences.

ABSTRACT

Scientific reasoning and writing skills are ubiquitous processes in science and therefore common goals of science curricula, particularly in higher education. Providing the individualized feedback necessary for the development of these skills is often costly in terms of faculty time, particularly in large science courses common at research universities. Past educational research literature suggests that the use of peer review may accelerate students' scientific reasoning skills without a concurrent demand on faculty time per student. Peer review contains many elements of effective pedagogy such as peer-peer collaboration, repeated practice at evaluation and critical thinking, formative feedback, multiple contrasting examples, and extensive writing. All of these pedagogies may contribute to improvement in students' scientific reasoning.

The effect of peer review on scientific reasoning was assessed using three major data sources: student performance on written lab reports, student performance on an objective *Scientific Reasoning Test* (Lawson, 1978) and student perceptions of the process of peer review in the scientific community as well as the classroom. In addition, the need to measure student performance across multiple science classes resulted in the development of a Universal Rubric for Laboratory Reports. The reliability of this instrument and its effect on the grading consistency of graduate teaching assistants were also tested. A spplication of the Universal Rubric to student laboratory reports across multiple biology classes revealed that the Rubric is further useful as a programmatic assessment tool. The Rubric highlighted curricular gaps and strengths as well as measuring student achievement over time.

This study demonstrated that even university freshman were effective and consistent peer reviewers and produced feedback that resulted in meaningful improvement in their science writing. Use of peer review accelerated the development of students' scientific reasoning abilities as measured both by laboratory reports (n = 142) and by the *Scientific Reasoning Test* (n= 389 biology majors) and this effect was stronger than the impact of several years of university coursework. The structure of the peer review process and the structure of the assignments used to generate the science laboratory reports had notable influence on student performance however. Improvements in laboratory reports were greatest

when the peer review process emphasized the generation of concrete and evaluative written feedback and when assignments explicitly incorporated the rubric criteria. The rubric was found to be reliable in the hands of graduate student teaching assistants (using generalizability analysis, $g = 0.85$) regardless of biological course content (three biology courses, total n = 142 student papers). Reliability increased as the number of criteria incorporated into the assignment increased. Consistent use of Universal Rubric criteria in undergraduate courses taught by graduate teaching assistants produced laboratory report scores with reliability values similar to those reported for other published rubrics and well above the reliabilities reported for professional peer review.

Lastly, students were overwhelmingly positive about peer review (83% average positive response, n = 1,026) reporting that it improved their writing, editing, researching and critical thinking skills. Interestingly, students reported that the act of *giving* feedback was equally useful to receiving feedback. Students connected the use of peer review in the classroom to its role in the scientific community and characterized peer review as a valuable skill they wished to acquire in their development as scientists.

Peer review is thus an effective pedagogical strategy for improving student scientific reasoning skills. Specific recommendations for classroom implementation and use of the Universal Rubric are provided. Use of laboratory reports for assessing student scientific reasoning and application of the Universal Rubric across multiple courses, especially for programmatic assessment, is also recommended.

DEDICATION

This thesis is dedicated to my husband, Dr. Henry Philip Crotwell, whose support and encouragement made finishing the thesis a reality instead of a hope. It is also dedicated to my daughters, Morgan Aileen and Charlotte Mayt, who are the sunshine in my days and often without knowing it, have kept me going with their smiles and laughter. May you never let your dreams escape you.

TABLE OF CONTENTS

	page
Cover Page	
Declaration……..	*ii*
Acknowledgements……	*iii*
Abstract……..	*iv*
Dedication……..	*vi*
List of Tables……	*x*
List of Figures……..	*xii*
Chapter 1: Introduction……..	1
Overview……..	1
Rationale……	1
Background…….	3
Why peer review was selected as a pedagogical strategy……..	6
Problem Statement……..	9
Research Questions……..	10
Limitations……	12
Significance……...	12
Summary……..	13
Chapter 2: Literature Review…….	15
Overview……...	15
What is scientific reasoning?……	17
How can educators facilitate the development of students' scientific reasoning abilities?…….	23
What is peer review?…….	30
Why is peer review likely to improve scientific reasoning?…….	30
How do we measure students' scientific reasoning abilities?.....	35
How has the literature informed this study?.....	50

page

Summary…….. 53

Chapter 3: Methodology……..... 54

Overview……. 54

Research design…… 54

Research studies……... 55

Study context …… 57

Data sources and instruments…….. 64

Ethics compliance…….. 77

Limitations of the study…….. 78

Summary……. 79

Chapter 4: Results from Achievement Data…….. 81

Overview……... 81

Study 1: Consistency and effectiveness of undergraduate peer reviewers… 83

Study 2: Reliability of the Universal Rubric for Laboratory Reports…….. 89

Study 4: Student achievement of scientific reasoning skills in laboratory reports (cross-sectional sample)…… 102

Study 5: Student achievement of scientific reasoning skills in laboratory reports (longitudinal sample)…… 106

Study 3: Reliability of the *Scientific Reasoning Test* in this undergraduate biology population……. 109

Study 7: Relationship between *Scientific Reasoning Test* scores and peer review experience……. 111

Study 6: Reliability of scores given by graduate teaching assistants in natural grading conditions…….. 117

Study 8: Graduate teaching assistants' perceptions of the Universal Rubric 125

Summary of achievement results……. 127

Chapter 5: Results of the Survey of Student Perceptions…… 128

Overview……... 128

page

Study 9: Undergraduate perceptions of peer review in the classroom…….... 131

Study 10: Undergraduate perceptions of the role of peer review in the scientific community…….. 145

Summary of students' perceptions of peer review……. 150

Chapter 6: Conclusions and Recommendations……. 151

Summary of the study context and problem statement…….. 151

Results of literature review and significance of the study……. 151

Summary of the components of the study……. 152

Summary of results of each study and discussion……. 156

Summary of conclusions…… 166

Study limitations and recommendations…… 168

Literature Cited…….. 171

Appendix 1. Department of Biological Science Curriculum Goals 185

Appendix 2. Handout given to students to encourage useful feedback 186

Appendix 3. Universal Rubric for Laboratory Reports 188

Appendix 4. Scoring Guide for the Universal Rubric given to Trained Raters 198

Appendix 5. Criteria list and instructions given to Natural Raters 215

Appendix 6. *Test of Scientific Reasoning* 216

Appendix 7. Student Peer Review Survey Fall 2006 226

Appendix 8. Student Peer Review Survey Spring 2007 228

Appendix 9. Single rater criteria reliabilities using trained raters 235

List of Tables page

No Tables in Chapter 1

Table 2.1. *Criteria Used in Professional Peer Review.* 38

Table 2.2. *Common Themes in Published Criteria for Science Writing or Scientific Reasoning.* 43-45

Table 2.3. *Published Reliability Scores for the* Scientific Reasoning Test. 49

Table 3.1. *Research Studies and Questions* 56

Table 3.2. *Examples From Handout Provided to Students to Encourage Them to be Effective Reviewers.* 61

Table 3.3. *Universal Rubric Criteria Codes and Definitions.* 66

Table 3.4. *Example of a Universal Rubric Criterion (Hypotheses: Testable and Consider Alternative) and Corresponding Performance Levels.* 67

Table 3.5. *Descriptions of Assignments Used to Generate Student Papers for Rubric Reliability Study* 68

Table 3.6. *Gender and Experience Levels of Graduate Student Raters* 70

Table 3.7. *Sample Sizes and Response Rates for Student Peer Review Surveys.* 76

Table 4.1. *Student Performance on Draft and Final Lab Reports and Changes Made as a Result of Peer Feedback* 88

Table 4.2. *Reliability of Individual Universal Rubric Criteria Using Generalizability Analysis (*g*)* 93

Table 4.3. *Reliability of Professional Peer Review and Relevant Rubrics for Writing.* 97

Table 4.4. *Inclusion of Rubric Criteria in Course Assignments* 99

Table 4.5. *Correspondence Between the Inclusion of Criteria in an Assignment and Criterion Reliability* 100

Table 4.6. *Longitudinal Performance of Individual Students Using Laboratory Report Total Scores.* 108

Table 4.7. *Distribution of Biology Majors' Prior Peer Review Experiences as a Function of Course Enrollment* 111

List of Tables page

Table 4.8. *Distribution of Students' Prior Peer Review Experience Once Students Who Repeated Courses are Removed.* 112

Table 4.9. *ANOVA Results When* Scientific Reasoning Test *Scores are Sorted by Cumulative Collegiate Credit Hours* 114

Table 4.10. *ANOVA Results When Transfer Credits are Excluded and* Scientific Reasoning Test *Scores are Sorted by University of South Carolina Credit Hours* 116

Table 4.11. *ANOVA Results When* Scientific Reasoning Test *Scores are Sorted by the Number of Peer Review Experiences in Which a Student Has Engaged* 117

Table 4.12. *Effect of a Few Hours Training on the Reliability of Scores Given by Graduate Teaching Assistants* 119

Table 4.13. *Reliability Scores for Individual Criteria under Natural Conditions* 121

Table 4.14. *Variability of Scores Awarded by Trained vs. Natural Raters.* 124

Table 4.15. *Summary of Achievement Data Results* 127

Table 5.1. *Average Percentage of Students' Positive Responses Regarding the Impact of Peer Review Across Three Introductory Biology Courses.* 130

Table 5.2. *Top Three Reasons Why Students Believe They Were Asked to Use Peer Review in the Classroom.* 134

Table 5.3. *Top Three Reasons Why Students Believe Scientists Use Peer Review* 147

Table 5.4. *Comparison of Students' Perceptions of the Functions of Peer Review in the Classroom and in the Scientific Community* 148

Table 6.1. *Brief Summary of Data Sources and Methodological Details for Each Study.* 155

Table 6.2. *Summary of Research Findings from this Study.* 167

Table 6.3. *Summary of Recommendations for Implementing Peer Review.* 169

List of Figures

page

Figure 1.1. Past research in science education provides a context for this investigation into the effect of peer review on students' scientific reasoning abilities. 2

Figure 1.2. Overview of research and relationships among individual studies. 11*

Figure 2.1. A portion of Table 3 from Halonen et al. (2003) describing the performance levels for a criterion. 47

Figure 3.1. Example of a tiered pair of questions from the *Scientific Reasoning Test* (Lawson et al., 2000) with a biological context. 73

Figure 4.1. Effect of using peer feedback on the quality of students' final papers. 85

Figure 4.2. Relationship between three rater reliability and single rater reliability using data derived from this study. 91

Figure 4.3. Student performance across a cross-section of biology courses. 103

Figure 4.4. Average scores earned by laboratory reports across multiple courses from longitudinal sample 107

Figure 4.5. Relationship between academic maturity and students' *Scientific Reasoning Test* scores. 114

Figure 4.6. Relationship between students' scores on the *Scientific Reasoning Test* and time spent in the USC curriculum. 115

Figure 4.7. Relationship between students' scores on the *Scientific Reasoning Test* and the number of peer review experiences in which they have engaged. 116

Figure 4.8. Comparison of the reliability scores of Trained vs. Natural raters 122

Figure 4.9. Comparison of stringency of Natural vs. Trained raters 123

Figure 5.1. Student perceptions of the role and impact of peer review by Introductory Biology course and term (total n = 1026). 135

*Also provided again on page 154 for the reader's convenience.

CHAPTER 1
INTRODUCTION

Overview

Two major sources of motivation exist for peer review as a subject of investigation. Firstly, past research suggests that peer review would be an effective pedagogical tool for improving scientific reasoning. Secondly, science educators desire their students to have functional working knowledge of the major components of the scientific process and peer review is one of those practical competencies. The context of this research is described in problem and purpose statements and the explicit research questions are outlined. Both quantitative and qualitative data were used to triangulate between evidence found in students' written work, their performance on a two-tiered scientific reasoning test and their self-reported perceptions of the peer review process. The limitations of these data sources and approach and the significance of this line of research are discussed.

Rationale

This research focused on the impact of peer review on students' scientific reasoning skills in a college biology curriculum. As indicated above, it is likely to be an effective pedagogical strategy for improving research ability as well as a skill required of practicing scientists and therefore desirable in students. Faculty in higher education institutions in particular and educators in general are unlikely to invest in new pedagogical strategies however unless significant evidence exists that such innovations will produce notable gains in student performance. While much research has investigated the pedagogical effectiveness of various components of peer review such as peer-peer collaboration, writing to learn, development of scientific process skills, there appear to be few explicit studies of the impact of peer review of science writing on students' scientific reasoning abilities (Figure 1.1). Thus, this research was required to satisfy the need for evidence and insights as to some of the effects of peer review on student scientific development. Further, for those university science departments around the United States that have already implemented peer review, this research may identify mechanisms for increasing its beneficial effects on student performance or reducing its frustrations by highlighting the relative strengths and weaknesses of the different aspects of the process.

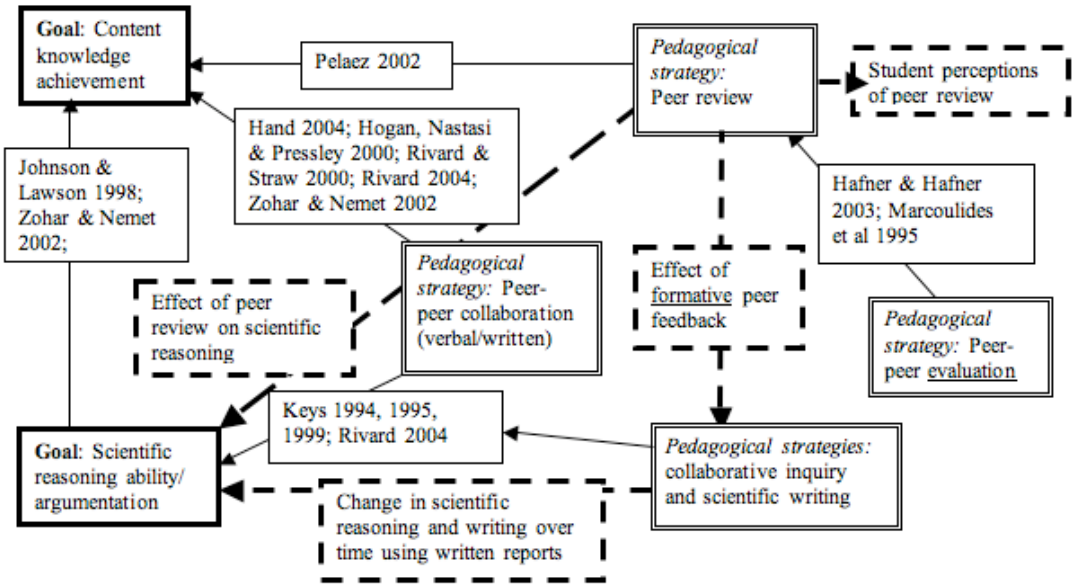


Figure 1.1. Past research in science education provides a context for this investigation into the effect of peer review on students' scientific reasoning abilities. Dashed lines indicate aspects and connections pursued by this research

For science faculty willing to consider incorporating new pedagogical strategies, peer review is a particularly attractive intervention because it is part of authentic scientific practice. Despite large volumes of literature on the benefits of inquiry-based teaching, such a pedagogical revolution has yet to broadly impact upon the pedagogy of many of higher education institutions, even in laboratory courses (Basey, Mendelow et al. 2000). This is likely due to the large time investment required to make such shifts, especially when higher education faculty and graduate teaching assistants are not generally provided with much pedagogical training or support for incorporating new methods into their teaching (Bianchini, Whitney, Breton, & Hilton-Brown, 2001; Carnegie Initiative on the Doctorate, 2001; Gaff, 2002; Luft, Kurdziel, Roehrig, Turner, & Wertsch, 2004; Volkmann & Zgagacz, 2004).

As peer review is a procedure that many science faculty already understand and have personally experienced, it does not impose the same time or cognitive demands inherent in other less intuitive pedagogical innovations. A contrasting example is the recent classroom innovation of *student response systems.* While a powerful pedagogical tool, student response systems require significant investments of time and effort by faculty to modify their teaching approach and materials. Adoption of student response systems has been slower despite massive financial

investments from book publishers. Specifically, student response systems are a commercial product which allow an instructor to pose questions in large lecture auditoriums and receive instantaneous feedback from students who answer using small hand-held wireless devices. They allow an unprecedented degree of interaction and feedback for both faculty and students even when class sizes are in the hundreds to thousands (Powell 2003). Compared to peer review, however, adoption of student response systems has been slow despite the investment of notable commercial resources such as bundling student response systems with textbooks, offering free trials and other publisher incentives.

Peer review has already been initiated in over 3800 courses at 900 institutions in the United States using the *Calibrated Peer Review (CPR)* website alone with a user-base of over 140,000 student accounts (Russel 2007). Other academic and commercial peer review websites exist with their own distinct user-bases. In contrast, the largest commercial student response system (Classroom Performance System – CPS) is in use by only slightly more than twice the number of students. This is surprising given that CPS is backed by Pearson Education and the formidable marketing and advertising divisions inherent in a global publishing company serving 100,000 million people (www.pearsoned.com/about/index.htm). In comparison, the fact that the user-base of the *Calibrated Peer Review* website has grown to nearly 50% the size of that of the Classroom Performance System through academic word-of-mouth with absolutely no commercial advertising indicates that science faculty appear to have an affinity for peer review as an instructional strategy. Thus, there is both a substantial audience interested in research on the effect of peer review as well as a large reservoir of science faculty who might adopt peer review if evidence of its effectiveness were available.

Background

Strengthening scientific reasoning skills improves content knowledge

One of underlying purposes of science education is the development of scientific reasoning skills (American Association for the Advancement of Science, 1993; Committee on Undergraduate Biology Education, 2003) both as a pre-requisite for a scientifically literate society, and as an end-goal in itself. Scientific reasoning skills also correlate with students' abilities to learn content knowledge. Past research

has shown that a focus on students' reasoning skills improves or is sometimes a pre-requisite to students' ability to learn content knowledge. Students classified as possessing greater formal reasoning ability showed much larger gains on a concept knowledge test especially on items dealing with more abstract biological concepts such as evolution or cellular level processes (Lawson, Alkhoury, Benford, Clark, & Falconer, 2000). Use of an inquiry-based approach also increased students' reasoning abilities over the course of the semester (Lawson et al., 2000). Johnson and Lawson (1998) compared end-of-term content achievement for students in inquiry vs. expository sections of a non-majors biology course and found that reasoning ability was a better predictor of performance than prior knowledge or coursework. Students in sections that used an inquiry-based approach also showed larger gains in reasoning ability. Zohar (2002) identified that explicit instruction in argumentation produced greater knowledge of genetics and resulted in students being able to transfer reasoning abilities to everyday situations. These prior studies provide an explicit connection between scientific reasoning skills and content knowledge gains and therefore motivate university science departments to improve students' scientific reasoning skills both for their intrinsic value and because it is also likely to improve their performance in future courses.

Peer-peer collaboration improves content knowledge and/or scientific reasoning

Pelaez (2002) required half of the content for her physiology course to be taught by students completing online research and writing assignments followed by anonymous peer review using the same software system employed in this research (*Calibrated Peer Review*). She compared student achievement in content knowledge for topics taught by peer reviewed online research projects with scores for topics taught by standard lecture followed by group work. Pelaez found significantly greater student achievement for topics that were taught by peer review compared to didactically taught topics ($p < 0.001$, paired t-test). These gains were realized by both top-scoring and low-scoring students in both multiple choice and essay-based assessments suggesting that meaningful learning was occurring. This is a striking and compelling finding that self-study and peer-peer interaction caused greater gains in content knowledge than those produced by a standard lecture format. Pelaez suggested the peer review process helped to identify hidden misconceptions that usually are not addressed by more didactic pedagogies. As identification of

inaccurate prior knowledge and confrontation of erroneous ideas are pre-requisites for conceptual change and meaningful learning (Posner, et al. 1982), peer review may be a particularly powerful pedagogical strategy.

Pelaez further suggests that such peer-peer collaboration may be of particular benefit to low-achieving students who traditionally have difficulty identifying such gaps or inaccuracies in their own knowledge or comprehension. She cautions that the difficulty of the peer-review task must be matched to the students' scientific reasoning abilities however. She further elaborates that the greater formative feedback provided by the peer-review system may also be of particular advantage to students who traditionally under-perform on assessments which require synthesis and critical thinking. As longer writing assignments are a more time-intensive means of assessing learning than multiple choice exams, students from under-resourced K-12 educational systems may have far less experience and therefore less opportunity to gain these evaluative and synthesis skills. Peer-review as a formative assessment helps to level this playing field by allowing students multiple opportunities to practice these skills as well as receive productive feedback thus negating the common pattern where performance on essay exams favours students from more privileged educational backgrounds. Pelaez's (2002) work therefore demonstrates that peer review is a powerful pedagogical strategy for improving students' content knowledge and strongly suggests that equally benefits their critical thinking and evaluative skills as well.

Focusing on the effect of peer-peer interaction on critical thinking, Hogan, Nastasi, and Pressley (2000) compared peer-led discussions vs. teacher-led discussions in Grade 8 science classes. They found that students asked three and a half times as many questions (18% vs. 5% of total verbal statements) and made twice as many metacognitive statements (18% vs. 9 %) in peer led group discussions compared to teacher led discussions (Hogan, Nastasi et al. 2000). In particular, peer led discussions contained twice the number of statements categorized as *justification* (25% vs. 13%) and *synthesis* (40 % vs. 21%) than did teacher led groups indicating that students in peer groups were engaged in critical thinking for larger percentage of the time (Hogan, Nastasi et al. 2000). Thus, peer-peer collaboration appears to encourage students to engage and practice reasoning skills much more frequently than do teacher-centred teaching methods.

Rivard (2004) found that peer-peer interactions improved all students' science knowledge, though the mode of improvement varied by student achievement level; verbal collaboration and exploration produced greater gains in low achievers while high achieving students benefited more from writing explanations. Peer-peer collaboration can also improve students' writing as well as content knowledge. Specifically, being forced to employ the criteria in peer evaluations has been shown to lead to greater understanding and implementation of the assignment criteria by students as writers at the college level (Bloxham and West 2004). Thus, peer-peer collaboration and the feedback provided by peers during such collaboration are useful pedagogical strategies that facilitate student learning.

Connection between writing and learning

There is a general recognition that writing is an important component of science and that the act of writing often improves or structures scientific reasoning (Florence & Yore, 2004; Yore, Hand, & Florence, 2004). Rivard and Straw (2000) tested the relative and combined effects of talking and/or writing on students' understanding of ecology and found that analytical writing helps to transform rudimentary ideas into coherent and structured knowledge. Hand, Hohenshell & Prain (2004) found a direct connection between the number of student writing experiences (especially when students were writing for an audience other than the instructor) and conceptual gains as well as retention of that knowledge eight weeks later. Furthermore, Keys (1994) found repeated collaborative writing assignments also could demonstrate an increase in scientific reasoning skills among Grade 9 science students. Thus, writing is a productive venue for peer review in a pedagogical sense, as well as being an authentic process of science.

Why peer review was selected as a pedagogical strategy

Components of peer review as a pedagogical strategy

Beyond its already demonstrated value as a pedagogical tool for improving content knowledge (e.g. Pelaez, 2002), past research suggests that peer review is likely to improve a student's critical thinking skills for the following reasons. Firstly, peer review provides exposure to multiple contrasting examples helping students to determine the salient criteria for a given task (Bransford, Franks et al. 1989). Secondly, peer review potentially provides relevant formative feedback. Formative

feedback has been shown to have significant benefits for improving student work and learning (Chinn & Hilgers, 2000; Ravitz, 2002; Topping, Smith, Swanson, & Elliot, 2000; Yorke, 2003). Thirdly, providing feedback to three separate peers requires students to quadruple the extent to which they practice critical thinking skills over the course of an assignment. Normally students would review and revise only their own paper, but with peer review, they must engage, evaluate and construct suggestions for three papers in addition to their own, thereby gaining four times the practice at this skill. Concerted and repeated practice over time is an important component of developing expertise (Ericsson and Charness 1994). Lastly, students' comments on end-of-term evaluations indicate that the process of peer review often stimulates the reflection and self-assessment that has been shown to lead to metacognition and meaningful understanding (Baird & White, 1996; Bloxham & West, 2004; Pope, 2005).

There has been a great deal of work on how 'writing to learn' improves both content and argumentation skills, as well as how peer collaboration improves reasoning ability and content knowledge. With the exception of Keys and colleagues' work (Keys, 1994, 1995, 1999b, 2000; Keys, Hand, Prain, & Collins, 1999), however, little has been done to explicitly address how scientific writing or peer review affects scientific reasoning (Figure 1.1). Further as, Keys's work focused on collaborative writing, not just collaborative inquiry, it therein departs from the model used in this study (collaborative inquiry and individual writing). Further, while research has been conducted on the reliability of peers as reviewers (Cho, Schunn, & Wilson, 2006; Hafner & Hafner, 2003; Marcoulides & Simkin, 1995), to this researcher's knowledge, little work has been published on the nature of undergraduates' *formative* feedback (except see Cho, Schunn et al. 2006) nor its effect on students' scientific writing and reasoning skills. Thus, investigation of the explicit connection between peer-review and improvements in students' scientific reasoning skills will connect previous research in new and fruitful ways.

Support for the underlying assumption that peers can be effective evaluators

Undergraduate peers have also been found to be reliable evaluators. When peer reviewers assessed oral presentations in a large-scale study of college biology students, the aggregated mean peer review total score (scores averaged over approximately 30 reviewers) was found to have nearly a 1:1 correspondence with the

instructor generated scores for all three years of the study (Hafner & Hafner, 2003). Individual reviewer scores were highly variable (generalizability analysis attributed only 25% of score variance between any two raters to actual differences in presentation quality), but mean student scores across all reviewers were found to be highly consistent and informative (Hafner & Hafner, 2003).

In a study more similar in structure to this research, 60 undergraduates were found to be reliable in their assessment of the quality of peers' written papers with 65% to 75% of the variability in score being attributed to real differences in quality among papers according to a generalizability analysis (Marcoulides and Simkin 1995). Cho, Schunn and Wilson (2006) also found that peer review using six referees produced a reliability of 0.78 and a high correlation with instructor scores ($r = 0.89$) when a rubric was employed. It should be noted that all the studies reported above measured the reliability of *numerical* scores (i.e. grades) assigned by peers rather than measuring the quality or impact of more subjective comments regarding writing substance and quality. Cho, Schunn and Charney (2006) investigated the helpfulness of peer comments and found that when undergraduates rated the helpfulness of the feedback they received, there was no distinction between the average value assigned to instructor comments versus peer comments ($p = 0.36$). Thus, peer can effectively apply criteria and provide useful evaluations of written work and peer feedback, especially when aggregated, can be comparable in quality to instructor feedback.

Rationale for peer review to be taught as a scientific skill

Given the demonstrated effectiveness of peer-peer collaboration, writing and peer assessment in improving students' knowledge and reasoning abilities, the next logical step is to combine those functions into a single pedagogical strategy of peer review. Beyond its inherent value as a real-world skill, there are multiple reasons why peer review is likely an effective pedagogical strategy. Advocated best practices for science teaching in general have converged into categories that correspond to the fundamental components of peer review: active engagement (Linn 1997), collaborative learning (Cabrera, Colbeck et al. 2001), formative feedback (Yorke 2003), reflection (Baird and White 1996), and a focus on incorporating assessment as an integral part of the curriculum (Linn 1997). A review of the literature on self, peer and co-assessment of student work indicates that such a focus develops the

competencies needed by students to engage in professional practice (Sluijsmans, Dochy et al. 1999).

Additionally, while both practicing scientists as well as science educators explicitly desire that students develop their scientific reasoning skills as a fundamental goal of science laboratory experiences (Goodman et al., 2007; Pelaez & Gonzalez, 2002) many science faculty also desire that students become aware of the integral nature of writing to the process of scientific thinking (Yore et al., 2004). Science faculty also desire their students to understand the fundamental role peer review plays in maintaining the integrity of science (Abd-El-Khalick and Lederman 2000). Indeed, Ford (2008) provides a convincing argument that the scientific community and peer review is the source of authority in science and that an explicit focus on the role of peer review is crucial to effective construction of content knowledge as well as development of students' scientific reasoning. He specifically focuses on the need to teach students to *critique* ideas and claims, not just *construct* them and that a realistic understanding of how scientific knowledge is generated is necessary for effective learning (Ford, 2008). Additionally, writing and peer review are real world skills desirable in scientifically literate people as well as future scientists. While engagement in authentic practices is important to the development of scientific inquiry skills (O'Neill & Polman, 2004) it does not necessarily lead to understanding of scientific process (Schwartz, Lederman et al. 2004) if its purposes as both a pedagogical tool and desirable real world skill are not made explicit to students. Thus, the value of peer review will be bolstered if instruction explicitly addresses the rationale for including peer review in the curriculum. Students' perceptions of the role of peer review in the classroom and as a real-world skill were therefore also addressed as part of this research.

Problem statement

Curriculum goals developed by the Department of Biological Sciences at the University of South Carolina and similar large, research-based science departments focus on developing functional scientific competencies in their students. Foremost among these goals is developing students' scientific reasoning skills and writing skills. Defined components of scientific reasoning include: identifying assumptions, creating and evaluating hypotheses, designing relevant experiments, analysing data,

evaluating results, assessing the validity of conclusions, identifying gaps in knowledge, learning from mistakes and deciding on steps to be taken in the future (see Appendix 1 for full list of Curriculum Goals). In short, two major curriculum goals are for students to be able to engage in effective scientific reasoning and communication of findings. Departmental review of the biology curriculum in 2003 identified two major challenges to achieving these goals. Firstly, the majority of undergraduate biology laboratory experiences revealed a lack of emphasis on scientific reasoning skills. Secondly, no systematic means of evaluating the effectiveness of the curriculum existed.

Consequently, curriculum reform efforts ensued to provide more opportunities for students to engage in open-ended research investigations in the introductory biology and sophomore-level courses for majors. Effective science writing was also a desired competency as well as a rich data source for assessing student scientific reasoning abilities. The process of peer review appeared to address both these challenges simultaneously. Peer review is both an authentic scientific competency in and of itself as well as being a pedagogical tool that can engage students in significant opportunities to practice scientific reasoning, evaluation and writing skills. Further, peer review as a pedagogical tool increases the level of formative feedback provided to students thereby increasing opportunities for learning, without placing further time demands on faculty or graduate teaching assistants. Peer review was thus selected as one pedagogical innovation to address the problem of improving students' scientific reasoning skills as well as training them in a professional competency. The focus it placed on student science writing also provided a rich data source (draft and final versions of papers) facilitating a solution for assessing students' scientific reasoning abilities in a meaningful way.

It quickly became evident however, that a single course was insufficient for the development of these skills and that a means of assessing student performance across multiple courses was needed. Thus, this research also reports on the development and testing of a Universal Rubric for Laboratory Reports as a means of measuring the change in students' scientific reasoning abilities over time.

Research Questions

This research focused on the following broad research questions. Can undergraduate students consistently and effectively engage in peer review? Is the

Universal Rubric for Laboratory Reports a reliable measure of student scientific reasoning abilities? How do students' scientific reasoning abilities in their laboratory reports change with course topic, assignment details and over time? Do students' scientific reasoning abilities improve with additional peer review experiences? Lastly, do students perceive peer review as a worthwhile educational activity? These broad questions were divided into the ten studies outlined in Figure 1.2 and summarized below. The relationship between the ten studies and each of the related research questions is presented in table 3.1.

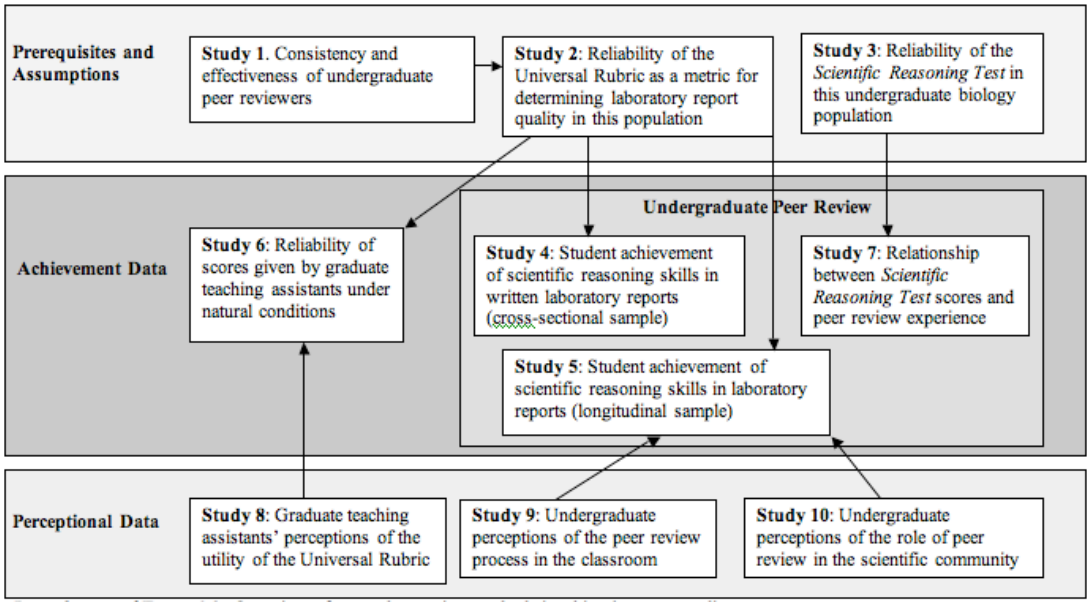


Figure 1.2. Overview of research and relationships among individual studies.

Firstly, this research established the required foundational condition that students are capable of effectively and consistently engaging in peer review (Study 1). Next, the reliability of the Universal Rubric was established (Study 2) and it was used to assess changes in student scientific reasoning abilities over time in a cross-sectional sample (Study 4) and a longitudinal sample (Study 5). Additionally, the reliability of scores generated by graduate teaching assistants under natural grading situations was assessed (Study 6) and graduate student opinions regarding the utility of the Universal Rubric were collected (Study 8). An external, objective measure of student scientific reasoning ability was also employed using the *Test of Scientific Reasoning* (Lawson, 1978; Lawson et al., 2000). Its reliability in this population was established (Study 3) and the relationship between *Scientific Reasoning Test* score

and students' peer review experiences was investigated (Study 7). Lastly, undergraduates' perceptions and understandings of the purpose, utility and impact of peer review in the classroom (Study 9) and the role of peer review in the scientific community (Study 10) were investigated.

Limitations

Students' written laboratory reports, while a rich source of data, inherently miss some aspects of reasoning which would be more clearly seen in students' small group discussions, by direct questioning or observation of students in the laboratory. Further, using only written work as the primary data source definitely biases the results towards students who are able to better articulate their scientific reasoning onto paper. There likely are students within this sample who have scientific reasoning skills and gains that were not evident because those students have more difficulty expressing themselves in writing than verbally or by action and decision. This research also focuses on a product (written report) that is often the result of scientific reasoning. Focusing on the end-product rather than the process itself means that some nuances of how reasoning develops many not be noticed. Use of an authentic outcome such as written reports is an effective compromise, however, given the realities of limited resources and investigator time for such a large population of students. This research approach allows the assessment of the effect of peer review; further research may then illuminate the real-time mechanisms by which peer review impacts the development of reasoning skills.

Significance

Past research has investigated the effectiveness of various *components* of peer review such as peer-peer collaboration, writing to learn and formative feedback on the development of students' scientific skills (Fig. 1.1) suggesting that peer review would be a powerful pedagogical strategy as it combines many of these elements. Peer review is also a skill required of practicing scientists and therefore desirable to develop in students as an authentic competency. As little direct research exists on the effect of peer review on students' scientific reasoning abilities, however, these studies will contribute to valuable insight to our understanding of students' scientific development. Besides contributing new knowledge to the field of science education,

this research may have practical implications for faculty who have not adopted peer review, by providing sufficient evidence to encourage them to incorporate it into their own classes. Further, for those university science departments around the United States that have already implemented peer review, this research may identify mechanisms for increasing its beneficial effects on student performance or reducing its frustrations by highlighting the relative strengths and weaknesses of the different aspects of the process.

Lastly, development of a rubric of core criteria for written laboratory reports that emphasises scientific reasoning is of interest to a wide variety of science faculty and to date, no generalised rubric for written laboratory reports appears to be available in the literature. Assessment of the effect of peer review on students' writing is of interest to both science and writing faculty and language and literacy faculty interested in argumentation.

Summary

Writing and critical thinking are ubiquitous practices in science and goals of science education. A rich and varied literature also exists on both the benefits of peer collaboration and formative feedback as well as on the benefits of analytical writing to learning and critical thinking. Peer review emphasises learning by writing and provides multiple additional opportunities for students to practice scientific reasoning and evaluative skills. It also increases the level of formative feedback provided to students three-fold without a concurrent increasing demand on instructor time. Peer review may therefore accelerate the development of students' scientific reasoning skills.

This thesis focuses on assessing the effect of peer review on students' scientific reasoning abilities. Peer review is hypothesized to be a mechanism for stimulating the discourse and reflection necessary for development of scientific reasoning skills. Predicted changes in scientific thinking skills can be measured via an objective quantitative test as well as qualitative analysis of students' written laboratory reports and peer reviews. Comparisons of the correlation between scientific reasoning and generalized undergraduate academic experience (number of credit hours earned) compared to the number of peer review experience allows the effects due to the peer review process to be distinguished from the effect of

increasing academic maturity. Additionally, investigation of scientific reasoning evidenced in laboratory reports using both cross-sectional and longitudinal samples allowed rich and direct measurement of student achievement of curricular goals. Lastly, a survey investigating student perceptions of the peer review process would enable the impact of peer review to be viewed through the students' eyes. The context for the study is undergraduate biology laboratory courses at the freshman, sophomore and upper division levels at a large (25,000 students) state university in the United States.

Peer review thus provides an interesting link between the areas of scientific writing and reasoning and investigation into its effect on students' scientific reasoning and writing skills will provide useful data to faculty and administrators in higher education concerned with student achievement as well as programmatic assessment and demands on faculty time.

CHAPTER 2
LITERATURE REVIEW

Overview

This chapter is structured around the relevant literature supporting and defining the concepts of scientific reasoning, science writing, and peer review as well as discussing what tools are currently available for measuring scientific reasoning in written form and what tools must still be developed. Scientific reasoning in particular is a concept that is differently defined in various subfields. Here it refers to the mental processes necessary to design, implement and interpret scientific research. Therefore, the initial portion of this chapter is spent more fully discussing what those mental processes might be and what we know about how those mental abilities develop in students.

Scientific reasoning can be demonstrated in a wide variety of contexts such as observation of a person's actions, recording verbal discussion, probing by interviews or analysing scientific outputs, such as written reports. While other contexts are briefly discussed, writing is the logical pragmatic data source to investigate scientific reasoning in higher education given the relatively well-developed generalized writing ability of students at that age (compared to K-12 abilities), the ubiquity of written laboratory reports in the curriculum, the authenticity of writing as a means of communicating scientific endeavors and the availability of appropriate instruction and mentoring from other science writers (graduate students and faculty teaching science courses). A discussion of what distinguishes science writing from other writing and how it relates to scientific reasoning is therefore included.

The process of peer review, as it is a familiar process to scholars, is only briefly described. Instead time is spent investigating the potential effectiveness of peer review as a pedagogical strategy. It should be noted that peer review is doubly valued in that it is a desirable skill for students to learn simply because it is an authentic scholarly activity as well as being a pedagogically powerful tool.

Having now identified the learning outcome of interest (scientific reasoning), the instructional intervention (peer review) and the primary data source (science writing), it is clear that a tool is needed with which to measure the resulting student achievement. A review of the criteria used in professional peer review as well as criteria and measurement tools available in the published literature follow. These

resources were used to compile a list of consensus criteria for effective scientific reasoning in written formats. It became clear that no appropriate published rubric was available. Criteria from the published literature were then combined with input from departmental faculty, graduate students and external educational researchers over the course of 18 months. The result was Universal Rubric for Laboratory Reports. Out of consideration for the reader, the various versions, revisions and discussions required to produce the rubric were not detailed here. The relevant literature components were mapped onto the final Universal Rubric criteria so that the reader may see whether support for each criterion derives from the scientific community and/or the research literature. The reader should note that all criteria included in the final version of the rubric were also reviewed and received support from the faculty of the Biological Sciences Department at the University of South Carolina.

Regardless of its reliability and validity, however, the rubric provides only a single data perspecitve. A construct as complex as scientific reasoning requires data to be triangulated across multiple perspectives to increase confidence in the validity and generalizability of conclusions (Mathison, 1988). The potential use of a different methodology (published pencil and paper test of scientific reasoning) as a means of sampling a greater proportion of the student population was therefore investigated. This more distal measurement could determine if the effect of peer review on scientific reasoning was detectable beyond the confines of student written reports.

Lastly, research literature demonstrated that information on student perceptions of the utility and impact of peer review would be worthwhile as student perceptions often have a strong impact on and instructor's willingness to implement instructional innovations. Previous research on students' perceptions of peer review was sparse and indicated a need for a large scale, quantitative survey. There also appeared to be a gap in the research literature concerning students' perceptions of the role of peer review in the scientific community.

The chapter then concludes with a discussion of additional insights gleaned from the literature about how to best implement peer review in the classroom and an overall summary.

What is scientific reasoning?

Historical perspective

The field of study concerned with scientific reasoning can be divided along one's belief in the extent to which content knowledge and problem context affect reasoning strategy and ability. At one end of the spectrum are studies that focus on the development of domain specific content knowledge and the development of conceptual knowledge and procedures within a field. The other end of the spectrum considers scientific reasoning to exist across all scientific domains and focuses on scientific problem solving e.g. "hypothesis generation, experimental design and evidence evaluation" (Zimmerman, 2000, p. 101) independent of content knowledge. Zimmerman (2000) describes most of the domain specific research as focusing on identification of naïve theories (e.g. misconceptions) and the process of conceptual change in specific topic areas. This researcher concurs that challenging students' alternative conceptions or misconceptions and development of content knowledge is more accurately described as learning rather than reasoning. Similarly, at the other end of the spectrum, early attempts to quantify scientific reasoning separated a person's content knowledge from the strategies that they might use to solve problems (e.g. Ward & Overton, 1990) found little relationship between a person's education or occupation and their performance on such tasks. This type of "knowledge-lean" or "domain-general" knowledge was believed to be distinct from a person's knowledge of particular subject matters, but subsequent work in the field indicated that reasoning abilities were clearly affected by the context in which the reasoning task was set, even if no specific knowledge was required (see review in Zimmerman 2000). Namely, when presented with a logic problem set in an everyday context (e.g. students being punished for rule-breaking at school), students' correct answers were greater than if presented with the exact same logical structure in an "if a, then b" format (Ward & Overton, 1990).

As the ability to conduct science is of interest here, scientific reasoning here is viewed as process mid-way between those two extremes. The strategies and abilities with which this study was concerned transcend specific scientific context (hypothesis generation, analysis of evidence, etc.), but this researcher firmly acknowledges that those processes are most realistically measured in contexts with which the subject has at least some familiarity (e.g. everyday contexts or course material).

Characteristics of scientific reasoning in experts

A first step in a discussion of the development of scientific reasoning is to identify what is meant by the term. Kuhn (1989) defines scientific thinking as the ability to consciously use and coordinate theory and evidence.

> The scientist (a) is able to consciously articulate a theory that he or she accepts, (b) knows what evidence does and could support it and what evidence does or would contradict it, and (c) is able to justify why the coordination of available theories and evidence has led him or her to accept that theory and reject others purporting to account for the same phenomena. (Kuhn 1989, p. 674).

Zimmerman (2000, p. 104) expands the definition to encompass the action of problem solving in addition to the justification of conclusions and defines scientific reasoning as "problem solving *strategies* [emphasis original] that are involved in the discovery and modification of theories about categorical or causal relationships."

Within domain general models of scientific reasoning, there exists another relevant distinction however. Early work focused on *knowledge lean* or context independent reasoning problems that usually took only minutes to solve (Zimmerman 2000). Such problems are much more a test of formal abstract reasoning skills rather than scientific reasoning skills however. Content knowledge, while not the goal of scientific reasoning here, is required in order to frame and inform ones decisions about which hypotheses are likely to be fruitful and what techniques are available for experimental design. Later work on scientific reasoning shifted to more *knowledge rich* tasks that required multiple steps over longer time periods. In relevant work, Klahr and Dunbar (1988) identify three major interrelated conceptual spaces in scientific reasoning in their Scientific Discovery as Dual Search model: 1) hypothesis generation, 2) experimental design and testing of hypotheses and 3) evidence evaluation. They note specifically that scientific reasoning is not a linear process, but a recursive coordination and integration of these processes.

Much of the cognitive psychology literature on scientific reasoning focuses on expertise as it is from observing experts that we derive the qualities and traits that comprise the definition of scientific reasoning. Dunbar extends the work begun with Klahr, but makes the logical but radical suggestion that studying people of various

backgrounds under controlled experimental conditions on artificial tasks in a psychology laboratory fails to capture the aspects of scientific reasoning that produce new scientific discoveries. He suggests specifically that the failure to observe scientists as they actually work creates two gaps in our understanding: 1) actual scientific problem solving may differ from solving arbitrary psychology tasks, and 2) the large contribution made by the social collaboration among colleagues is excluded (Dunbar 2000). When he observed scientific experts in their natural environment (here defined as well-funded molecular biologists heading up their own research laboratories), Dunbar (1997; 2000) generated several surprising findings: 1) that practiced scientists attend to *un*expected findings as the major source of new hypotheses and experiments, and 2) that they engage in distributed (group) reasoning (e.g. group discussions in laboratory meetings) to overcome challenges in their research. In particular, Dunbar concludes that group reasoning surmounts the difficulties that individuals have generating explanations for unexpected results.

> This pattern of challenging inductions was ubiquitous across all labs…. Individual subjects have great difficulties in generating alternate inductions from data, and also have great difficulties in either limiting or expanding inductions. Distributed reasoning helps circumvent these difficulties. When distributed reasoning occurs, the group quickly focuses on the reasoning that has occurred and the other members of the laboratory will generate different representations. These new representations will make it possible for members of the lab to propose alternate inductions, deductions and causal explanations. Thus, distributed reasoning provides new premises and models that a particular individual might not be able to generate when reasoning alone. (Dunbar 1997, p. 13)

Not surprising to those who have attended laboratory meetings, however, none of the members of the group recalled or could identify the contributions of various members to the solution once the challenge had passed.

> Once a new concept is generated the cognitive scaffolding is thrown away and scientists cannot reconstruct the cognitive steps that went into the

> discovery. Because of this scientists and historians reconstruct their creative moments, often from their lab books. Unfortunately many of the key cognitive steps made in a discovery do not end up in the lab books. Thus, many of these reconstructions are based on partial information and, as a result, myths surrounding the creative process develop. (Dunbar 1997, p. 16)

This work provides three important insights. Firstly, the focus on collaboration and group work in science classrooms is further justified as a skill required of practicing scientists (Dunbar 2000) in addition to being an effective pedagogical tool. Dunbar's work also makes clear that focusing on unexpected results and/or conflicting data is an intentional action or skill of practicing scientists (1997). University instruction should therefore include addressing conflicting data and consideration of alternate explanations as an important part of the curriculum and such an emphasis should be included in whatever criteria are used to assess student performance. Thirdly, this is a stark reminder that even the most comprehensive performance assessment cannot capture all the skills required to be a practicing scientist. Student performance in science writing is a robust measure of scientific reasoning skills, but does not capture the process of science in its entirety.

Another means of characterising scientific reasoning in experts is to compare the strategies of experts versus novices when solving problems. Common characteristics of novices appear to be the obvious lesser content knowledge, as well as a tendency to focus on the surface qualities of a problem, rather than work with the underlying principles (Dhillon 1998; Schunn and Anderson 1999). Experts tend to frame problems using the abstract underlying principles and solve problems by dividing them into functional sub-problems (Schraagen 1990). This tendency of novices to miss the underlying structure of a problem is a familiar phenomenon for any instructor who has successfully led students through one problem solving exercise only to have them completely stymied when presented with the same type of problem set in a different context. Novices also often tend to be unable to correctly represent problems in diagram form (Dhillon 1998; Schunn and Anderson 1999) (likely to do the same lack of understanding of underlying principles described above), use many, short, less informative and potentially random means of attempting

to solve the problem as well as be less aware or capable of the need to control variables (Schunn and Anderson 1999).

Further, there is a notable distinction in how novices and experts approach a problem in that experts spend more time and effort identifying the underlying paradigm and framing the question than do novices. For example, four groups of subjects ranging in their level of expertise and content knowledge were presented with a real world problem. Undergraduate experimental psychology majors (novices), experimental psychology graduate students (intermediates), PhDs in areas outside of sensory psychology (design experts) and PhD in sensory physiology (domain experts) were presented with the problem that the taste of Coke-Cola no longer appears to be preferred by much of the Dutch public and asked to design research programs which would suggest how the product should be improved (Schraagen 1990). The presence of two groups of PhDs was an attempt to determine the extent to which problem solving is domain specific. Namely, PhDs from other areas have plenty of expertise in conducting scientific research, but using content knowledge that is less relevant here. Subjects were asked to 'talk-aloud' describing their thinking as rationale as they worked through the problem. The domain experts spent nearly fives times as long in framing the question as did beginners (10.5 vs. 2.2 statements, $p = 0.02$) and nearly twice as long as did the design experts (5.7 statements). The research programs produced by the subjects varied in that the PhDs all produced much more structured goals than did the novices, out-of-area PhDs and graduate students all used mental simulation a great deal and only the PhDs problem conception schema contained abstract principles (Schraagen 1990). In-area PhDs broke the problem down into sub-problems and designed a means of investigating each sub-problem. Out-of-area PhDs also framed the problem in abstract terms, but had to resort to mental simulation, working through the results that would be generated by each experimental approach and then checking the outcome against the initial goal to determine if it was a fruitful approach (Schraagen 1990).

Thus, the identification of an appropriate paradigm and subsequent approach to a research problem appears to be the most challenging portion of scientific reasoning, particularly because novices are often completely unaware of the need to identify the underlying principles before designing an experimental approach. When research is set into more scholarly settings, this issue of needing to identify the underlying paradigm would translate itself into a need to understand the context or

knowledge landscape of a field and propose research that would fill interesting gaps in that knowledge in a fruitful way.

Summing across these works, scientific reasoning thus appears to have several layers. The first layer of reasoning is the ability to take static data and make logical conclusions or to design an experiment that would test a given hypothesis. This type of reasoning can be measured by paper and pencil tests which present students with scenarios for which they are asked to select appropriate methodologies or to interpret the outcomes of experiments already run (e.g. Scientific Reasoning Test, Lawson 1978). The second layer, the development of the more complex skill of identifying underlying principles and framing a scientifically interesting question is likely more challenging and requires a more comprehensive methodology for its assessment. Appropriate methodologies might include direct observations of students while engaged in research, or analysis of students' written reports on their research project.

The last layer, which is missing from these scholarly attempts to investigate how science is conducted is the ability to be conversant in ones' scientific field and recognize when conceptual frameworks are incompatible with the existing evidence and to propose a new paradigms (Kuhn 1970). While superficially similar to the ability to recognize that a given set of data may or may not support a given conclusion and suggesting or selecting alternative tests (which is common in paper and pencil tests), such constrained scenarios miss the more fully developed process that occurs when investigators take a step back in their consideration of the problem and re-examine not just the localized concept, but the conceptual framework from which it is derived. This pinnacle of scientific reasoning produces new thoughts or insight that lead to scientific discovery. While true scientific discovery is difficult to achieve within even the higher education classroom, it should be included in the definition of scientific reasoning, or else the definition cannot encompass the best examples of scientists at work.

Therefore, for the purposes of this study, scientific reasoning is defined as: the ability to generate, manipulate, evaluate and reconcile data within conceptual frameworks. Additionally, scientific reasoning includes the ability to note disparities among data or between data and theoretical frameworks and test and revise to those conceptual frameworks in a continuous attempt to

generate an internally consistent understanding of phenomena. The challenge now arises as to how to facilitate the development of scientific reasoning in our students as well as measure the effectiveness of such a curriculum.

How can educators facilitate the development of students' scientific reasoning abilities?

What is known about the trajectory of how scientific reasoning skills develop?

Zimmerman (2000) provides a comprehensive review of literature regarding scientific reasoning and the following highlights emerge from her work.

> Children's performance (third to sixth graders) was characterized by a number of tendencies: to generate uninformative experiments, to make judgments based on inconclusive or insufficient evidence, to vacillate in their judgments, to ignore inconsistent data, to disregard surprising results, to focus on causal factors and ignore noncausal factors, to be influenced by prior belief, to have difficulty disconfirming prior beliefs, and to be unsystematic in recording plans, data, and outcomes. (Zimmerman 2000, p. 129)

Adults appear to differ from children in that they typically "needed to see the results of several experiments. Rather than ignoring inconsistencies, adults tried to make sense of them. Adults were more likely to consider multiple hypotheses (e.g., Dunbar & Klahr, 1989; Klahr et al., 1993)." (Zimmerman 2000, p. 134). Both children and adults used multiple strategies and Zimmerman accurately points out that the transition to more productive strategies is gradual rather than abrupt. Experience improves performance among children whether that experience is gained over chronological age or over multiple sessions with the same experimental system and education improves performance among adults (Amsel and Brock 1996).

Several pedagogically important distinctions arise from this set of work. The first distinction is that hypothesis generation or properly framing the question appears to be a more challenging task than selecting or designing appropriate experimental protocols. Even elementary school children (first graders) are facile in selecting the appropriate experimental design to differentiate between two conflicting hypotheses and can generate empirical procedures on their own to test two given alternative hypotheses (Sodian, Zaitchik et al. 1991). Specifically, when 1st and 2nd graders were presented with two mutually exclusive and exhaustive hypotheses, over 50% of the

1st graders and 86% of the 2nd graders could correctly design an experimental test to distinguish between the two proposed explanations. When presented with a problem but no suggested explanations, however, only about 25% of the children in both grades could generate spontaneous solutions (e.g. appropriate tests). Thus, the skill of properly framing the question appears to be more difficult than the skills of controlling variables or identifying causal relationships (see again the Coke-Cola taste study Schraagen 1990; Sodian, Zaitchik et al. 1991), though it should be noted that control of multiple variables (causal and non-causal) is still quite challenging for elementary age children (Kuhn 2007). This outcome makes sense as properly framing the question requires contextual knowledge to select the plausible solutions from the infinite universe of possible solutions, but selection of an appropriate methodology to test two hypotheses requires considerably less contextual knowledge. Phrasing the proposed solution in a productive way (to allow easy differentiation and refutation) is likely also a learned skill.

Pedagogical implications of research on expertise

Thus, when attempting to teach scientific reasoning, it appears to be productive to initially scaffold students by providing them with hypotheses and having them focus on developing methodological competency (control of variables, understanding of co-variation, replication, etc.). More experienced students can then be given the more challenging task of framing their own questions as well as determining appropriate methods. Lajoie (2003) in particular points out a weakness in the field that has invested much in differentiating between the abilities of novices vs. experts while neglecting to develop corresponding pedagogical methods to help students better develop expertise. She recommends that the trajectory and qualities of expertise must be made explicit to students and that they benefit from a focus on metacognitive elements because experts have "a better awareness of what they know and do not know" (Lajoie 2003 p. 21). In particular, she recommends that students be facilitated by a "continuous interacting hierarchy of novice to intermediate learners" supervised by an expert in a collaborative real-world setting (Lajoie 2003 p. 22).

Students should be provided with multiple representations of a problem to allow comparison [and] frequent situations that force them to reflect on the results of their actions. Frequent embedded formative feedback and expert intervention

highlighting the need for ongoing practice and refinement are also key issues for success (Lajoie 2003). These elements have been shown to greatly accelerate students' development of the qualities associated with expertise such as greater pattern recognition, metacognition regarding the limitations of one's own knowledge and deeper, more highly structured knowledge. For example, 20 hours of training using a simulation presenting new avionics technicians with multiple problems to be solved and formative feedback including suggested strategies improved performance to a level equivalent to almost four years of experience (Nichols, Pokorny et al. 1992). In particular, such training simulations accelerate learning because they present learners with both common and unusual situations allowing learners to perceive underlying principles more rapidly than with on the job training wherein long time periods are required to experience such contrasting and unusual situations (Ong 2007).

Thus, research on expertise appears to suggest that collaboration, reflection and metacognition facilitate the development of scientific reasoning. In the scientific community, however, the scientific reasoning does not end when the last data point is collected, but continues as outcomes are communicated to colleagues via scientific writing (Yore, Florence et al. 2006) and similar to scientific reasoning, scientific writing is also an explicit skill that must be taught (Campbell, Kaunda, Allie, Buffler, & Lubben, 2000; Keys, 1999a; Lerner, 2007). As both are specialized skills, any measure of scientific reasoning via scientific writing is confounded; effective reasoning may be obscured by poor writing or simply a lack of proficiency with science writing (Lerner 2007). Thus the use of science writing as a data source for scientific reasoning is a more conservative measure than direct observation. As scientific writing is a highly authentic and readily available data source for assessing scientific reasoning in higher education however, this constraint is acknowledged, but does not alter the decision to use scientific reports as a major data source.

Defining features of scientific writing

Scientific writing varies notably from other forms of formal writing and from informal writing as well. Keys (1999a) identifies scientific writing as differing from other forms of writing in that scientists use:

[1.] grammatical metaphor, or the condensation of several words that describe an action or process into a single noun, such as *photosynthesis, metamorphosis,* or *polymerization*…

[2.] expansion, the building of semantic relationships between events by the use of additional clauses that further specify, define, or extend the initial clause. Expansion includes three main types: (a) elaboration—further defining or clarifying an idea; (b) extension—joining two unique but related ideas; and (c) enhancement—qualifying with further information such as time, place, cause, or condition…

[3.] lexical density: a high number of content words per clause;

[4.] writing in the voice of third person; and…

[5.] many [explanatory words] for events, such as *cause, represent, produce*, and *form*. [numbers and all underlined emphases added, italicized emphases are original] (p. 1046)

In addition to the differences delineated above, scientific writing also commonly uses an identifiable organizational format with an introduction/ purpose, methods, results, conclusions (Keys 1999), cites references to other scientific work and contains figures, tables or other graphical representations of data. Scientific writing also strives to avoid value-laden adjectives or adverbs in an attempt at objectivity. A complete absence of bias is impossible however, as even the questions scientists choose to pursue are affected by societal and personal influences (Kuhn 1970; Simonton 2004). It has been recognised since the early 1900s that scientific writing differs sufficiently from other forms of writing and that to learn to write scientifically, students must either be taught its conventions explicitly or be notably enculturated to scientific writing as a genre (Keys 1999; Lerner 2007).

How science writing can facilitate science reasoning

Writing in and of itself has historically been viewed as an effective pedagogical tool due to the creation and ownership of knowledge and reflection it encourages and has evolved into movements such as the 'Writing Across the Curriculum' and 'Writing to Learn' which have grown in higher education and K-12 institutions over the last few decades (Connally & Vilardi, 1989; Keys, 1999b; Klein, 1999) and in science laboratories in particular (Lerner 2007). There has been debate

over whether expressive creative (more personal) writing vs. more constrained, scientifically focused writing has greater power to engender learning and many advocates of writing to learn suggest expressive or informal writing is beneficial in science classrooms as well (Hand, Prain, & Wallace, 2002; Keys, 1999b; Keys, Hand, Prain, & Collins, 1999). To some degree, that debate is irrelevant to our purposes here. Even if creative writing were to be more effective at stimulating engagement, the specific skill of comparing and evaluating scientific ideas and justifying scientific conclusions with data in a persuasive written form is a desirable skill that can only be acquired by practice in the genre of scientific writing (Keys 1999).

Most active research scientists in academic settings appear to view writing as an integral component of their science and a process that stimulates them intellectually. Yore, Hand and Florence (2004) surveyed and interviewed tenured and tenure-track faculty from life-sciences, physical sciences and engineering disciplines. Explicitly, the scientists offered the idea that writing serves more than just a communication function; it also helps them to improve the clarity of their ideas, generate new insights and synthesise the information in new ways (Yore et al., 2006; Yore et al., 2004), though the recognition of the transformative effect of writing was more tacit for some individuals than others (Yore, Hand et al. 2002). In particular, scientists interviewed by Yore and colleagues valued the reflection and metacognition that writing encouraged and the resulting self-assessment. "[W]riting helped them to clarify ideas and detect faults in logic, inconsistencies in claims, evidence and warrants, and voids in background" (Yore et al., 2004, p. 364).

Developmental trajectories of student scientific writing

Past research provides important clues as to how writing is best incorporated into the curriculum. Writing can stimulate scientific reasoning and knowledge generation in students as young as middle school (Keys 2000) and indeed explicit instruction on the role of writing in scientific knowledge generation is heavily recommended (Campbell et al., 2000; Keys, 1994, 1999b, 2000). The simple act of writing alone is insufficient however. When students are asked to communicate the outcome of their scientific investigations without any specific writing prompts, most students simply regurgitate factual information with little interpretation or discussion. Explicit identification of writing goals to students, such as the consideration of

alternative explanations, or the need to justify conclusions or provide rationales, appears to be necessary to facilitate synthesis, reflection and scientific knowledge generation (Keys 1999).

For example, Campbell and colleagues (2000) gave incoming university freshman a physical science problem set in a real world context and asked ~~to~~ them to work in collaborative groups to differentiate between two alternative hypotheses. The outcome of their work was to be written into "…a full report detailing all aspects of the experiments, measurements, calculations and graphs as well as your findings on who is right or wrong." (Campbell, Kaunda et al. 2000, p. 842). The purpose of the study was to determine students' baseline abilities prior to instruction. This contextualized, real-world problem solving situation contained all the elements desired in a quality scientific report – alternative explanations that must be tested and refuted, multiple possible methodologies, a need for justification of conclusions, etc. Student reports fell far short of hopes and expectations however. Most students failed to see the connection between methodology and resulting data. No student's report contained methodological rationales and several lacked any description of methods at all. The authors did not report on the degree to which students' conclusions were supported by data, but did recommend that "[i]f an ability to communicate [ones' scientific process] is considered a necessary element of science learning at university level then the communication of science should be taught explicitly and alongside the procedures and concepts of science." (Campbell, Kaunda et al. 2000, p. 851).

Similarly, Keys (1999a) found that when 8th graders were asked to conduct two inquiry investigations and "provide a written report detailing the behaviors that you observed while watching your animal." (p. 1047) and "evaluate the creek water…based on what you know about physical characteristics, chemical characteristics and macroinvertebrates," (p. 1048) that 50% of the individual reports and 75% of the collaborative reports results in simple "knowledge telling" in which content knowledge could be dense, but inferences and syntheses were rare or non-existent ($\leq$ 3 inferences per report). Students also failed to provide information on their methodologies and only reported results, often without any interpretation or conclusion (Keys, 1999a). In contrast, when Keys (2000) provided a similar population of 7-9th graders engaged in a soil erosion inquiry project with a writing prompt that requested:

1. Your scientific opinion of how bad the erosion is…
2. Detailed evidence supporting your opinion, including the results of specific observations and measurements.
3. A description of how you carried out the observations and measurements.
4. How your findings compared with your predictions…
5. Possible causes of [the] erosion… (p. 689)

All students wrote reports that included appropriate data and observations, methodologies and reasoned conclusions. Most "…generated new knowledge and explanations specifically from the act of writing…. Therefore students as young as eighth grade can engage the mature cognitive process of making the 'return trip' from the [written] discourse space back to the content space." (p. 687). Further, Keys asserts that middle school students when properly supported can "engage in high levels of scientific thinking including generating hypotheses, evidence, meaning for patterns, and knowledge claims. Thus, they learned science from the writing experience…" (p. 688). The act of writing and conveying ones' scientific journey can therefore be seen as an important component of that journey. Practicing scientists find that the act of writing enhances or participates in their processes of scientific discovery.

Summary of the role of writing in scientific reasoning

Students (novice scientists) should therefore be encouraged to generate both scientific insight and knowledge while learning to effectively communicate through writing. Writing serves as both a means of facilitating scientific reasoning as well as providing a data source for evaluating students' abilities. The benefits of writing are further likely to be multiplied when writing is combined with peer review. Peer review may accelerate the development of both scientific reasoning and writing because of the critical thinking and evaluation skills required and repeatedly practiced as students evaluate the claims and evidence of their peers. The inevitable comparisons that will be made when students evaluate the work of others may also stimulate self-assessment and metacognition.

What is peer review?

Definition of peer review

Peer review is the evaluation of scientific work and reasoning by scholars who work in similar or complementary areas (peers) to determine whether or not proposed work should be funded or published (National Science Foundation 2008) and is a ubiquitous process in science (Ziman 1998). Success as a scientist is predominantly defined by one's ability to publish in peer-reviewed journals and secure grant funding (Mervis 2000). Despite its near universal use as a means of providing active feedback and exerting subsequent influence on the forefront of research, only a small amount of research has been conducted on how scientists respond to peer review. When interviewed, scientists have reported that addressing reviewers' comments about their writing forced them to assess, monitor, and regulate their science inquiries and research reports (Yore, Florence et al. 2006). Hence, engaging students in peer review is likely to also stimulate reflection and metacognition, thus facilitating the development of scientific reasoning skills.

Why is peer review likely to improve scientific reasoning?

Peer-peer collaboration improves student learning

The positive impact of collaborative learning is generally well accepted (Boyer Commission, 2001; Committee on Undergraduate Biology Education, 2003; Duit & Confrey, 1996) and has a strong empirical research base in science classrooms (e.g. Hogan, Nastasi, & Pressley, 2000; Jensen & Finley, 1996; Osborne, Erduran, & Simon, 2004). Peer review is a specialized form of peer collaboration, but even such asynchronous, online, written collaboration can cause learning gains (Hoadley and Linn 2000). Mechanisms by which peer feedback can stimulate learning or insight include identifying misconceptions, gaps in logic and unrecognized assumptions. Even when peers are novices, peer-peer collaboration is helpful so long as each peer possesses different inaccuracies or inadequacies (Schwarz, Neuman & Biezuner, 2000).

In fact, when Schwarz and colleagues (2000) paired 44 students who both held misconceptions about the relative value of fractions, three quarters (77%) of the students could answer correctly after working through collaborative inquiry tasks with a misinformed peer (a significant gain at the $p = 0.05$ level) and for nine of the

pairs, both students overcame their misconceptions by the end of the task. In contrast, of the 10 students with misconceptions who were paired with a competent peer (one who had answered the pre-test question correctly), only half could answer the post-test question successfully which is a non-significant gain ($p > 0.05$). The authors attribute the greater success of peers who were paired with another incorrect peer to the fact that peers who held contrasting misconceptions identified and corrected those misconceptions through the collaborative process because they engaged in argumentation and justification. In contrast, social dynamics appear to differ when one peer is competent and the other not; far less justification and testing of ideas ensued in these pairs (Schwarz et al., 2000).

While the collaboration of peers reviewing written work will not contain so much active justification and back and forth discussion, this result is important because it indicates that both ends of the peer spectrum can provide useful feedback. Namely, the active requirements of the peer review process will overcome the lack of interaction in 'right-wrong' pairs found by Schwarz and colleagues. Competent peers are thus quite likely to provide useful feedback, while these results suggest that less competent peers may also stimulate positive revisions through contrasting misconceptions. Thus, even less competent peers can likely provide useful feedback and discussion points to their peers. Similarly, work of Rijlaarsdam and colleagues (2006) focusing on typcial students demonstrated that receiving *any* type of feedback caused significant gains, but students receiving *written* feedback *specific* to their papers had the largest gains. While the use of collaborative work is therefore well acknowledged in K-12 pedagogical literature, the benefits of formative feedback on student learning are less commonly acknowledged in higher education (Yorke 2003).

Peer review of science writing is likely to be a particularly effective source of collaboration and formative feedback both because of its authenticity as a scientific skill and because past research suggests it may provide increased opportunities to practice evaluative skills, increase engagement and has a tendency to cause reflection and metacognition.

Peer review provides multiple opportunities to practice scientific reasoning skills

Each peer review experience exposes a student to multiple contrasting examples in the form of peers' work. At the institution where this research took place each student reviewed three peers' papers. In addition, most instructors who

implemented peer review also provided three additional exemplars identified as poor, average and high quality to assist students in understanding how the specified criteria could be enacted with varying degrees of success. Thus, each student who engaged in peer review during this research project saw at least three unique examples and often six unique examples in addition to their own paper.

Multiple contrasting examples of peers' work can also be a critical aid to helping students to determine the salient criteria for a given task (Bransford, Franks et al. 1989). Multiple contrasting examples have been shown to be important in helping students identify which aspects of a phenomenon are relevant and which are incidental (Driver and Scott 1996). These multiple opportunities to apply one's knowledge of the criteria in a relevant context would facilitate students finding weaknesses in their own work. In addition, peer review simply increases the time that students spend comparing and evaluating scientific thoughts. In and of itself, time on task has been found to correlate with greater achievement regardless of instructional method (Admiraal, Wubbels, & Pilot, 1999; Timmerman, Strickland, & Carstensen, 2008; Trowbridge & Wandersee, 1994). Moreover, concerted and repeated practice over time is an important component of developing expertise (Ericsson and Charness 1994).

Beyond just the additional opportunities to practice critical thinking and evaluation skills, peer review potentially provides relevant formative feedback. Formative feedback has been shown to have significant benefits for improving student work and learning (Chinn & Hilgers, 2000; Ravitz, 2002; Topping, Smith, Swanson, & Elliot, 2000; Yorke, 2003). Without feedback, students cannot assess whether or not their conceptions are accurate or indeed whether or not they are learning at all.

Peer review encourages reflection and metacognition

Reflection and metacognition are also critical facilitators of meaningful learning (Baird & White, 1996; Bloxham & West, 2004; Pope, 2005; Yore et al., 2002). Students' comments on end-of-term evaluations at our institution and elsewhere (Stefani 1994) indicated that the process of peer review often stimulates such reflection and self-assessment possibly leading to metacognition. Metacognition is the conscious control of one's learning; when students are metacognitive, they are aware of where and when and how they have learned

something fostering greater construction of knowledge or conceptual change (Schraw and Dennison 1994; Baird and White 1996). The act of writing and particularly the act of revision itself often leads to metacognition (Keys et al., 1999; Klein, 1999; Yore et al., 2004). Other researchers have suggested that undergraduates may gain as much from the act of reviewing as from the peer feedback they receive because of the self-reflection which is stimulated when student view others' work and make internal comparisons with their own (Cho, Schunn et al. 2006). Peer review may thus stimulate metacognition on both levels: because of the concrete acts of writing and revision as well as the forced evaluation of other's work.

Peer review increases engagement in coursework

Increased personal relevance and increased student involvement are commonly indicated preferences of students in a wide variety of classrooms (Fraser 1998; Fraser 2002). As peer review and scientific writing are authentic scientific skills and desired professional competencies (Cicchetti, 1991; Marsh & Bazeley, 1999; Yore et al., 2004), it is plausible that students, as aspiring scientists, would perceive opportunities to practice these skills as personally relevant. Assessing the quality of peers' papers and receiving feedback also is likely to make students feel more actively engaged in the assignment. There is a general perception among faculty who use peer review in their courses that students respond positively. Marcoulides and Simkin report student comments such as "This was very useful to me. Why don't other professors do this?" (1995, p. 223). More systematic and quantitative evaluations of the effects of peer review are "alarmingly sparse" however in the words of Hanrahan and Isaacs (2001) who conducted the only evident quantitative study of student perceptions of peer review in science writing.

When students were given the query "What do you think were the pros and cons of doing peer and self-assessment on the essay assignment?" a survey of students in a third year health psychology course found eight major themes in student perceptions of the peer review process (Hanrahan and Isaacs 2001). For the reader's convenience, these themes are highlighted in italics. Students reported that peer review was *difficult*, though most students focused on either the difficulty of being objective in one's self-assessment, or perceived lack of credibility of peers or unfamiliarity with subject matter (essay topics varied and the course contained students in multiple programs of study) (Hanrahan and Isaacs 2001). As all students

write on the same topic when doing peer review in biology courses these concerns are not terribly applicable to this study. Students also reported that peer feedback provided a *better understanding* of the assignment and its criteria and that being able to *compare their work to others* was instructive (Hanrahan and Isaacs 2001). They reported that the process overall was *productive* and *caused an improvement in their writing* either through self-reflection or by developing critical thinking skills. Students identified the following *problems with implementation*: the process was time consuming, some peers put little effort into their assessments, and reviewers did not receive any feedback on the helpfulness of their comments. Lastly, students expressed that they felt *empathy* with their instructors, *motivation* to write well and impress their peers and *discomfort* with having their work exposed to others as well as with being placed in the role of critiquing someone else. Hanrahan and Isaacs (2001) did not quantify the frequency of these student perceptions, but do suggest that future work quantify and validate these themes to determine if the benefits observed by their study are universal to the process of peer review or specific to their context and situation.

In more informal surveys, students' comments were positive, and faculty perceived them to be more engaged and more motivated than in courses without peer review (Stefani 1994). In first year biology courses where students peer reviewed written laboratory reports (n = 120), "100% of the students said that [peer review] was more time consuming and over 75% said that it was hard," but "100% of the students said [it] made them think more, 85% said it made them learn more and 97% said that it was challenging" (Stefani 1994). In graduate psychology courses where students peer reviewed potential manuscripts (n = 33 students) they rated the process of peer review and revision as 8.9 ± 1.3 (on a scale where 1 = worthless, 10 = learned a lot) and the value of reviewing other people's papers as 7.9 ± 2.3 (Haaga 1993). In comparison, students rated the value of giving and watching oral presentations less favorably (7.0 ± 2.4 and 4.4 ± 2.1, respectively) indicating that they perceived peer review as a more useful exercise (Haaga 1993). Notably, even non-science majors more often identified peer-review as a preferred learning tool compared to other learning aids such as graphic organizers, videos, case studies, personal experience, or study group discussions (Pelaez, 2002).

Thus, there are anecdotal and qualitative data indicating that students find peer review beneficial though it is challenging and time consuming. In particular, students

believe it improves their writing and their ability to think critically, as well as stimulates reflection and metacognition by encouraging evaluation and comparison. A current gap in this research base is the lack of quantification of these perceptions however. Do the majority of students feel peer review is beneficial? Is improvement in critical thinking a rare or common perception? No information is provided in the literature as to how students perceive the role of peer review in the scientific community, nor if they see connections between the classroom and scientific community. Student perceptions are valuable for the insight they provide into motivation and effort. Such affective dimensions can also affect performance, but even if students were unanimous in their appreciation of peer review, such perceptions may or may not be accurate reflection of actual changes in students' performance as a result of peer review. To determine the effect of peer review on actual performance of scientific reasoning tasks, it is necessary to turn to other sources of data.

How do we measure students' scientific reasoning abilities?

Multiple measures for a complex concept

Scientific reasoning, by definition, occurs in the inaccessible interior of the mind. What is visible to the researcher are the outcomes or actions generated by these acts of reasoning. Potential data sources range from ethnographic observations of a person's actions while in the laboratory, to intentional communications such as written reports or oral explanations to direct questioning by an interviewer. Given that reasoning occurs within the mind, it is plausible that asking subjects to articulate their thought processes would be a direct measurement of their reasoning. Research shows that self-reports of reasoning are often inaccurate or incomplete however. Scientists have been shown to be: unaware of automated aspects of their thought processes and therefore leave out critical portions of their scientific reasoning processes (Feldon 2007) or oblivious to the synergies and distributed reasoning provided by collaborators (Dunbar 2000). Other professionals such as teachers are similarly shown to not be aware of disjunctions between their voiced intentions and their actions (Simmons et al., 1999; Wubbels, Brekelmans, & Hooymayers, 1992).

Reasoning is therefore best measured by a variety of data sources that provide triangulation for conclusions (Mathison 1988; Johnson and Onwuegbuzie 2004). Beyond the student perception of the impact of peer review already discussed above,

this study will focus on measuring student reasoning abilities using two different data sources: a detailed investigation of students' reasoning as evidenced in their science writing and a more conservative measure of their abilities on a previously published multiple-choice test. Subsequent sections focus on research relevant to the substance and reliabilities of rubrics for measuring reasoning in science writing and the previous findings of the *Scientific Reasoning Test* applicable to this purpose and student population.

The laboratory report is the chosen data source for determining student reasoning because it is commonly employed in many science courses (providing a natural source of data over several years for each student) and because it most closely approximates a real performance assessment: the scientific manuscript. In order to measure scientific reasoning using students' writing, relevant criteria must be selected or developed. Criteria used by professional referees for scientific journals combined with past research on rubrics for student science writing and discussions with biology faculty provide an appropriate context for the development of criteria that would apply across a wide variety of biology laboratory reports. Such criteria and a rubric built from them would be termed *universal* because it would identify the attributes of effective scientific reasoning and writing regardless of the subject matter of the assignment or course (within a science major). It would also provide students with a consistent and explicit set of criteria against which to measure their performance and therefore also function as an effective learning tool for students. As described above, when students are asking to communicate their research through scientific writing and are not provided criteria or goals for what should be included, the outcome is primarily just "knowledge telling" (exhaustive lists of factual statements or the final outcome of the work without rationales or explanations) (Keys, 1999a) even at the university level (Campbell, Kaunda et al. 2000). Development of a rubric thus provides a pedagogical as well as methodological benefit.

Criteria used in professional peer review

The National Science Foundation provides only two criteria applicable to all grant proposals: *intellectual merit* and *broader impacts* (2008). Intellectual merit combines a reviewer's assessment of the appropriateness of the research design with the qualifications of the researchers in light of the significance of the research topic

and the likelihood that it would have a "transformative" impact on knowledge in the field (National Science Foundation 2008). The broader impacts criterion is concerned with the synergistic potential of the proposal and likely societal impacts. Thus, for university students, these criteria indicate that students must develop a sense of the *context* and subsequent *significance* of their work as well as the *methodology* and outcomes.

Another source of information on the criteria used by professional referees are the instructions provided to reviewers of manuscripts submitted to journals. Reliability studies of professional peer review have for the most part focused on highly prestigious journals with high rejection rates in the social science and medical fields (Cicchetti, 1991). These studies are the primary source of information on the criteria used by professional peer review. In addition to the expected request for an overall recommendation on whether to publish the manuscript, two criteria were consistent across all journals for which information was available: *significance*, and *methodology* (Cicchetti, 1991; Marsh & Ball, 1989; Marsh & Bell, 1981; Petty, Fleming, & Fabrigar, 1999) (Table 2.1). Thus, it appears that these two criteria are broadly agreed upon in the scientific community and should be a focus of pedagogical strategies aimed at developing university students' scientific writing and reasoning abilities.

Other criteria commonly included in professional peer review included *writing quality* (Cicchetti, 1991; Marsh & Ball, 1989; Marsh & Bell, 1981), *literature review* (Cicchetti, 1991; Marsh & Ball, 1989; Marsh & Bell, 1981), *succinctness* (Cicchetti, 1991), *originality* (Cicchetti, 1991) and *theoretical context* (Petty, Fleming et al. 1999). To further investigate the criteria on which professionals judge science and science writing, Marsh and Ball (1989) did a content of analysis of written critiques of journal submissions and created an evaluation sheet with a total of 21 criteria. They then solicited at least two reviews for each of 278 manuscripts submitted to the *Journal of Educational Psychology* using these 21 criteria. Using factor analysis, they determined that all 21 items condensed back down to four criteria commonly stated in instructions to reviewers: 1) research methods, 2) relevance to readers, 3) presentation clarity, and 4) significance which were already identified by that and other journals (Table 2.1). Relevant or appropriate for a journal's readership is the only criterion listed above which is not relevant to student

papers. Thus, there appears to be a general agreement within the scientific community that *methodological competency, appropriate context, adequate literature review, strong significance* and *writing quality* are fundamental attributes of high quality research and science writing (Table 2.1 and Sternberg & Gordeeva, 1996). Consequently, there is strong support for including these criteria in the Universal Rubric for Laboratory Reports and such inclusion will assist students in identifying and developing the skills desired of practicing scientists.

Table 2.1. *Criteria Used in Professional Peer Review.*

	Journal of Abnormal Psychology[1]	**Journal of Personality and Social Psychology[2]**	**British Medical Journal[1]**	**Journal of Educational Psychology[3]**
Criteria				
Significance/ Importance	X	X	X	X
Design/Analysis/ Methodology	X	X	X	X
Writing quality	X			X
Literature review	X	X		
Appropriateness for this journal's readers	X		X	X
Succinctness	X			
Theoretical context		X		
Originality			X	

Note. All journals also asked reviewers to provide an overall recommendation regarding publication. [1](Cicchetti, 1991), [2](Petty et al., 1999), [3](Marsh and Ball, 1981, 1989)

Reliability of professional peer review

The selection of criteria and development of a rubric is only half of the process however. The utility of the rubric for both research data collection and use by instructors in biology laboratories must be evaluated by testing the reliability of the scores it generates when the rubric is applied to students' laboratory reports. As a means of providing context, the reliability of the rubric developed for this study was compared to the consistency of professional referees for science journals and against previously published relevant rubrics for measuring student written reasoning and

argumentation. Marsh and Ball (1989) reviewed 15 studies on professional peer review of manuscripts submitted to various social science journals from 1973 to 1984 and found that single rater reliabilities while significantly different from chance, were distressingly low (mean single-reviewer reliability of 0.27 $\pm$ 012; range 0.08 (n= 216 manuscripts) to 0.54 (n = 0.87)). "Single-reviewer reliability [was] defined as the correlation between two independent reviews of the same manuscript across a large number of manuscripts submitted for publication" (Marsh and Bell 1981). This finding of low reliability among referees has been replicated by other focused studies (r = 0.29 Petty, Fleming et al. 1999) and meta-analyses on scientific manuscripts (r = 0.07 to 0.37 Cicchetti, 1991). Grant proposals have higher reliabilities, which may be partially due to the greater number of reviewers (4 rater reliability for total score = 0.49, Marsh & Bazeley, 1999). Such reliability scores are still below those considered acceptable if a researcher attempted to publish an instrument with such scores however. At least, overall single rater recommendations to publish or fund are consistently more reliable than individual criteria (Cicchetti, 1991; Marsh & Ball, 1989; Marsh & Bell, 1981; Petty et al., 1999) but still hovered around 0.30 to 0.34.

Many explanations have been given for the disturbingly low consistency of professional reviewers. For example, editors may intentionally select reviewers of contrasting viewpoint for manuscripts on controversial topics, (Cicchetti, 1991), reviewers have been shown to have bias towards research in current "hot" topics, even at the expense of appropriate methodology (Wilson, DePaulo et al. 1993) and to favor some papers over others based on the prestige of the institution or author (Petty et al., 1999; Ross et al., 2006), author gender (Petty, Fleming et al. 1999), primary language of the author (Ross et al., 2006) or even text length (Petty, Fleming et al. 1999). One explanation conspicuously missing from these discussions however is that reviewers were provided simply with a list of criteria, rather than a rubric.

The distinction between criteria and rubrics

Criteria are central to evaluation. Without explicit identification of the qualities that are valued and sought after, evaluation cannot occur. These qualities may be defined to varying degrees and range from highly subjective (e.g. "outstanding research which advances the field") to highly objective (e.g. "text length in words"). The distinction between criteria and rubrics is that rubrics provide descriptions of the performance at level for each criterion (Kuhs, Johnson et al.

2001). These guidelines inform the reviewer as to where the decision lines between different levels should occur and thereby encourage reviewers to interpret the scale in the same way and use the same decision break points. Without a rubric, reviewers may have different expectations or definitions of performance levels leading to vastly different applications of the point scale. Even undergraduate peers from a wide variety of university settings can produced highly reliable ratings ($\alpha = 0.88$) when they use a well-defined rubric (Cho, Schunn, & Wilson, 2006).

Use of rubrics in higher education

Whilst faculty in many higher education institutions require graduate teaching assistants to use criteria or evaluation worksheets when grading undergraduate laboratory reports (our institution and Kelly and Takao 2002), there is a paucity of published research on the reliability or validity of rubrics in that context or on the natural consistency of graduate students in assigning grades. The only information found by this author to date indicates that graduate teaching assistant and faculty instructor grades correlate poorly when using the same detailed list of 14 criteria (Kelly and Takao 2002). A rubric designed around an epistemic model of students' claims and justifications produced high reliability ($r = 0.80$) when two trained raters applied it to the same papers graded by the teaching assistants. The rater's relative rankings did not correlate however with either the graduate teaching assistant scores ($r = 0.12$), nor the faculty instructor's relative ranking of the case study papers. This finding suggests that there may be a strong difference in the efficacy of a list of criteria versus a rubric and that graduate teaching assistants and instructor would benefit from the use of rubric.

Need for an appropriate rubric for university science writing

Educators and mentors frequently identify achievement goals for students in terms of writing and reasoning and advocate the use of rubrics across a great variety of fields (Arter & Mctighe, 2001; Kuhs et al., 2001; Trevisan, Davis, Calkins, & Gentili, 1999), but this author has yet to find a rubric in the published literature applicable to university science writing that has been psychometrically tested. Extensive and well-reasoned published rubrics for scientific reasoning exist, but lack reliability testing (Halonen et al., 2003) or are designed for venues other than writing (observing students in the lab, Baxter, Shavelson, Goldman, & Pine, 1992; Germann & Aram, 1996; oral presentations, Hafner & Hafner, 2003; verbal discussions, Hogan

et al., 2000). For example, Hogan, Nastasi and Pressley (2000) specifically developed a rubric to assess reasoning complexity, but it was designed for the less structured venue of group discussions. Consequently, while several of the six criteria are relevant (*justification, detail, explanation, logical coherence*) because the performance levels of this rubric largely count the number of instances in which any student exhibits the behavior, it could not be usefully applied to student written work. For example, the performance levels for the criterion of *justification* (which is concept central to the Universal Rubric) are based on the number of justifications provided per idea, without any overarching sense of the quality or priority of the justifications. Such criteria make sense with the more free-ranging nature of group discussion, but this rubric is largely uninformative when transferred to written work.

It should be noted however that Hogan and colleagues' (2000) scheme does provide direct support for the Universal Rubric criterion of *Discussion: refuting alternative explanations* as their definition of synthesis is "a measure of how and if opposite views are accounted for, which is a hallmark of dialectical and higher order thinking" (Table 5, p. 398). Besides being hard to translate in practice, when Baxter, et al. (1992) compared rubric reliabilities from observations of students performing laboratory experiments vs. those same students' laboratory notebooks they found that reliability scores varied based on medium. Therefore, rubrics developed for other media suggest general concepts or priorities, but cannot be borrowed directly as measurement tools for written laboratory reports.

Rubrics designed specifically to assess writing abound and many have been reliability tested, but they are for non-science writing forms such as narratives or persuasive essays (Baker, Abedi, Linn, & Niemi, 1995; Marcoulides & Simkin, 1995; Novak, Herman, & Gearhart, 1996; Penny, Johnson, & Gordon, 2000) or even if they can be applied to science writing, are so general as to prevent them from being useful for assessing scientific reasoning and other domain specific abilities. For example, Cho, Schunn and Wilson (2006) have a rubric which has been applied to writing in 16 different courses ranging from history to psychology at four different higher education institutions. This rubric has been shown to be highly reliable both in terms of agreement among peer reviewers ($\alpha = 0.88$) and between peers and instructor assessments ($\alpha = 0.89$). But the three criteria comprising that rubric (*flow, logic* and *insight*) do not address many of the qualities valued in the scientific community (e.g.

intellectual context, significance, methodology) and are therefore insufficient for evaluating scientific reasoning in particular.

What are available for science writing at the university and post graduate levels are criteria lists (Haaga, 1993; Kelly & Takao, 2002; Topping et al., 2000), but because they lack performance levels they consequently have poor reliability among raters. While Topping (2000) found high similarity in the counts of positive, negative or neutral comments made between peer reviewers and instructors, most other studies investigating the actual point values assigned by peer raters versus instructors had very little consistency (Haaga 1993), even when just ranking papers (Kelly and Takao 2002). Specifically, the teaching assistant and the course professor agreed on the ranking of only one out of four case study papers and assigned similar total points scores for only two out of four (Kelly & Takao, 2002). In contrast, when the researchers reviewed the same papers and assessed the epistemic levels of student argumentation using a complex rubric, the inter-rater reliability was $r = 0.80$.

It thus becomes clear that rubrics generate far more reliable and therefore, informative assessments of students' scientific writing than do lists of criteria. Use of a rubric is therefore advocated for any measure of student performance. The question then becomes, what criteria should be included in the rubric and what performance levels should be defined?

Table 2.2. *Common Themes in Published Criteria for Science Writing or Scientific Reasoning.*

	(Kelly and Takao 2002)[1]	**(Halonen et al., 2003)**	**(Haaga, 1993)[2]**	**(Topping et al., 2000)[3]**	**Professional peer review or other research literature**
Study Context	Undergraduate oceanography scientific report	Desired psychology curriculum outcomes	Graduate psychology manuscripts	Graduate psychology term papers	
Instrument reliability	Not tested	Not tested	r = 0.55	Not tested	
Performance levels	None specified	(5) "Before training" to "Professional"	None specified	None specified	
Universal Rubric for Laboratory Reports Criteria					
Context	"Clear distinction between portions of the theoretical model supported by data/ background knowledge and those which are still [untested.]"	Degree of theoretical/ conceptual framework (Table 2, p. 199)	"background (primary lit) is covered adequately"	"clear conceptualization of the main issues" "literature review"	
Significance					(Cicchetti, 1991; Marsh and Ball, 1981, 1989; Petty et al., 1999; Sternberg & Gordeeva 1996)
Hypotheses are Testable	"A clear, solvable problem is posed…				
Hypotheses have Scientific Merit	…based on an accurate understanding of the underlying theory."				

	(Kelly and Takao 2002)[1]	(Halonen et al., 2003)	(Haaga, 1993)[2]	(Topping et al., 2000)[3]	Professional peer review or other research literature
Experimental Design	"Multiple kinds of data are used when available."	Sophisticated observational techniques, high standards for adherence to scientific method, optimal use of measurement strategies, innovative use of methods (Tables 1 & 3, p. 198-199)			(Cicchetti, 1991; Marsh and Ball, 1981, 1989; Petty et al., 1999)
Data Selection	"Available data are used effectively. Data are relevant to the investigation."			"new data (type, range, quality)"	
Data Presentation	"Observations are clearly supported by figures"				
Statistics		"Uses statistical reasoning routinely" (Table 3, p. 199)			
Conclusions based on data	"Conclusions are supported by the data." "Text clearly explains how the data support the interpretations."	"uses skepticism routinely as an evaluative tool" seeks parsimony (Table 5, p. 200)	"conclusions follow logically from evidence and arguments presented"	"conclusions/ synthesis"	
Alternative explanations					(Dunbar 1997; Hogan, Nastasi, & Pressley, 2000)
Limitations		Understands limitations of methods, "bias detection and management" (Table 3, p. 199)			
Primary Literature	"Data adequately referenced."	Selects relevant, current, high quality evidence, uses APA format (Table 6, p. 201)		"references" "Literature review"	

	(Kelly and Takao 2002)[1]	(Halonen et al., 2003)	(Haaga, 1993)[2]	(Topping et al., 2000)[3]	Professional peer review or other research literature
Writing Quality	"Clear, readable focused and interesting. Accurate punctuation and spelling. Technical paper format [complete and correct]."	Organization, awareness of audience, persuasiveness, grammar (Table 6, p. 201)	"well written (clear, concise, logical organization and smooth transitions"	"structure (headings, paragraphs); precision and economy of language; spelling, punctuation syntax"	
Additional criteria expressed in literature, but not included in the rubric as they either lacked universality, or were prioritized less by faculty as discrete concepts.	Clear distinction between observations and interpretations. Epistemic level: Arguments build from concrete data to more abstract theory. Each theoretical claim supported by multiple data sources.	Awareness, evaluation of and adherence to ethical standards, and practice. (Table 4, p. 200) Scientific attitudes and values: enthusiasm, objectivity, parsimony, skepticism, tolerance of ambiguity (Table 5, p. 200)	"Goals of the paper are made clear early" "Scope of the paper is appropriate (not over-reaching or over broad)"	"psychology content" "Advance organizers (abstract, contents)" "originality of thought" "action orientation"	

Note: If not indicated directly in the table, quotations were found as follows [1]Kelly and Takao, 2002, Table 1 p. 319; [2]Haaga, 1993, Table 1 p. 29; [3]Topping et al., 2000, Appendix 1, p. 167

Selection of criteria for a universal laboratory report rubric at the university level

Beyond the demonstrated need to use rubrics instead of simply lists of criteria when evaluating papers, what can be gleaned from these studies are the qualities valued in science writing at the university level. A survey of research literature on the subject of university-level science writing found four relevant papers that indicated the criteria by which students' scientific reasoning skills were judged (Table 2.2). When the criteria espoused for scientific writing and reasoning at the university level are compared, consensus is achieved for *context, conclusions solidly derived from data,* and *writing quality* (Haaga, 1993, Halonen, 2003, Kelly, 2002, Topping, 2000). Broad support is generated for *use of primary literature*, and *experimental design* (Halonen et al., 2003; Kelly & Takao, 2002; Topping et al., 2000). Surprisingly, while *context, methodology, primary literature* and *writing quality* appear in both the pedagogical and professional peer review criteria lists, *significance* is conspicuously absent from the classroom based lists despite its ubiquitous use as a criterion in the scientific community (compare Tables 2.1 and 2.2). Conversely, the criterion *conclusions justified by data* is found in all the pedagogical criteria lists and is absent from the professional referee considerations. This author hypothesizes that the absence of *significance* from classroom evaluations is a likely result of instructors feeling that students lack the content background to fully appreciate the implications of scientific work and see gaps in knowledge or inconsistencies in the field. Why professional peer review criteria do not list *extent to which conclusions are justified by data* is considerably less clear and open to investigation.

This failure to make clear to students that an explanation of the *significance* of scientific work is a desirable quality hinders their development as practitioners of science. Making the *significance* of completed work clear should be identified as a goal of science writing at the university level for two reasons. Firstly, as the scientific community appears to universally value significance when evaluating scientific writing, it must be included in any honest attempt to develop students' scientific reasoning abilities. Omitting it would hinder students' development as practicing scholars. Secondly, students will not strive to understand or consider *significance* as an issue in their work or writing unless it is identified to them as a valuable attribute. The values of the scientific and science education communities thus provide an important foundation for the development of the Universal Rubric.

Review of Table 2.2 thus indicates that all the criteria comprising the Universal Rubric have support from the science education and/or scientific community.

Historical perspective on rubric criteria

The reader may also care to recall the discussion of the role of content knowledge in scientific reasoning begun at the beginning of this chapter and note that while the criteria selected for the rubric are domain general (not dependent on content knowledge in any particular area of science), they explicitly acknowledge that proficiency in scientific reasoning requires a strong knowledge of the subject matter and familiarity with the context and procedures of the field. These criteria thereby represent a shift in the definition and values of scientific education. Earlier works focused on a dichotomy between reasoning strategies vs. conceptual knowledge. The current consensus of priorities and values described here suggests that neither of those viewpoints is sufficient and that the ability to integrate formal reasoning and contextual knowledge now comprises a major component of scientific reasoning.

Performance levels for scientific reasoning rubrics

Rubrics are differentiated from lists of criteria by the inclusion of descriptions of possible student performance at designated levels. A literature search produced one published rubric for scientific reasoning with relevant performance levels. Halonen et al. (2003) performance levels range from *before training* to *professional graduate and beyond* (Figure 2.1).

Table 3. Description of Proficiency Levels in the Problem Solving Skills Domain

	Levels of Proficiency				
Components of Problem Solving Skills	Before Training	Basic Introductory Psychology	Developing	Integrating Advanced Undergraduate	Professional Graduate and Beyond
Methods skills: Recognition, evaluation, generation	Does not rely on scientific method	Recites steps in conducting research; articulates basic knowledge of correlational and causal techniques; acknowledges value of controlled comparisons	Selects and applies appropriate method in simple projects; operationalizes and isolates variables; identifies influence of extraneous variables	Selects and applies appropriate method to maximize validity and reduce alternative explanations	Develops unique applications of research methods; establishes a research focus that identifies and builds on primary interests in behavior

Figure 2.1. A portion of Table 3 from Halonen et al. (2003, p. 199) describing the performance levels for a criterion. Publisher provides permission for reproduction in theses and dissertations free of charge.

Similar performance levels were selected for the Universal Rubric developed for this study. Student performance was expected to range from *not addressed* (no evidence that the student attempted to accomplish the criterion) through *novice*, and *intermediate* to *proficient* (performance expected of an outstanding undergraduate or beginning graduate student).

Research significance of a Universal Rubric for science writing

Given that no psychometrically tested rubrics for experimentally based science writing have been found in the literature, it appears that the development and testing of such a Universal Rubric would make a notable contribution both as a research instrument and as a pedagogical tool. University faculty, teaching assistants and other practitioners might find it applicable to their pedagogical goals and implement it directly. Other researchers might benefit from using criteria which align with those used in professional peer review and other research (including this study) to compare the reliabilities of student peer reviewers or the reliability of various pedagogical groups (teaching assistants, faculty). Testing of such a rubric using graduate teaching assistants would provide faculty and department chairs with sorely lacking information as to the natural consistency of these ubiquitous instructors who so far have been mostly vastly overlooked in terms of professional development and pedagogical support (Gaff, 2002; Golde, 2001; Luft, Kurdziel, Roehrig, Turner, & Wertsch, 2004). Finally, a rubric independent of subject area allows comparison of student performance across multiple courses and assignments providing a previously impossible longitudinal analysis of the development of students as scientists.

The Scientific Reasoning Test

Such a fine-grained and detailed analysis of student performance restricts the investigator to a smaller sample sizes (tens of students) however due to the intense time and effort that is required to produce each datum. When one desires to sample a larger proportion, or perhaps the entire student population in question (hundreds to thousands of students) and one does not have vast resources, a coarser grained means of assessing student scientific reasoning ability is useful. One such instrument is the *Scientific Reasoning Test (SRT)* (Lawson 1978). Developed to assess university students' scientific reasoning abilities across a variety of subject matters (biology and physics), it has been applied repeatedly in higher education biology courses

(Lawson, 1978; Lawson, 1979, 1980, 1983, 1992; Lawson, Alkhoury, Benford, Clark, & Falconer, 2000; Lawson, Baker, Didonato, Verdi, & Johnson, 1993; Lawson, Banks, & Logvin, 2007) and non-biological high school settings (Norman, 1997; Westbrook & Rogers, 1994) and found to be reliable in such contexts (Table 2.3). Positive correlations have also been found between student performance on the *Scientific Reasoning Test* and self-efficacy (Lawson, Banks et al. 2007), computational ability (Lawson 1983) and biology achievement (Lawson et al., 2000) as well as using it as a means of assessing the effectiveness of curriculum reform efforts on student scientific reasoning ability (Lawson et al., 1993; Westbrook & Rogers, 1994).

The *Scientific Reasoning Test* is based on a Piagetian understanding wherein student reasoning abilities vary across of spectrum from concrete reasoning which "…makes use of direct experience, concrete objects and familiar actions…" to formal reasoning which "…is based on abstraction and that transcends experience…" (Karplus, 1977, p. 364). It presumes that reasoning is independent of content knowledge, but uses examples that are reasonably familiar to secondary and tertiary students in western nations.

Table 2.3 *Published Reliability Scores for the* Scientific Reasoning Test.

Citation	**# students**	**Context**	**Reliability score**
(Lawson 1978)	513	Year 8, 9, 10 science, English and biology	0.86
(Lawson 1983)	96	Undergraduate biology	0.76
(Norman 1997)	60	Year 11 and 12 chemistry	0.78
(Lawson, Baker et al. 1993)	77	Undergraduate biology	0.55[1]
(Lawson et al., 2000)	663	Undergraduate biology	0.81
(Lawson, Banks et al. 2007)	459	Undergraduate biology	0.79[2]

Note. Reliability scores are Cronbach's alpha ([1]indicates a split half reliability) unless indicated to be [2]Kuder Richardson (KR20).

It is further useful as most of the questions on the *Scientific Reasoning Test* use a physical science context thereby avoiding bias towards any one biology class when the test is applied across the curriculum. Therefore, as a more distal measure of scientific reasoning ability, the *SRT* would also offer insight as to the transferability

of the scientific reasoning skills gained by peer review. A robust finding of the cognitive psychology literature is that performance declines whenever people are asked to solve abstract logic problems or real-world problems outside of the knowledge domain in which they learned the reasoning strategy (21 studies reviewed in Zimmerman, 2000). This decline occurs even when the principles behind the problems are identical. Therefore, as it uses mostly non-biological contexts and examples, the *SRT* functions as a highly conservative measure of students' ability to transfer their scientific reasoning to new situations.

How has the literature informed this study?

The measurement of scientific reasoning

The research literature has informed this study in multiple ways. Past work illustrates that scientific writing differs from other genres (preventing the ready adoption of already published rubrics) and that the measurement of scientific reasoning via writing is still a developing field. Past research by cognitive psychologists on the development of scientific expertise as well as investigations into the evaluative criteria used in the scientific community help to define scientific reasoning and identify broadly supported criteria for measuring its development in our students. Reference to the professional scientific community as well as past pedagogical research suggest that the criteria of *methodology, context, literature, significance, justification of conclusions* and *writing quality* are highly valued components of scientific expertise which are also measurable in science writing.

Past research has also strongly indicated that peer review is likely to be an effective pedagogical tool for stimulating scientific reasoning. Effective peer review and the collaboration it requires are real world skills and thus desirable learning outcomes as well as useful pedagogical strategies. In particular, peer review encompasses several of the strategies identified by Lajoie (2003) as accelerating the development of scientific reasoning expertise. Peer review is an authentic activity in the scientific community that provides multiple contrasting representations of the same task, collaboration among students with a range of abilities and individualized formative feedback. The multiple representations and formative feedback also both stimulate reflection, revision and metacognitive awareness which are necessary for meaningful learning and the development of expertise. Anecdotal and qualitative reports of students' comments suggest that students believe peer review improves

their engagement and reflection. Given the support for its use, the question then becomes, how is peer review best enacted in the classroom? Are there specific instructional scaffolds to improve student performance and enhance outcomes?

Peer review as a pedagogical tool: suggestions for implementation

Past work on students' perceptions of peer review suggest that they find it to have a positive impact overall (Haaga, 1993; Hanrahan & Isaacs, 2001; Stefani, 1994), but that students have concerns about the ability of their peers to assess them effectively (Cho, Schunn, & Wilson, 2006; Hanrahan & Isaacs, 2001). Students may perceive their peers comments as being less valuable or helpful than those of a subject matter expert such as an instructor. This perception is inaccurate however. When the average peer reviewers' scores correlate strongly with instructor scores($r = 0.89$, $n = 254$ students over 5 separate courses Cho, Schunn, & Wilson, 2006; $r = 0.62$ to 0.88, $n = 107$ students over three years Hafner & Hafner, 2003). Cho, Schunn and Wilson (2006) also calculated the reliability of *among* peer reviewers and found that any three to four reviewers had an effective reliability of $r = 0.55$ while using all six peer reviewers produced a reliability of $r = 0.78$ (95% confidence interval = 0.46 to 0.92). It should be noted that these correlations and reliabilities are significantly higher than those produced by professional referees (Cicchetti, 1991; Marsh & Ball, 1989; Marsh & Bell, 1981; Petty et al., 1999) or between graduate teaching assistants and faculty instructors (Kelly and Takao 2002 515).

Peer reviews are thus viewed as both valid and reliable from the standpoint of an instructor who can see the range of variation in paper quality across the whole course (and who has access to these statistics). Cho, Schunn and Wilson (2006) make the salient point however that student perceptions may differ because students cannot see the variation in student paper quality across the whole class. In 75% of the 16 courses at four institutions studied, the variation among raters on a single student's paper exceeded the variation in quality that that same student was exposed to as a reviewer (Cho, Schunn, & Wilson, 2006). Namely, the smaller subset of papers available to students combined with the relatively greater variation found in a small sample of reviewers skewed students' perceptions of the reliability of peer scores. Students should therefore be granted access to the instructor's viewpoint and these research data on peer reliability should be made an explicit part of instruction.

Cho, Schunn and Charney (2006) also conducted the only identified quantitative study on students' perceptions of the *usefulness* of feedback in a science class. They studied three classes (two undergraduate and one graduate). In the first undergraduate class (n = 28) students received blind feedback from either peers or a faculty member. While students raised in authoritarian educational systems may complain that their peers are unqualified to rate their work, undergraduates at this major US university could not distinguish between the helpfulness of comments provided by peers vs. those provided by a faculty member ($p = 0.36$, Cho, Schunn et al. 2006). The average usefulness scores were at least 4.0 (maximum point value was 5.0) regardless of rubric criterion or source indicating that students found the feedback useful regardless of the identity or expertise of the reviewer. Thus, future implementations of peer review in classrooms should provide explicit instructional background and/or research data to proactively address student concerns about the quality of peer feedback.

Whilst undergraduates did not perceive any differences in the *usefulness* of the feedback provided by faculty vs. peers, the *function* of the comments does vary based on the identity of the reviewer. These differences provide insight as to how undergraduate students and graduate teaching assistants should be guided in their development as reviewers. When the review comments from two undergraduate and one graduate class were coded as to whether or not they made constructive suggestions for change, the frequency of each comment type varied as a function of the reviewer's identity (Cho, Schunn et al. 2006). Faculty comments varied from graduate and undergraduate peer comments by being both nearly twice as long (and consequently containing nearly twice as many idea units) ($p < 0.001$) and also having the highest frequency of *directive* comments to any other type (3:1, $p < 0.005$, Cho, Schunn et al. 2006). Directive comments were defined as "suggesting a specific change particular to a writer's paper" (Table 1, p. 269) and could highlight either strengths or weaknesses. Undergraduate comments contained 70% more *praise* comments (positive comments lacking suggestions for change) than faculty. Graduate student comments had the highest frequency of *criticism* (negative comments lacking a suggestion for improvements) though criticism was relatively uncommon overall (Cho, Schunn et al. 2006). Cho, Schunn and Charney (2006) therefore recommend that instructors implementing peer review provide explicit

instruction and support to encourage undergraduates to be more directive (specific and suggest changes that would improve that paper in particular) and to encourage graduate teaching assistants to use more praise. It should be noted that the feedback provided by graduate students in this study were written for graduate peers however. Therefore, the tendency towards criticism may not be representative of the comments that graduate students would provide to undergraduates when they are teaching.

The findings from these studies indicate that when peer review is used in the classroom, it is critical that students be informed that peers *are* effective reviewers, as well as provided with support for *how* to further improve the quality of their feedback by being more directive.

Summary

The development of scientific reasoning skills in students is a complex and multi-layered process requiring spans of several years (Ericsson and Charness 1994; Zimmerman 2000). Science writing is an integral component of scientific reasoning or at least an important product produced by such reasoning. Peer review appears likely to accelerate the development of scientific reasoning and writing due to its collaborative, metacognitive, and comparative nature as well as the formative feedback it provides. Measuring the development of students' scientific reasoning skills is also a challenge. As scientific reasoning develops over time, measurement tools independent of assignment and course are necessary to track students' longitudinal progress. The development of a rubric based on attributes valued in the scientific community and applicable to a wide variety of science writing would provide many fruitful research opportunities. Not only could acceleration of students' scientific reasoning due to peer review or other instructional interventions be measured, but also questions concerning the explicit trajectory of how students develop expertise (which skills develop easily, which are more challenging) could be addressed. Lastly, triangulation using other metrics of scientific reasoning is necessary and information on students' perceptions of peer review would be useful for facilitating classroom implementation. Students' perceptions of the role, function and consequences of peer review in both the classroom and the scientific community are also relevant as they affect motivation, self-efficacy and transferability of reasoning skills.

CHAPTER 3
METHODOLOGY

Overview

The purpose of this chapter is to enable the reader to evaluate the methodologies employed in data collection to assess reliability and validity of the data from which conclusions are drawn in the discussion sections. This chapter is consequently organized by a general description of the research design (mixed methods) followed by a delineation of the research questions and a description of the components of the study that are consistent across all data types (such as the population of biology majors or the enactment of peer review). Next, the three major data sources and accompanying instruments are described: 1) the Universal Rubric for Laboratory Reports which when applied to student laboratory reports assesses student achievement in the area of scientific inquiry and critical thinking skills, 2) the *Scientific Reasoning Test* (Lawson *et al* 2000; Lawson 1978) and 3) the *Peer Review Survey* which elicits student perceptions of the process, purpose and impact of peer review. Sections on each of the data sources include a description of the instrument, the means of administering the instrument and data collection, followed by a description of the statistical analysis.

Research design

Multiple data sources and measurement types were used to assess the impact of peer review on students' scientific reasoning skills. In particular, an effort was made to incorporate both broad scale quantitative measures as well as more detailed qualitative perspectives to allow triangulation and increase confidence in conclusions (Mathison 1988; Johnson and Onwuegbuzie 2004). Specifically, three major types of measurements were made: 1) broad quantitative measures of scientific reasoning ability using cross-sectional cohorts of students to search for the overarching impact of peer review, 2) cross-sectional and longitudinal assessments of student scientific reasoning ability using laboratory reports and student writings as data sources, and 3) student perceptions of peer review using a survey tool. The inherently subjective nature of the laboratory report-based data was greatly reduced by using multiple independent raters and other methods of replication. The broad quantitative

assessment of students' abilities to reason scientifically was made using a pre-published multiple-choice instrument that was not biology specific. The use of both cross-sectional and longitudinal populations allows for further triangulation of results. Lastly, collecting student perceptions of the effect of peer review allows an additional level of insight not otherwise afforded as to whether or not students recognised the pedagogical aims and outcomes of the instructional innovation.

Research studies

The overarching topic of this project is divided into ten separate studies whose inter-relationships were illustrated in Figure 1.2. Firstly, prerequisite conditions and assumptions had to be tested. Study 1 investigates the degree to which students are capable of productively engaging in peer review – specifically, that the time and cognitive demands of the task are reasonable and that peer feedback can cause improvement in student writing. Study 2 tested whether the Universal Rubric produced consistent and reliable scores when implemented by trained raters. While it had been demonstrated reliable in other similar student populations, Study 3 confirmed the reliability of the *Scientific Reasoning Test* in this undergraduate population. Next the primary thrust of the research was to determine the impact of peer review on students' scientific reasoning abilities and how those abilities change over time as a result of peer review. Study 4 assessed changes in student scientific reasoning abilities in a cross-sectional sample and Study 5 was the same methodology using a longitudinal sample. The relationship between *Scientific Reasoning Test* scores and the number of peer review experiences in which students had engaged was investigated in Study 7. Studies 6 and 8 investigated the reliability of the Rubric when used by science graduate students under natural grading conditions and graduate students' perceptions of the utility of the Rubric as the Rubric could potentially be an effect pedagogical as well as research tool. Lastly, undergraduates' perceptions and understandings of the purpose, utility and impact of peer review in the classroom (Study 9) and the role of peer review in the scientific community (Study 10) were investigated because they would have a direct impact on student motivation and effort which would affect the achievement results from studies 4, 5 and 7.

Table 3.1. *Research Studies and Questions*

Study	Study Title	Research Question
1	Consistency and effectiveness of undergraduate peer reviewers	Can first year undergraduates enrolled in Introductory Biology be effective (consistent and useful) peer reviewers?
2	Reliability of the Universal Rubric for Laboratory Reports	Is the Universal Rubric a reliable metric of scientific reasoning and writing skills in this population across a variety of biology courses with graduate teaching assistants scorers?
3	Reliability of the *Scientific Reasoning Test*	Is the *Scientific Reasoning Test* a reliable metric in this population?
4	Student scientific reasoning skills in laboratory reports (cross-sectional sample)	To what degree do undergraduates evidence scientific reasoning skills in their laboratory reports and does their achievement vary by course?
5	Student scientific reasoning skills in laboratory reports (longitudinal sample)	To what degree do individual undergraduates evidence scientific reasoning skills in their laboratory reports and how do their skills change over time?
6	Reliability of scores given by graduate teaching assistants under natural conditions	How does the reliability and stringency of scores given by graduate teaching assistants vary with pedagogical training and support?
7	Relationships between *Scientific Reasoning Test* scores and peer review experience	Does peer review have a greater influence on students' *Scientific Reasoning Test* scores than academic maturity as measured by academic credit hours and institution type?
8	Graduate teaching assistants' perceptions of the utility of the Universal Rubric	How do graduate teaching assistants perceive the Universal Rubric as a pedagogical tool and would they advocate its use to others?
9	Undergraduate perceptions of the peer review process in the classroom	How do Introductory Biology students perceive the role of peer review in the classroom and its effects on them personally?
10	Undergraduate perceptions of the role of peer review in the scientific community	How do Introductory Biology students perceive the role of peer review in the scientific community and its effects on practicing scientists?

Note. See also Figure 1.2 (p.11 or 154) and Table 6.1 (p. 155) for overviews of the research design.

Study context

Study population

The university is a large (18,000 undergraduates, 8,600 graduate students) partially state-funded institution with approximately 1600 faculty, a medical school, law school and business school in addition to eleven undergraduate colleges. Ninety percent (90%) of the students are state residents, 82% of freshmen continue to their senior year and 62% graduate within six years. Classes are on a 14 week semester system with Fall terms beginning in late August and finishing in early December and Spring semesters begin in early January and finish in early May (www.sc.edu).

The population of biology majors has had relatively consistent demographics over the last five years (2002 to 2007, n=10,396 students for all five year averages). A notable majority of biology majors are women (62 $\pm$ 2% female) with an average age of 20 years. Most biology majors are Caucasian (63 $\pm$ 2%) or African-American (19 $\pm$ 2%); other ethnic groups ranged from less than 1% (Native American), 2% (Hispanic) to 7% (Asian). Eight percent of biology majors did not report an ethnic group. Categories of student race or ethnic origin used are those defined by the National Center for Educational Statistics (NCES) collected by the university as part of the admissions process. Ethnic categories are self-reported by the student. Only one racial code was recorded per student. For comparison, the overall student body population at this institution is has fewer women (54 $\pm$ 0.3% female) and slightly fewer African-Americans (71 $\pm$ 1% white and 15 $\pm$ 1% black) than the biology major for the same time period (n = 170,427 students). Thus, the biology major is populated by more women and more African-American students than the institution as a whole. Any positive outcomes from peer review as an instructional innovation may therefore be of interest to those concerned with underrepresented groups in science.

Student sample

Demographics for the courses from which the data were collected do not vary notably from the biology major patterns (61% of the biology majors were female and 63% of the total sample was white, 19% was black with single digit percentages for all other ethnic groups) but details are provided in each relevant section. The

courses from which study samples are drawn begin with the year-long sequence, Introduction to Biological Principles I and II (BIOL 101 and 102) which serves as the entry level course for biology majors. It should be noted that a large proportion (~55-65%) of the students enrolled in the introductory sequence (BIOL 101/102) are not biology majors, but belong to related health science fields (pharmacy, exercise science, students intending to apply to medical school but who are majoring in other fields). Thus, sample sizes vary for specific sub-populations depending on whether the measure was restricted to biology majors or utilized all students enrolled.

Subsequent to Introductory Biology, biology majors are required to enroll in three courses: BIOL 301 (Ecology and Evolution), BIOL 302 (Cell and Molecular Biology) and BIOL 303 (Genetics). BIOL 301 and 302 have corresponding optional laboratories that are quite popular with majors and from which samples for some portions of the study were drawn. The remainder of the Biology curriculum is composed of upper division courses of the student's choice (400, 500 and 600 level courses). Samples for this project were also taken from one upper division course BIOL 530 (Histology) which has a mandatory laboratory.

Software

Peer review was accomplished using *Calibrated Peer Review*™ (*CPR*) (http://cpr.molsci.ucla.edu) an online software program developed in the mid-1990s by Orville Chapman and other practicing scientists at the University of California Los Angeles as part of the National Science Foundation Molecular Science Project. Currently, several hundred institutions across the US use *CPR* and over 140,000 students' accounts existed in the system (Russel 2007). Contrary to those who have used *CPR* as a peer grading system, this research used peers predominantly for formative feedback and little if any portion of a student's grade was derived from points assigned by the *CPR* software. In this research, students were graded on their efficacy as reviewers during the peer review process and writers were encouraged to incorporate the formative feedback they received and improve their paper before turning a final version of the paper into the instructor for a grade.

All final papers were checked for plagiarism using the commercial software *Safe Assignment* (www.safeassignment.com) which ran through Blackboard©.

The peer review process.

The process of peer review was defined for the purposes of this study as students' exchanging written work and feedback via an online, web-based system that affords anonymity to both writer and reviewer (but which is transparent to the instructor and researcher). Each peer review process began with students participating in an open-ended research project. Projects were usually collaborative among pairs or groups of three to four students. Students then wrote their findings individually in a format similar to that used for science publications (Introduction, Methods, Results, Discussion, Literature Cited: hereafter termed a 'laboratory report.'). Students are provided at the onset of the project with the criteria and goals on which they will be judged both by the peer reviewers and the grading instructor. These criteria and goals come largely from the Universal Rubric (see below) with some assignment-specific modifications. Writers upload drafts of their written assignments with identifying information removed. If the instructor has included them, students read and score *calibration* papers (exemplars provided by the instructor) using the assignment criteria. The instructor has also scored these practice papers. The software then distributes each writer's paper to three peer reviewers. Reviewers are stimulated by written prompts in the online system (input by the instructor) that encourage them to focus their feedback on the given criteria that were backbone of the assignment.

Two forms of feedback are possible in the *CPR* system. Reviewers can rate student papers on a scale of 1-10 and provide other numerical ratings of the quality of a writer's work by clicking a rating choice for each criterion or if the instructor has set the reviewing prompts to include open-ended text boxes, they can write detailed comments and explanations to the writer. For example, in response to a criterion prompt such as "Are the writer's conclusions based on the data?" students could respond by clicking either "yes or no." If the instructor included a text box for the criterion, the reviewer could also provide a justification or explanation of how well the writer met that criterion. The *CPR* software tracks those numerical scores and flags reviewers whose numerical evaluations deviate more than one standard deviation from other peer reviewers responding to that same paper.

Once the deadline for peers to provide feedback has passed, these numerical and written pieces of feedback are then made available to writers online. Writers are

encouraged to use the feedback to improve their paper prior to handing it in to be graded by the instructor.

Instructions given to students for producing useful feedback

Two forms of accountability exist to encourage students to provide useful feedback. The *CPR* system compares the numerical ratings made by reviewers and assigns a reviewer competency index score based on how closely aligned the reviewers are to one another. For students enrolled in BIOL 101 a large proportion of their peer review grade was based on this consistency rating, the score reviewers give their papers and how closely aligned the student's self-assessment was to the reviewers' assessment. In BIOL 102, these numerical ratings comprised little to no proportion of the student's grade. Instead, the emphasis was on the quality of the written feedback comments. In BIOL 102, graduate teaching assistants randomly selected one review written by each student and assigned points based on the quality of the comments.

For both courses, students were provided with a set of instructions explaining the quality of useful feedback. The instructions included the following definition of useful feedback as well as reminders to be respectful and professional in the written comments they provide to peers. Students received the following information in class and again within the online *CPR* system just prior to reviewing peer papers.

***Useful feedback*:**

- *is specific and concrete,*
- *focuses on the quality of the author's argument (are conclusions logical and well supported by the evidence/data?) rather than on mechanics of writing such as grammar or spelling,*
- *identifies assumptions behind or consequences of author's ideas which the author has not explicitly discussed and*
- *would likely result in meaningful new content being added to or revised in the paper.*

For the terms included in this study, in BIOL 102, the *CPR* process was also preceded by an in-class exercise on how to produce useful feedback. Students were given examples of feedback and asked to score them as *useful, partially useful*, or

not useful. A class discussion followed concerning what were appropriate scores for each feedback example. A handout summarizing this exercise was provided to students for their convenience and use (Table 3.2 and Appendix 2).

Table 3.2. *Examples From Handout Provided to Students to Encourage Them to be Effective Reviewers.*

Feedback item	**Useful?**	**How to improve the feedback**
1. Your paper is GREAT! How did you come up with your idea?	No	Provides no actual information to the writer on HOW to improve the paper
2. At the end of paragraph 2, you say you think this was a sex-linked cross. Is this your hypothesis? What traits do you think the parents had? Why do you think this is the best explanation?	Yes	Full of detail about where and why the reviewer was lost and if the writer answers the reviewer's questions, the paper will have a clearer statement of the hypothesis, a consideration of alternative explanations and logical connection between hypotheses, data and conclusions.
3. Your argument makes no sense. What is your evidence?	Partially	Asking for evidence is useful, but reviewer does not indicate which part of the paper is confusing them or what exactly they didn't understand.
4. Your argument depends on weight being an inherited trait. What evidence do you have to support this assumption?	Yes	The reviewer has identified an assumption made by the writer and pointed out how the validity or invalidity of this assumption could impact the writer's conclusion.
5. Which of your hypotheses is best supported by the data?	Partially	The reviewer is specific in indicating that the writer did something well (posed multiple explanations) and indicates that no clear conclusion was made but without specifying how or where they felt the writer's conclusions were lacking.

Note: See Appendix 2 for full Handout.

Enactment details of peer review in specific courses

Students were supported in their development as effective reviewers through gradual increase in expectations and repeated exposures to the peer review process. A transitioning emphasis from the rote procedures of peer review to the quality of feedback was employed. The laboratory portion of the year-long introductory biology sequence highlighted peer review as a central skill and student learning was coordinated across the two courses. The peer review process was begun in our

curriculum in Spring 2002 and has been enacted using *Calibrated Peer Review* every semester in the introductory biology courses (BIOL 101 and 102) since Fall 2003.

First semester Introductory Biology (BIOL 101). In the first semester course, students were first exposed to the procedures and purpose of peer review using a relatively intellectually unchallenging assignment: write an introductory paragraph for a hypothetical laboratory report on a recently completed laboratory experiment and provide feedback to their peers using the *CPR* website. The purpose of this assignment was to allow students to focus on the mechanisms and procedures of peer review without undue worry about the nature and extent of the writing or feedback they were providing. Students were also asked to gradually build skills in writing all aspects of a laboratory report over the course of the semester. Namely, after the introductory paragraph, they next write just the methods section for a subsequent laboratory activity, the results section for an activity after that, etc. The culminating experiment at the end of the semester was a *Drosophila* (fruit fly) genetics experiment in which students had to determine the mode of inheritance of an unknown phenotypic trait. For this experiment, students were asked to write a full laboratory report and provide peer review feedback using the *CPR* system. In this course, a minor portion (<5%) of students' laboratory grades were affected by their ability to successfully complete the peer review and give assessments which were consistent with (within one standard deviation) other peer assigned to the same paper. This is the assignment on which peer review occurred each semester for the BIOL 101 course.

Second semester Introductory Biology (BIOL 102). In each iteration of the second semester course students were provided with an educational dataset, *Galapagos Finches* (Reiser, Smith et al. 2001; Reiser, Tabak et al. 2003), derived from real datasets collected by Rosemary and Peter Grant in the early 1970s (see Grant and Grant 2002). Students are told that a mass mortality event occurred on the island of Daphne Major and are asked to determine the cause and if evolution occurred in the finch population as a result. As it is a real ecological dataset collected for other purposes, there are a variety of defensible conclusions and interpretations as well as irrelevant portions to the data. Students pose their own hypotheses, locate, analyze and interpret relevant data and therefore must argue and justify their data selection decisions and conclusions. Written reports are then uploaded to the *CPR* system. In this round of peer review, very few of the points

associated the peer review assignment were earned by successfully navigating the software (the exact number of points varied from semester to semester, but were < 1% of each courses' total), but instead were focused on the quality of feedback and writing produced. Indeed, reviewers were now graded on the quality of the feedback they provide. Using the "useful/partially useful/ not useful" schema indicated previously (column 2 of Table 3.2 as well as detailed in Appendix 2), instructors randomly chose a single review written by each reviewer and grade the quality of the feedback as a full point, a half point or no points respectively. Providing ten useful pieces of feedback in a single review earned 100% of the (10) points possible. Instructors were science graduate students hired as teaching assistants. Students were allowed to write as many pieces of feedback as they desired per review to earn the 10 points.

Ecology and Evolution Laboratory (BIOL 301 L). In Spring 2005, peer review was also incorporated in the BIOL 301 laboratory courses and continued in subsequent semesters thus providing students a third opportunity to engage in the process. Uploading of final papers via *SafeAssignment* began in Fall 2005 in this class as the University did not make *SafeAssignment* available in the Spring 2005 semester. BIOL 301 lecture is a required course for all majors. BIOL 301 laboratory is an optional (but popular) laboratory for biology majors. Many transfer students bring in credit for introductory biology and thus enter into the biology curriculum at this level. In BIOL 301 Laboratory, students engage in peer review 2-3 times per semester as there are three experiment-based portions of the laboratory that result in written laboratory reports. In some instances, peer review also occurred in other several upper division courses, but not in any systematically reportable way. Portfolios including an upper division course in addition to the 301 L and introductory biology courses can be constructed for a handful of students. The intent is that students should encounter peer review each time they are asked to do an experiment and subsequently write it up, but coordination among the diverse faculty members in the department who teach the upper division courses has been sporadic. For the purposes of this study, a sufficient number of students experienced peer review multiple times (up to 3) for an effect to be discerned. It is expected that greater effects will be seen in future years as a greater proportion of students have three or more experiences with peer review.

Other courses sampled (BIOL 302 Laboratory and Histology) did not engage in peer review during the semesters when data collection occurred. These samples were used to include additional students at later stages in their academic career who had participated in one of the three aforementioned courses.

Data sources and instruments

The effect of peer review on students was determined by three major data sources: 1) students' performance on the *Scientific Reasoning Test* (*SRT*), 2) student performance in laboratory reports collected from the three described courses (BIOL 101, 102 and 301L) and 3) student perceptions of the peer review process collected in an anonymous online Survey. Additionally, two instruments were developed to assist with the data collection. Firstly, in order to compare student performance in laboratory reports across multiple courses, a Universal Rubric for Laboratory Reports was developed and its reliability as a measurement tool is investigated. Secondly, a Survey was constructed to measure students' beliefs and perceptions of the usefulness and value of the peer review experience.

Universal Rubric for Laboratory Reports

Instrument development

Rubric criteria were derived from biology department's curriculum goals and therefore intended to be independent of any particular content area within biology (e.g. Trevisan, Davis et al. 1999). The curriculum goals and subsequent criteria were derived through a series of discussions with colleagues on the Departmental Curriculum Committee. Members of the committee were the principal authors of the Department's goals. This researcher encapsulated those discussions and used them and the written goals to define an initial set of 15 criteria. The desired performance at the high end of the scale was also based on those discussions. The low end of the performance scale was based on this researchers' personal experience with struggling freshman. The interim performance levels were developed according to instructor experience and a desire for an internally consistent and parallel range of performances. The end result was a four-level scale ranging from *not addressed* which included behaviors often observed in first semester freshman, through *novice* and *intermediate* and culminating at *proficient*. The *proficient* level of performance

was conceptualized as the level of performance expected from a top-ranking undergraduate or beginning graduate student.

Preliminary testing occurred by incorporating the rubric criteria into assignments given in this researchers' own courses (BIOL 102 laboratory at that time). This researcher also continued to share and discuss the criteria and performance levels with a wide variety of science faculty, graduate students and educational researchers both within and outside the institution over an 18 month period. At the end of that period of recursive review and revision, the criteria and performance levels were piloted on laboratory reports from courses taught by faculty other than this researcher. Nineteen biology graduate teaching assistants from a variety of biological sub-fields were asked to apply the criteria and performance levels to a variety of actual student papers and provide explicit written feedback on the relevance and usefulness of each criterion and performance level in a single sit down session. Criteria definitions and performance level descriptions were subsequently revised again. This level of review, discussion, testing and revision either meets or exceeds that currently described for other published rubrics (Hafner & Hafner, 2003; Halonen et al., 2003; Trevisan et al., 1999).

Final rubric description

Rubric criteria were structured around the foundational components of professional scientific writing: introduction, methods, results, and discussion. To assist students, additional explicit criteria were created to focus on hypothesis quality, data use and presentation, statistical competency, use and understanding of primary literature, significance of research and writing quality (Table 3.3).

Table 3.3 *Universal Rubric Criteria Codes and Definitions.*

Criteria	**Code**	**Definition**
Introduction		
Context	I:C	Demonstrates a clear understanding of the big picture; Why is this question important/ interesting in the field of biology?
Accuracy	I:A	Content knowledge is accurate, relevant and provides appropriate background including defining critical terms.
Hypotheses		
Testable	H:T	Hypotheses are clearly stated, testable and consider plausible alternative explanations
Scientific merit	H:S	Hypotheses have scientific merit.
Methods		
Controls and replication	M:C	Appropriate controls (including appropriate replication) are present and explained.
Experimental design	M:E	Experimental design is likely to produce salient and fruitful results (actually tests the hypotheses posed.)
Results		
Data selection	R:S	Data chosen are comprehensive, accurate and relevant.
Data presentation	R:P	Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete.
Statistical analysis	R:St	Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained.
Discussion		
Conclusions based on data selected	D:C	Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.
Alternative explanations	D:A	Alternative explanations (hypotheses) are considered and clearly eliminated by data in a persuasive discussion.
Limitations of design	D:L	Limitations of the data and/or experimental design and corresponding implications for data interpretation are discussed.
Significance of research	D:S	Paper gives a clear indication of the significance and direction of the research in the future.
Primary Literature	PL	Writer provides a relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided) using primary literature. Primary literature is defined as: peer reviewed, reports original data (not a review), authors are the people who collected the data, and a non-commercial scientific association publishes the journal.
Writing Quality	WQ	Grammar, word usage and organization facilitate the reader's understanding of the paper.

In addition to the criteria, performance levels were described for each criterion to comprise a rubric. The full final version of the Universal Rubric for

Laboratory Reports is attached as Appendix 3. An example of a single criterion showing the four performance levels is given in Table 3.4.

Table 3.4. *Example of a Universal Rubric Criterion (Hypotheses: Testable and Consider Alternatives, H:T) and Corresponding Performance Levels.*

Criterion	***Performance Levels***			
	Not addressed	*Novice*	*Intermediate*	*Proficient*
Hypotheses are clearly stated, testable and consider plausible alternative explanations	• None indicated. • The hypothesis is stated but too vague or confused for its value to be determined • A clearly stated, but not testable hypothesis is provided. • A clearly stated and testable, but trivial hypothesis is provided.	• A single relevant, testable hypothesis is clearly stated • The hypothesis may be compared with a "null" alternative that is usually just the absence of the expected result.	• Multiple relevant, testable hypotheses are clearly stated. • Hypotheses address more than one major potential mechanism, explanation or factors for the topic.	• A comprehensive suite of testable hypotheses are clearly stated which, when tested, will distinguish among multiple major factors or potential explanations for the phenomena at hand.

Source of student papers

Student papers were selected from three different university biology laboratory courses to represent student performance at the freshman and sophomore levels. These courses included the first and second semesters of the introductory biology course sequence for majors (BIOL 101 and 102) as well as from the laboratory on Ecology and Evolution associated with a required majors course (BIOL 301) intended to be taken by sophomores. Similar to the overall major demographics, course demographics were predominately female (60-64%) and the top two dominant ethnic groups were Caucasian (55-70%) and black (13-24%) regardless of course.

The assignment details that generated the student papers are presented in Table 3.5.

Table 3.5. *Descriptions of Assignments Used to Generate Student Papers for Rubric Reliability Study*

Course (term)	Description of assignment	# papers selected
BIOL 101 (Fall 04)	Genetics: Determine the Mendelian inheritance pattern of an unknown phenotypic trait in fruit flies (*Drosophila melanogaster*) based on data collected from a live cross.	49
BIOL 102 (Fall 04)	Evolution: Determine whether or not evolution occurred in a population of birds as the result of a drought using a pre-existing multi-year dataset (*Galapagos Finches*) (Reiser, Tabak et al. 2003).	45
BIOL 301 (Fall 05)	Ecology: Determine whether shade/sun affects the abundance and distribution pattern of dandelions in a field.[1]	48

Note. This 301 assignment was *not* peer reviewed in this term. It was the only assignment completed in the course by the time of the rubric reliability study. Peer review occurred for subsequent assignments in this class.

BIOL 101 and 102 papers selected from those written in Fall 2004 and BIOL 301 were selected from those written in Fall 2005. From each of the three classes (BIOL 101, 102 and 301 laboratory), a subset of 45 to 50 papers were selected based on the following criteria:

1) paper and graphs were complete, on topic and without plagiarism;
2) paper was authored by a biology major who was still enrolled in the biology program at the time of selection;
3) no more than 5 papers were selected from any one laboratory section (maximum enrollment of 24 students per section, 33 sections total sampled) and
4) within each section at least one paper was selected from a student who earned an "A" in the course and at least one from a student who earned a "D." Efforts were made to select papers representing the available spectrum of quality (as determined by course grade) within each laboratory section.

Selected papers were then stripped of all author-identifying information, assigned an anonymous ID code and standardized for font type and size, margins, and line spacing before printing.

Graduate student raters

The papers were scored by two independent groups of raters. Raters were drawn from the biology graduate students, most of who had served as teaching assistants in the relevant classes and regularly graded laboratory reports from those courses. One group was trained for a formal reliability test of the rubric and one group remained untrained to assess the usefulness of the rubric under more natural conditions. Graduate teaching assistants are ubiquitous as laboratory instructors in Research 1 institutions (Golde 2001) and while large research institutions comprise only three percent (3%) of all higher education institutions in the US, they produce 32% of the nation's undergraduate degrees (Boyer Commission, 1998). Graduate teaching assistants thus have a significant impact on undergraduate education and to this researcher's knowledge, no previous measure of their natural grading consistency has ever been reported.

The two sets of raters were divided among treatment and naturalized conditions to ensure the most similar distribution of experience possible (Table 3.6). Both sets of raters received identical sets of student laboratory reports, record sheets for recording scores, a copy of the assignment that was given to the undergraduate students who wrote the papers and verbal instructions on the purpose of the project and their role in it as well as monetary compensation for their time and effort.

Trained Raters. Nine raters received five hours of training on how to score using the Universal Rubric (henceforth referred to as "trained" raters). Trained raters were provided with a Scoring Guide version of the rubric (Appendix 4). In the Scoring Guide, each criterion was followed by examples of student work at various performance levels. Training was facilitated by Dr. Robert Johnson, a specialist in rubric design and assessment (University South Carolina, College of Education) and began with a whole-group discussion of the rubric rational and intent (3 hours). Raters then broke into their assigned teams and individually scored their three example papers for that course. Discussion within the teams occurred until consensus scores were reached for each criterion in each exemplar paper (2 hours).

Scoring of all papers by trained raters occurred within 24 hrs after training and under supervised conditions. Trained scorers kept track of the time they spent in scoring and worked an average of 7.8 hours averaging 6.2 papers per hour. The maximum duration of scoring was 8.5 hours and the minimum was 6.8 h. Scoring occurred in blocks of several hours with breaks for meals and sleep.

Table 3.6. *Gender and Experience Levels of Graduate Student Raters*

Course	Rater Type	Raters' genders	# semesters of teaching experience per rater	Raters have taught this assignment?
101	Trained	F, F, M	2, 2, 5	Y, Y, Y
101	Natural	F, F, F	1, 2, 4	Y, Y, Y
102	Trained	F, F, F	3, 3, 11	Y, Y, N
102	Natural	F, F, F	1, 4, 4	Y, Y, Y
301	Trained	F, F, M	3, 7, 7	Y, N, Y
301	Natural	M, M	3, 7	Y, Y

Note. Codes are as follows: (F) female, (M) male, (Y) yes, (N) no. Data are portrayed respectively. Namely, for "101/Trained," the first and second raters were both females (F,F), with 2 semesters of teaching experience (2,2) and had taught the *Drosophila* assignment in BIOL 101 before (Y,Y) while the third rater was male with 5 semesters of teaching experience who had also taught this particular assignment in the past (M,5,Y). The average number of semesters of teaching experience for trained raters was 4.8 $\pm$ 3.0 and for natural raters was 3.4 $\pm$ 1.9.

Natural Raters. Eight additional graduate students were hired to assess grading consistency of teaching assistants in a natural condition and hereafter will be referred to as *natural* raters. At the time that the rubric reliability study was conducted, upper division teaching assistants often received no pedagogical support for student written assignments. Graduate teaching assistants in the 300 level laboratories had occasional verbal instructions from the supervising faculty member regarding assignment details, but rarely received instructional materials (e.g. no criteria, rubrics, etc.). Introductory biology graduate students were provided with assignments and criteria and met regularly with faculty on pedagogical issues. Thus, to provide a standardized, but natural level of support, the comparison group of raters received a ten-minute verbal explanation of the assignment and a single page

list of criteria. Specifically, for the purposes of this study, the natural raters were provided with the same list of criteria as are found on the rubric, but without the performance level descriptions or examples of student work (Appendix 5). The same point scale was used (maximum 3 points per criterion) by the natural raters as the trained raters. Natural raters scored papers over the course of week, on their own time, in locations of their choosing so as to most accurately match normal grading conditions.

Data analysis

Rubric reliability data were analysed using generalizability (g) analysis (Crick and Brennan 1984) which determines the portion of the variation in scores which is attributable to actual differences in the quality of the papers rather than variation attributable to less relevant sources such as variation among raters, or assignments or variation due to interaction among factors such as student-assignment or rater-assignment (Shavelson and Webb 1991). For example, a generalizability score of 1.0 therefore means that all the variation in scores between papers was due to differences in quality among the papers and that no error was introduced (all raters were perfectly consistent regardless of student, assignment, etc.). In contrast, a generalizability score of 0.0 means that none of the variation in scores among papers was attributable to actual differences in quality and that all the variation in scores was entirely due to other sources of variation such as rater inconsistency instead.

Test of Scientific Reasoning

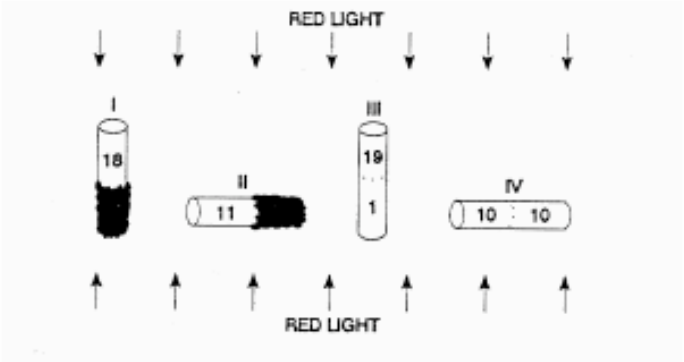
The instrument selected for quantitative measurements of students' scientific reasoning ability was the 2000 version of the *Scientific Reasoning Test (SRT)* (Lawson, 1978; Lawson, Alkhoury, Benford, Clark, & Falconer, 2000) (attached here as Appendix 6) because it had been previously validated as a reliable instrument in a similar context (university freshman), and it does not suppose any previous knowledge of biology, thereby avoiding a source of bias when administered to biology majors at different stages in the curriculum. In fact, most of the questions addressed students' reasoning ability and knowledge of experimental design principles using physical rather than biological contexts. The 2000 version of the instrument has 24 multiple choice questions designed in a two-tiered fashion so that

the second question in each pair asks students to identify the rationale behind their selection for the first question in the pair. The rationale choices used in the second tier questions are based on previously identified student misconceptions (Lawson et al., 2000). Eight of 12 pairs of questions use physical science concepts such as density, conservation of mass, pendulum swinging etc, as the context for questions. Types of reasoning required of students are those "associated with hypothesis testing (i.e., the identification and control of variables, correlational reasoning, probabilistic reasoning, proportional reasoning, and combinatorial reasoning)." (Lawson et al., 2000, p. 1001). In particular, four questions ask students to identify what results would falsify a hypothesis which is a primary component of scientific reasoning in our view. One of the few test items that does have a biological context is illustrated in Figure 3.1.

Administration

The *Scientific Reasoning Test* was administered to a cross-sectional cohort of students in the Fall of 2005 who were enrolled in the five classes mentioned above (BIOL 101, 102, 301L, 302L, 530; total enrollment = 942). Data from non-biology majors were winnowed out as the biology majors are the population of interest for this study. Further details on sample reduction decision are provided in the results section. Non-majors comprise more than half (59%) of the students in the introductory level and less than 10% of the students in the upper division courses. The test was administered in the laboratory portion of each class within three weeks of the beginning of the semester. The purpose and use of the data were explained to the students and an information sheet with contact information for the researcher was provided to the students. Completing the test was voluntary and there was no impact on the students' course grade as a result of their score on the test. In addition to the previously published *Scientific Reasoning Test*, students were asked to self report on the number of peer review experiences they have had in past courses, as well as their gender, ethnicity and prior biology background. For students who were enrolled in BIOL 102 or higher, they were also asked where they took introductory biology (within our program or elsewhere) and if they remembered engaging in peer review as part of the course.

11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



This experiment shows that flies respond to (respond means move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

12. *because*

a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
b. most flies did not go to the bottom of Tubes I and III.
c. the flies need light to see and must fly against gravity.
d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
e. some flies are in both ends of each tube.

Figure 3.1. Example of a tiered pair of questions from the *Scientific Reasoning Test* (Lawson et al., 2000) with a biological context. Correct answers are 11 (B) and 12 (A).

Data analysis

It is expected that *Scientific Reasoning Test* scores should increase with academic maturity; as a student progresses from freshman to senior year, their scientific reasoning score would be expected to increase simply as a result of their overall course work and greater academic experience. The effect of peer review on students' Scientific Reasoning Score was therefore compared to the effect seen merely from increasing academic experience as measured by credit hours. Additionally, as many of the biology majors transfer in after freshman year (where the major of the peer review experiences occur in our curriculum currently), there

could be a greater effect of academic maturity based on where students earned the credit hours currently on their record. Students were categorised into standard class years by credit hours earned (0-30 credit hours = freshman, 31-60 credit hours = sophomore, 61-90 credit hours = junior, 91-120 credit hours = senior, greater than 120 credit hours = fifth year). USC credit hours and credit earned elsewhere were tracked separately. ANOVA tests were used to test for significant differences among credit hours earned, type of credit hours (USC vs. transfer credit) number of peer review experiences and *SRT* score.

Survey of student perceptions regarding peer review

Overview

A Survey on students' perceptions of the process of peer review, its role in the classroom as a learning tool and its function in the scientific community was developed and administered in the introductory biology course during the 2006-2007 academic year. The instrument was titled *The Peer Review Survey* and hereafter will be referred to as the Survey. Each administration occurred at the end of the semester following the peer review process. The Survey focused on three major issues: 1) students' understanding of the purpose of peer review within and outside the classroom, 2) students' understanding of the mechanics of peer review, their understanding of function of each of the steps in the process, and their opinion of the quality of instructional support provided to help them engage in the peer review process, and 3) the perceived impact of peer review on their writing and critical thinking within and beyond the course. The Survey was developed with review and input from both the faculty responsible for the introductory course as well as from the Office of Program Evaluation, USC College of Education.

Fall 2006 Survey structure

The first semester, the Survey consisted of 23 statements to which students were asked to respond using a Likert scale with 6 values ranging from 'strongly disagree' to 'strongly agree.' An even number of Likert choices was intentional to force students to indicate a negative or positive reaction. Six questions focused on the purpose of peer review; ten focused on the mechanics of the peer review process and the effectiveness of the instructional supports; seven inquired about the impact of peer review on students' writing and critical thinking skills. Three open-ended

questions then followed with a fourth option for additional comments (See Student version Fall 2006, Appendix [7]). The open-ended questions were as follows:

- *Please describe why you think we asked you to use peer review in this class.*
- *Please describe how you think practicing scientists use peer review in their work.*
- *What changes would you recommend to improve peer review in this class?*

Student open-ended responses were collected and reviewed recursively until regular themes emerged which were defined into categories of responses. Responses were then categorised with more detailed sub-themes identified as appropriate. Once the coding scheme appeared stable because all responses fit into an existing category and sub-theme, the number of responses in each category was tabulated (Anderson 2002).

Fall 2006 Administration

The Survey was made available to the 562 students still enrolled in the BIOL 101 course by the end of the Fall 2006 semester (November) and 444 students responded (response rate of 79%). The Survey was made available by use of an online Survey tool (*Survey Monkey*) and student responses were anonymous. Students were encouraged to participate by use of a single bonus point for completing the Survey (<1% of overall laboratory grade) and bonus points were awarded when students emailed to their instructor a "secret code" that was revealed to them at the end of the Survey. This online Survey tool and same reward system were used in Spring 2007 (see Table 3.7 for samples sizes and details).

Revision and expansion of Survey and Spring 2007 administration

After the Fall 2006 semester, student responses to the three open-ended queries described above were reviewed and categorized. For the Spring 2007 administration, "select the top 3 reasons" items replaced these open-ended questions. In the "select the top 3 reasons" format, each choice that could be selected was derived from one of the categories of response that emerged from the Fall 2006 data collection. Students were asked to choose their top three responses from the resulting list. The refined student version (See Student Peer Review Survey Spring 2007 in

Appendix [8]) was re-administered in the Spring of 2007 to students in both BIOL 101 and BIOL 102 (n = 638).

For the Spring 2007 administration, it was also expected that most of the students in BIOL 102 would have already participated in peer review because they were enrolled in BIOL 101 the previous semester, so two new items were added. Students were asked if the peer review process was 1) more or less difficult and 2) more or less useful the second time. In addition, four more demographic triangulation questions were posed to students in the Spring 2007 administration:

- *Are you a biology major?*
- *Did you participate in peer review in BIOL 101 at USC last fall (2006)?*
- *Did you fill out a similar Survey in Fall 2006?*
- *Including the current semester, how many semesters have you used peer review for biology classes?*

The revised Survey was then made available to all BIOL 101 and 102 students enrolled as of April 2007. An "Additional Comments" open-text box was also part of the Spring version of the Survey in case students wished to add any additional information or insight.

Table 3.7. *Sample Sizes and Response Rates for Student Peer Review Surveys.*

Term	Course	#students enrolled	Response rate #	Response rate %	#with 2nd PR experience	% biology majors
Fall 2006	101	562	444	79%	n/a	Not asked
Spr 2007	101	230	206	90%	15 (7%)*	42%
Spr 2007	102	408	376	92%	312 (84%)	38%
Totals/Average		**1200**	**1026**	**85.5%**		

Note. Survey administrations occurred at the end of each semester. PR = Peer Review. The students in BIOL 101, Spring 2007 who are reporting this to be their 2nd peer review experience are presumably students who failed to pass BIOL 101 in Fall 2006. Sample sizes for each item on the Survey vary slightly as not all students answered all questions, but variation is less than 2% of the total relevant number of respondents (e.g. for BIOL 102, individual item sample sizes ranged from 369 to 372.)

Ethics compliance

This work was conducted in accordance with the principles and philosophies of the Australian Code for the Responsible Conduct of Research and the policies and procedures of Curtin University of Technology in particular. All possible care was taken to maintain the highest possible standards of academic integrity and honesty as well as to respect the welfare, rights and privacy of all people involved in the research. Specific ethics review and compliance were also conducted by the institution responsible for the students who comprised the study population.

An Institutional Review Board (IRB) application for Approval of Human Subjects Research at the University of South Carolina was made in 2003 and this work received a designation of "Exempt." Therefore, written consent by subjects was not required. This work was deemed exempt because the "research [was] conducted in established or commonly accepted educational settings, involving normal educational practices, such as: (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods [or was] research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior." (USC Office of Research Compliance, http://www.orc.research.sc.edu/). This designation satisfies federal (US) regulations and National Science Foundation requirements for protecting human subjects.

The data generated by this investigation were based on artifacts from assignments that are part of the normal biology laboratory curricula. Artifacts included draft and final versions of student assignments and the corresponding peer reviews as well as results of the multiple choice scientific reasoning test and anonymous survey results. The peer review process was conducted using a software program that guaranteed anonymity from the student's perspective, but allowed the instructor to track and identify authors. The objective tests were administered using scantron sheets. All student assignments and artifacts are handled with the level of privacy and confidentiality required by the Family Educational Rights and Privacy Act (FERPA) of 1974 (http://registrar.sc.edu/html/ferpa/ferpa1.stm). All data were reported in aggregate forms to ensure anonymity. No data were generated which

would required the use of a pseudonym or consent as all quotes included were collected anonymously.

As an additional demonstration of respect for students' interest and welfare, students were informed by means of the course syllabi or as a direct handout, that their regular course work, or any voluntary measures of achievement (e.g. the *Scientific Reasoning Test*) might be collected and used in an aggregated or otherwise anonymous manner for the purpose of evaluating the effectiveness of teaching methods and curriculum (including peer review). Students were given contact information for this researcher as well as contact information for the Office of Research Compliance and encouraged to contact either if they had concerns or questions. Laboratory reports or *Scientific Reasoning Test (SRT)* measures of student performance were either natural components of the coursework regularly assigned for the class, or voluntary (participation or performance had no consequences or impact on the student's grade in the course). Specifically, cross-sectional *Scientific Reasoning Test* data were collected in two ways. In Introductory Biology, the *Scientific Reasoning Test* was incorporated as a minor component of the regular course grade (<1% of the total points) under the heading of participation. In the 300 and 500 level classes, the *SRT* was taken by students on a voluntary basis. Verbal and written information was provided to student volunteers describing the intended usage of the information, that participation or lack of participation would not affect their grade in any way and contact information should they have any questions or concerns later. All students opted to participate. In the case of the Surveys of student perception of the process of peer review, participation was encouraged by the use of a single bonus point (<0.5% of course grade). Response rates for the Survey administrations are reported in Chapter 4, but in general ranged from 80-90%.

Thus, all data sources were either a collection of pre-existing information (e.g. laboratory reports) re-analyzed outside of the context of the course or had zero to negligible influence on a student's grade in the course.

Limitations of the study

As with all research, conclusions were limited by the types and nature of the data collected. Scientific reasoning encompasses a broad category of skills and

abilities that can be demonstrated in a variety of ways. Students' written laboratory reports, while a rich source of data, inherently miss some aspects of reasoning which would be more clearly seen in students' small group discussions, by direct questioning or observation of students in the laboratory directly. In particular, using only written data sources selects for students who are able to articulate their scientific reasoning onto paper. There likely are students within this sample who have scientific reasoning skills and gains that were not evident because those students had more difficulty expressing themselves in writing than verbally or by action and decision. Nonetheless, communication of scientific ability and discoveries via written reports is ubiquitous and a primary means by which scientists share their findings making student written laboratory reports not only an appropriate data source, but also an effective pedagogical tool.

Use of multiple-choice tests suffers from similar confounding factors in that students may have strong reasoning abilities, but poor test taking abilities. Consequently, the results of the *Scientific Reasoning Test* are used only to assess broad-scale patterns without speaking to the ability of any individual or small subgroup of students.

Lastly, survey data must always be treated as the self-report data that they are. Namely, such perceptional data are only useful when it is perceptions that are informative and of interest. Namely, students perceive that peer review improves their reasoning skills, but that perception has little bearing on their actual achievement which must be measured separately. Here, students' perceptions of the pedagogical rationale and outcomes of peer review were of interest and so the use of a survey tool was appropriate.

Summary

The effect of an instructional strategy on student learning is always multi-faceted and complex as human being are dynamic creatures affected by internal factors (motivation, affect, self-efficacy, interest, etc.) as well as external factors (classroom environment, peer interactions, etc.) beyond the direct instruction. Thus, the effects of instructional strategies must be measured within the context of interest; achievement of absolute skills, performance on standardized measures, normative performance relative to others etc. It is with this perspective that the effect of peer

review on scientific reasoning was selected. Contextually appropriate data sources (written laboratory reports, rubrics focused on written demonstration of scientific skills) were chosen or developed and a continuous focus on real-world scientific skills and perceptions of how the classroom relates to scientific community were maintained.

CHAPTER 4
RESULTS FROM ACHIEVEMENT DATA

Overview

Conclusions about the effectiveness of peer review are drawn from three major data sources: student science writing in the form of laboratory reports, scores on a multiple choice *Scientific Reasoning Test* and student responses to an online Survey. The relationships among these three data sources and the research questions are illustrated in Figure 1.2. Results of studies on student achievement in scientific reasoning using laboratory reports and the *Scientific Reasoning Test* are reported here. Undergraduate perceptions of peer review are presented in Chapter 5. Use of both in-depth, proximal data sources such as written laboratory reports and amore distal measures of student performance such as the *Scientific Reasoning Test,* allowed for triangulation of results and provides greater confidence in the conclusions.

Before investigating student performance, however, some basic assumptions concerning the instructional innovation of peer review are investigated. Students' ability to produce useful feedback and whether or not writers can effectively implement that feedback to improve their papers were investigated in a small-scale study (Study 1). Once it was determined that even the least experienced students could productively engage in peer review, the effect of peer review on students' scientific reasoning abilities was then investigated. Laboratory reports provide an in-depth look at student scientific reasoning ability in a format similar to that used by practicing scientists and thereby constitute an authentic performance assessment likely to elicit relevant scientific reasoning skills of interest (National Research Council Committee on the Foundations of Assessment 2001)

The decision to use laboratory reports from a variety of courses necessitated the development of a common metric that could measure student achievement of scientific reasoning skills regardless of course content or level. Thus, the Universal Rubric for Laboratory Reports was conceived and developed. Before the Universal Rubric could be used to measure student achievement however, its reliability as a measurement instrument needed to be demonstrated (Study 2). The Universal

Rubric was tested using laboratory reports from three separate classes in order to demonstrate its applicability across a variety of biological content areas. It should be noted that these cross-sectional data on student laboratory reports from BIOL 101, 102 and 301 laboratories served an additional purpose beyond testing the psychometric reliability of the Rubric. The averaged scores produced during that reliability testing were used to evaluate student scientific writing and reasoning in a cross section of student enrolled in separate biology courses. As the BIOL 101 and 102 papers underwent peer review, while the 301 papers did not, this cross-sectional sample also provides a comparison of the performance of younger, less experienced students using peer review, against that of older more experienced students who did not participate in peer review (Study 4). Growth in scientific reasoning in the same students over time was also investigated using longitudinal portfolios of student work from multiple classes over various semesters (Study 5). These two in-depth views of student achievement of scientific reasoning skills were further supported by the more objective multiple-choice *Scientific Reasoning Test.*

The *Scientific Reasoning Test* is a conservative instrument for this study as it tests reasoning ability in contexts outside of and unrelated to the biology courses in which students are learning. Thus, if across hundreds of students from a range of biology courses it detects a relationship between improvement in scientific reasoning scores and the number of peer review experiences in which a student has participated, the effect of peer review must be rather notable (Studies 3 and 7).

Additionally, as the primary source of raters for the Rubric's reliability testing were biology graduate teaching assistants and little research on the consistency or grading habits of graduate teaching assistants has been published in the past, this portion of the study afforded a unique opportunity. Reliability and stringency of scores generated by graduate teaching assistants and the impact of training on those scores is also reported and discussed (Study 6). Graduate teaching assistant perceptions of the Universal Rubric utility are also reported to provide insight on how the Rubric might facilitate science graduate student teaching (Study 8).

Student perceptions of the effectiveness of peer review as a learning strategy are also investigated. Specifically, student perceptions of the role of peer review in the classroom (Study 9) and the scientific community (Study 10) were investigated. As motivation and self-efficacy affect performance as well, student perceptions provide additional insight into successes and continuing challenges of peer review.

Student comments also indicate whether students correctly perceive the instructor's motivation and intent for incorporating peer review into the curriculum. Do students perceive peer review as a useful skill for practicing scientists? Do they believe peer review improves their critical thinking skills or their scientific writing? Overall, these combined measures illuminate the impact of peer review on student achievement and the rate at which their scientific reasoning skills develop as well as their understanding of how peer review fits into the scientific community at large.

Study 1:
Consistency and effectiveness of undergraduate peer reviewers

Peer review capabilities of freshman and first semester transfer students were determined with a multiple data sources from one representative semester of introductory biology (BIOL 102, Fall 2004: n = 320 students). The assumption was made that if these introductory biology students were capable of peer review, students in subsequent courses would also be capable. The class sampled to answer this question had similar demographics to the biology major as a whole (64% female and 55% Caucasian, 24% African-American) and was 64% freshman and 21% sophomores (transfer students or change of major students). As is typical for introductory biology, only 49% of enrolled students were biology majors (35% were related health sciences majors for whom introductory biology is a required sequence). Data sources used were reviewers' ratings of peer papers, reviewers' comments, tracking changes in student writing from draft to final versions of the paper and student self-report data for time on task.

Characteristics of undergraduate peer review

Overall, introductory biology peer reviewers were found to be consistent and effective. Ninety-six percent (96%) of students in the class successfully completed the peer review process (n = 307). Given the detailed nature of several of the data sources, sub-samples were used for some analyses. The average numerical rating given to peers for their draft papers was 5.7 (on a scale from 1 to 10) with a standard deviation across writers of 1.5. As most papers were reviewed by three peers (average number of reviews per paper = 2.75), comparing the text ratings among the reviewers of a paper was a reasonable means of measuring the consistency of these

novice raters. The average standard deviation among reviewers looking at the same work was 1.6 on a scale of 10 (n = 335 reviews of 119 papers). Especially given students' inability to more finely differentiate scores (only integers values are allowed by the *CPR* website), a 16% variation among reviewers was viewed as reasonable. By way of comparison, graduate student raters of similar papers had average standard deviations of 2.0 (trained raters) to 5.0 points (natural raters) on a 45 point scale with average paper scores of 11.7 to 20.8 respectively. As the average standard deviation of scores produced by graduate students who determine grades in the course was 11% of their average total scores (5.0 / 45), it seems acceptable for peer reviewers to have a comparable range of variability (16%). Further, another study of peer reviewer consistency using three peer reviewers produced alpha Cronbach reliability scores of $\alpha = 0.55$ (Cho, Schunn, & Wilson, 2006). Introductory biology students were therefore deemed sufficiently consistent as peer reviewers and therefore, by extension, more advanced biology majors are also assumed to be consistent reviewers.

Introductory biology students took on average 32.4 $\pm$ 14.3 minutes per review (n=182 reviews) including the time required to read the paper (average paper length = 1,265 $\pm$ 448 words, n = 66 papers). Peer review thus did not appear to place an undue burden on students. Ninety-five percent (95%) of the writers who participated in the peer review process also viewed their results as indicated by the *CPR* website login data. Thus, on a broad scale, undergraduates appeared to competently navigate the peer review process and they were reasonably consistent in their estimations of the overall quality of a peer's work.

Incorporating feedback is a distinct process from peer review

Using a smaller sub-sample of papers from this course, an in-depth look at how the quality and nature of peer feedback was made by independent rating (not using the grades assigned by the graduate teaching assistants in the course). Both draft and final papers were scored using the criteria provided to the students. The individual pieces of feedback provided by the reviewers were also evaluated. Feedback was coded as to topic and nature, and the draft and final versions of the paper correlated with the feedback to determine which pieces of feedback appeared to have been used to revise the paper. The average reviewer gave 3.7 $\pm$ 2.6 (out of 10 possible) useful pieces of feedback per review. The average writer therefore

received an average of 10.4 ± 5.0 useful pieces of feedback across all three reviewers.

Receiving and implementing the feedback seem to be two distinct processes however. While 95% of students in the class appeared to have logged in and viewed their results (n = 308 students), but only 54 ± 31% of the feedback received by students in the intensive sub-sample was incorporated on average. Given this notable variation in the use of the feedback with the minimum use being 0% (2 students) and the maximum being 100% (1 student), a correlation between use and gain in points from draft to final version of the paper was plotted (Figure 4.1).

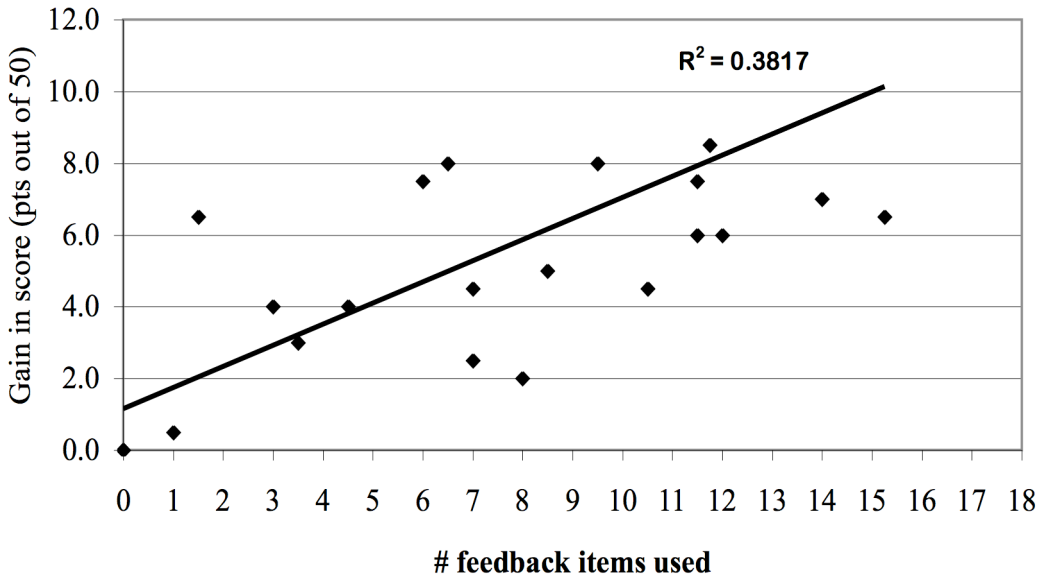


Figure 4.1. Effect of using peer feedback on the quality of students' final papers. Sample from BIOL 102, Fall 2004 (n=22 unique students.) Gain in score is the difference between the score achieved by the draft and final versions of the paper for each student. Number of feedback items used by each writer was determined by correlating all pieces of feedback received by a writer with changes evident in the final paper compared to the draft. By definition, only feedback items deemed useful (see Chapter 3 Methods) were used in this analysis as vague feedback did not generate changes that could be definitively traced back to a particular feedback item.

On average, for every three pieces of useful feedback incorporated, a writer saw a two point (4%) increase in his/her grade on the final paper. This graph was shown to all subsequent classes to encourage them to use the feedback as an effective means of improving their papers. The intercept on the regression line was not set at 0,0 as a writer could have presumably made revisions to the paper and improved it without incorporating any peer feedback. Peers therefore appear to also

be effective at providing useful feedback, both as defined for the purposes of this study and in terms of final grade outcome.

Improving the usefulness of peer feedback

This result supports previous findings that undergraduates perceive the helpfulness of peer feedback as similar to the helpfulness of instructor feedback (2006). Cho, Schunn and Charney (2006) also found that undergraduates were the largest source of praise-based feedback while instructors provided the bulk of the directive feedback (defined as makes specific suggestions for change). These authors therefore recommended that undergraduates be encouraged to be more directive in their reviews.

The handout and instruction provided to students as part of this study were developed prior to the publication of Cho and colleagues' work, but nevertheless were well-aligned with their recommendations. Students were instructed to focus on making their feedback as specific and concrete as possible with the ultimate determinant of feedback quality being whether the comments were sufficient to plausibly generate meaningful change in the writers' paper (see Chapter 3 for more detail). To encourage students to take these recommendations seriously, reviewers were graded by graduate teaching assistants on the quality of the feedback they gave (1 pt for each piece of useful feedback up to a maximum of 10 pts). The criterion of *useful* defined here is equivalent to Cho, Schunn and Charney's (2006) category of *directive.*

Cho, Schunn and Charney (2006) reported that approximately 20-60% of the comments made by undergraduates were directive depending on which sample and rubric criterion were considered. Far more directive feedback was provided in the area of *prose flow* (~40-60% of feedback) and far less for *argument quality* (~45%) or *insight* (~20%). Similarly, when the quality of undergraduates' feedback was reviewed by an independent rater as part of this study, approximately 37% $\pm$ 26% of the feedback was coded as useful. In contrast, instructor feedback was 80% directive for prose flow, 90% directive for argument and 50% directive for insight. So there appears to be room for improvement in undergraduate feedback, though this level of usefulness appears to be normal for undergraduate populations. No comparable information is available for professional peer review. Even if professional referees

written comments were collected (e.g. Marsh and Bell 1981), information on their topology or quality was not reported.

Cho, Schunn and Charney do not provide details however as to what supports were provided for undergraduates in writing their feedback, nor the undergraduates' experience level with peer review. Their online peer review system does allow writers to score the usefulness of the feedback they receive which would provide some motivation to reviewers similar to being graded by a graduate teaching assistant. One of their undergraduate samples was composed of juniors and seniors (who produced the largest frequency of directive feedback) and the other was not specified. These students were predominantly freshman. Presumably, the value of student feedback would be lesser without these supports. So at a minimum, students should be motivated to provide useful feedback by some sort of accountability system as well as being provided with rubric-based guidance and examples of what constitutes useful and productive feedback.

Changes in science writing as a result of peer review

Even if the proportion of useful feedback can be improved, peer feedback does still cause notable improvement in the quality of student writing and thinking. The student papers sampled here were generated as part of a BIOL 102 unit on evolution using data from populations of Galapagos Finches (same assignment as described in Table 3.5). Analysis of the changes made from the draft to the final versions of the paper and correlation with the feedback received by each writer indicated the location of the gain in points (Table 4.1). Overall, an average gain of 15% was seen for each criterion. Four major areas of weakness (defined as students earned < 50% of the possible points) were seen in the draft papers: 1) the comprehensiveness of the data used by the student, 2) refutation of alternative explanations, 3) explanation of evolutionary mechanisms and 4) future directions and significance of the research (Table 4.1, items in italics). While all the areas of weakness improved as a result of peer review, the greatest gains occurred in "Refutation of alternative explanations" (34% increase) and "Explanations of evolutionary mechanisms" (36% increase).

Thus, it appears that some weaknesses in students' science skills were more easily identified by peers. Additionally, peer reviewers' feedback caused greater improvement in those areas than others. Thus, this result suggests there may be a

developmental trajectory wherein some scientific reasoning skills (such as the ability to refute alternatives) develop before others. As students all worked on the same research project, peers likely felt comfortable suggesting alternative explanations and identifying when writers had effectively dealt with those alternatives.

Table 4.1. *Student Performance on Draft and Final Lab Reports and Changes Made as a Result of Peer Feedback.*

Assignment Criteria	**% possible points earned**		
	Draft	Final	Gain
Introduction/Background • Relevant Context • Content knowledge is accurate and relevant	64	72	9
Hypothesis(ses) are • Testable and relevant • Include multiple plausible alternative explanations	64	79	15
Methods: clear, concise description of data collection and analysis.	68	76	8
Results			
• Data presented in a logical, clear format	74	77	2
• Data are relevant to question and clearly tied back to hypotheses being tested	70	82	13
• *Data are complete and comprehensive*	*16*	*27*	*11*
Discussion • Clear, logical and persuasive discussion of why data support one hypothesis over the others	64	78	14
• ***Clear refutation of alternative explanations***	***38***	***71***	***34***
Explanation of evolutionary mechanisms	***19***	***54***	***36***
Future directions, implications	*16*	*26*	*10*
Writing quality • Clear, concise, direct and persuasive.	72	83	11
Total	53	68	15

Note. Sample size is 22 students who each wrote a draft and final version of their lab report.

Students had more difficulty determining whether or not *all* relevant alternatives had been considered in that the criterion of *data are complete and comprehensive* still showed room for improvement. This is not surprising as

identifying when data are incomplete can be more difficult if many students are missing the same data type. Students who make similar errors are less likely to identify the error when reviewing other students work (Schwarz, Neuman, & Biezuner, 2000). Further, the content knowledge necessary to explain evolutionary mechanisms clearly improved from draft to final as well. Discussing the future implications of a research project also appeared to be a challenging area for students at this level.

Summary of results from Study 1: The effectiveness and consistency of peer reviewers

Introductory biology students are capable of engaging in peer review and the process is a reasonable time commitment for an introductory level course assignment. Introductory biology peer reviewers can produce a reasonable number of useful feedback items per review (similar to the proportion reported by Cho, Schunn et al. 2006), though there is distinct room for improvement. Peer feedback was deemed useful both when reviewed by an independent rater and because it produced increases in student scores when writers choose to incorporate the feedback into their revisions. Four areas of weakness were found in student papers in an introductory biology course. Two of those areas improved as a result of peer feedback: students' consideration of alternative explanations and their explanations of evolution. Thus, introductory biology students were effective peer reviewers who stimulated both reasoning and content knowledge gains. It is further plausible that student capabilities will improve with experience.

Study 2:

Reliability of the Universal Rubric for Laboratory Reports

Once it was determined that introductory biology student can engage productively in peer review, subsequent research questions required a common metric for measuring student performance across multiple courses (either longitudinal or cross-sectional). The Universal Rubric for Laboratory Reports (hereafter "the Rubric") was developed to serve as this common yardstick for assessing student performance. Prior to widespread implementation or application however, the Rubric underwent psychometric evaluation to test its reliability as a measurement tool.

Reliability of the Rubric was measured at several levels and in two different contexts. The reliability of each individual criterion was calculated as well as the reliability of the overall total score assigned to each paper. To ensure universality of the Rubric, the reliability evaluation was replicated both within a course (n = 45 to 49 unique student papers per course) and across three different biology courses. In addition, the entire experimental design was implemented with trained raters and then replicated again (using the same papers) using graduate teaching assistants in natural grading conditions to allow extrapolation of the reliability results to more real-world contexts.

Generalizability analysis was used to assess reliability. It differentiates the portion of the variation in scores which is attributable to actual differences in the quality of the papers (reported as *g*) from the variation attributable to less relevant sources such as variation among raters, or assignments or variation due to interaction among those factors such as student-assignment or rater-assignment interactions (Shavelson and Webb 1991). All reliability scores reported here were generalizability scores. In general, only reliability scores generated from comparisons among three raters are reported in the text as they were generated from experimental data. Single rater reliabilities indicate the equivalent confidence one would place in a score generated by a single rater (e.g. an instructor grading a paper) but are function of three rater reliabilities and therefore have a predictable relationship (see Figure 4.2).

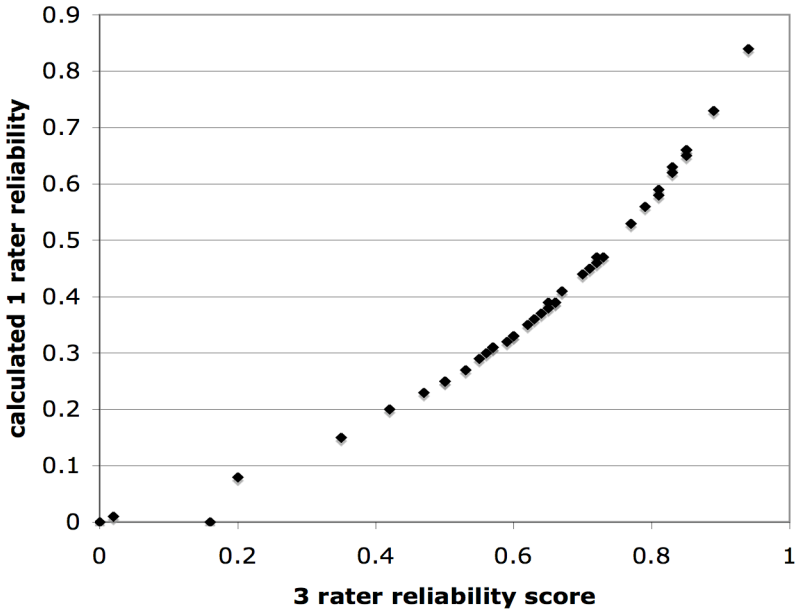


Figure 4.2. Relationship between average three-rater reliability and single-rater reliability scores using data derived from this study (n = 142 papers).

All results reported for rubric reliability refer to scores generated by three trained raters only unless otherwise specified. Data on the reliability of the rubric under natural grading conditions occurs in a separate upcoming section.

Reliability of individual criteria

Not surprisingly, reliability as measured by generalizability analysis varied by criterion (Table 4.2) but with a few exceptions, most criteria were reliable in a variety of contexts. The minimum three-rater reliability (*g*) across all three datasets was 0.20 and the maximum was 0.94 with an average reliability of 0.65 (this result excluded the criteria of Methods: Controls for all three datasets and Results: Statistical Analysis for BIOL 102 and Methods: Experimental Design for BIOL 101 as these criteria are discussed separately). Refer to Table 3.3 for full descriptions of criteria. Considering just the maximum reliabilities achieved, the range of reliability values as measured by generalizability analysis was 0.62 to 0.94 across all criteria (Table 4.2). These results indicate that each criterion is reliable in at least a subset of situations.

Other studies have reported individual criterion reliabilities as low as g = 0.151 for four raters (Baker, Abedi et al. 1995) so these individual criterion measures are quite encouraging in many cases. Baker et al. (1995) only report maximum criterion reliabilities of g =0.722 with four raters (and an average reliability of g

=0.53 over six criteria). It should also be noted that excluding the criteria mentioned above, when criterion reliabilities are averaged over multiple courses, the minimum reliability for any single criterion is 0.49 and the maximum is 0.79 (Table 4.2). Therefore, the reliabilities for these individual criteria are on par with others published in the literature are acceptable for use in this and similar contexts (see also Table 4.3 for published reliability values of total scores). Further, graduate student peer reviews in a course simulating publication in scholarly journals found a similar pattern to that shown here; individual criteria had quite low reliabilities (r = 0.26 to 0.47 for two raters), but that the overall score had a much higher reliability (r = 0.55) (Haaga 1993). This pattern of overall or total scores having equal or higher reliabilities than criterion scores was also found in other studies (Klein, Stecher et al. 1998) and professional peer review of journal submissions and grant proposals (Cicchetti, 1991; Marsh, Herbert W & Bazeley, 1999; Marsh, H W. & Bell, 1981).

Criteria with low reliability scores

A few of the criteria reported in Table 4.2 have reliabilities (g) at or near zero. Due to several of the assumptions behind generalizability theory, this appears to have occurred because the students uniformly failed to perform these criteria (resulting in rating scores of 0) rather than being a direct reflection of the effectiveness of the criteria. Specifically, for Methods: Control in BIOL 102, 132 of 135 scores given were zeros. For BIOL 301, 126 of the given 144 scores were zeros. For *Results: Statistics* for BIOL 102, 123 of the 135 scores were also zeros. Generalizability analysis assumes that there will be variation in performance and perceives such a lack of variation as an indicator of low reliability rather than poor actual performance. If the lack of variation in scores is an accurate reflection of student performance however, then the low reliability score is an artifact of the calculation process rather than an accurate assessment of the criterions' reliability.

Table 4.2. *Reliability of Individual Universal Rubric Criteria Using Generalizability Analysis* (g)

		Course			Overall
		101	**102**	**301**	**Ave**
Criteria	*n* =	*49*	*45*	*48*	*142*
Introduction					
Context		0.67	**0.83**	0.50	0.67
Accuracy and relevance		**0.67**	0.47	0.65	0.60
Hypotheses					
Testable		0.70	0.70	**0.81**	0.74
Scientific Merit		**0.76**	0.66	0.67	0.70
Methods					
Controls		-[1]	0.00[2]	0.16	n/a
Experimental Design		0.20	**0.89**	0.57	0.55
Results					
Data Selection		0.50	0.53	**0.66**	0.56
Data Presentation		**0.77**	0.72	0.64	0.71
Statistics[3]		0.59	0.02[2]	**0.62**	0.61
Discussion					
Conclusions based on data		0.63	0.60	**0.65**	0.63
Alternative explanations refuted		**0.73**	0.55	0.72	0.67
Limitations		0.57	**0.83**	0.60	0.67
Significance		0.56	**0.81**	0.79	0.72
Primary Literature		0.57	0.85	**0.94**	0.79
Writing Quality		0.42	0.35	**0.71**	0.49
Total Score		**0.85**	**0.85**	**0.85**	0.85

Note. Reliability values in bold are the maximum reliabilty score per criterion. Sample sizes reflect the number of unique papers scored per course. All values reported are three-rater reliabilities (*g*) using trained raters. See Appendix 9 for single rater reliabilities. [1]The trained raters for BIOL 101 did not perceive the genetics assignment as providing a traditional control and chose as a group to not rate this criterion. Natural raters scoring 101 papers achieved a three-rater reliability of 0.74 for this criterion however. [2]See section on low criteria reliabilities for explanation.

Two lines of evidence suggest that the uniformly low scores on these criteria were an accurate assessment of student performance rather than the result of a poorly designed criterion. Firstly, post-hoc review of the writing assignments given to the students indicates that while the use of a control was always implied/ conceptualized by the instructors, there was no explicit statement that students should include a control. Similarly, in BIOL 102, students were not instructed to perform statistical calculations or assessments.

This evidence came about because each assignment, while being based on the Universal Rubric, explicitly focused on fewer than 9 of the 15 criteria each. Criteria selection or exclusion was classroom decisions made by the faculty instructors in charge of the courses. Especially with introductory biology, the instructors' believed that students would be overwhelmed with having to provide peer review feedback on 15 substantial criteria and so the peer review, and consequently the written assignment handouts, emphasized a subset of the rubric criteria. If the instructor did not specify that students should incorporate those factors, it is likely that students did not attempt to perform them, and the resulting uniformly low scores were likely accurate. There were other criteria that were not explicitly identified, but on which the students' performance varied (and reliability scores were acceptable), so it is possible that the criterion of Methods: Control was simply a poor one. Alternatively, the inclusion of controls may be a specific skill that has to be learned and without explicit instruction, comes later in a student's development.

A second line of evidence supported this interpretation that the uniformly poor student performance was the cause of the low reliability rather than the utility of the criteria. Criteria with low reliabilities in one course where students were not instructed to incorporate that component had much higher reliabilities in the other two courses where those criteria were included in the assignment. Namely, reliability correlated with criterion inclusion to a certain extent. For example, Results: Statistics had a reliability score of 0.59 in BIOL 101 and 0.62 in BIOL 301 for trained raters. Additionally, in BIOL 101 the scores were notably consistent for the criterion Methods: Experimental Design (only 45% of the scores varied from the mode) and the reliability score was 0.20 again showing the correspondence between a lack of variation in scores and low reliability. Again, this was a situation where student performance was constrained by the assignment. The raters interpreted the students as not being involved in the experimental design (and therefore scored them

low on the performance of that skill) because much of the experimental design was instructor determined. After the completion of the rubric study, this criterion was revised to address both student derived and instructor derived experimental designs.

In the other two courses, the reliability of the Methods: Experimental Design criterion was higher (0.57 and 0.89, Table 4.2). Thus, each criterion performed well in at least one course and there was a correspondence between a failure to include rubric criteria in the assignment and low reliability scores. The low reliability scores for these criteria thus appeared to be more a reflection of poor alignment of the assignment and the rubric rather than an accurate assessment of the reliability of these criteria. Therefore, all criteria performed reasonably whenever they were included in the assignment in a manner that allowed student performance to vary with ability.

Another function of the Universal Rubric is thus to highlight curricular weaknesses or misalignments as advocated by Halonen et al. (2003). The instructors were previously unaware that they had not provided explicit instruction to the students to incorporate controls into their experimental design or to use statistics, nor were they aware that these omissions occurred in multiple courses. Application of a universal metric (such as the Universal Rubric) because it is comprised of curricular goals thus serves a dual purpose of assessing curriculum alignment and progression as well as student development. The corresponding data on student achievement must therefore be interpreted within the context of alignment between assignment and curriculum goals. If assignments do not ask students to perform various scientific skills, students are unlikely to develop those skills over time.

Reliability of the rubric as a whole

Reliability of the Total Score sum of the criteria scores earned for each paper divided by the number of criteria) based on three raters was much higher (g = 0.85 for all three datasets than that for individual criteria (Table 4.2). In general terms, these results mean that 85% of the variation in the total scores was reflective of actual differences in the quality of the papers (rather than rater inconsistency or other sources of error) and that scores generated with the rubric would produce highly reliable grades. Single-rater scores calculated from the three-rater scores were, of course, lower (g = 0.65 to 0.66, Appendix 9), but all are comparable to those found in the literature (see Table 4.3). In addition, the Universal Rubric's single rater

reliability is higher than the average or maximum single rater reliability reported for 16 studies on the reliability of professional peer review (r = 0.19 to 0.54, median = 0.30, (Cicchetti, 1991); $r = 0.27 \pm 0.12$, (Marsh and Ball 1989)) and nine studies of National Science Foundation grant submission reviews (r = 0.17 to 0.37, median 0.33 (Cicchetti, 1991)).

When attempting to compare the Universal Rubric to other published works, it should be noted that no other rubric for university science writing was identified which also underwent reliability testing. Rubrics listed in Table 4.3 are the closest relevant rubrics that could be located. These rubrics evaluate student writing when students are asked to explain, justify or persuade, but come from a variety of grade levels and subject areas outside of science. Several studies described relevant criteria for science writing at the university level (Haaga 1993; Kelly and Takao 2002) but were not rubrics. Halonen et al. (2003) describe an excellent and comprehensive rubric for development of students' scientific reasoning skills, but it is neither designed for science writing, nor reliability tested. Other rubrics found in the literature which focused on science skills either directly observe students in the laboratory (Baxter, Shavelson et al. 1992; Germann and Aram 1996) or study other communication media (laboratory notebooks: Baxter et al., 1992; oral presentations: Hafner & Hafner, 2003; verbal discussion: Hogan, Nastasi, & Pressley, 2000). Baxter, et al. 1992 specifically found that reliability scores varied based on medium of communication so reliabilities for non-written student performances were not included in Table 4.3. By comparison with published results, the Universal Rubric is therefore deemed reliable for written laboratory reports in biology at the university level.

Table 4.3. *Reliability of Professional Peer Review and Relevant Rubrics for Writing.*

Citation	Statistic	# Criteria	# Raters	Reliability Value
Rubrics				
(Baker, Abedi et al. 1995)[1]	α	6	4	0.84 to 0.91
(Cho, Schunn, & Wilson, 2006)[1]	α	3	5	0.88[2]
(Haaga 1993)[3]	r	4	2	0.55
(Marcoulides and Simkin 1995)[1]	g	10	3	0.65-0.75
(Novak, Herman et al. 1996)[1, 4]	g	6	1[5], 2	0.6, 0.75
(Penny, Johnson et al. 2000)[1]	phi	6	2	0.6 to 0.69
Professional peer review				
Meta-analysis (Cicchetti, 1991)	r	various	1[5]	0.33
(Marsh and Bell 1981)	r	5	2	0.51
(Marsh and Ball 1989)	r	4	1[5]	0.30
Meta-analysis (Marsh and Ball 1989)	r	various	1[5]	0.27 ± 0.12
(Marsh, Herbert W & Bazeley, 1999)	phi	Holistic	4	0.704
This study	g	15	3	0.85
			2	0.79
			1	0.65, 0.66

Note. Professional peer review employs lists of criteria rather than rubrics with defined performance levels which may account for the difference in reliability scores. [1]Non-scientific writing. [2]Relability produced by undergraduate peers rather than trained raters. [3]List of criteria only, not a rubric. [4]Multiple rubrics reported in this study, these results refer to the WWYR rubric. [5]Single rater reliabilities were calculated from two-rater data, but reported as single rater reliabilities.

In general, reliability scores increase as the number of measurements (writing samples per student) increases or the number of raters increases (Brennan, 1992; Hafner & Hafner, 2003; Novak et al., 1996) though an increase to four raters from three raters produces a negligible increase in the generalizability co-efficient (Brennan 1992). Longer scales (number of performance levels) do not produce a similar increase in reliability. The optimum number of performance levels appears to center around four (Penny, Johnson et al. 2000) which corresponds to the number

of performance levels in the Universal Rubric. In addition, augmentation of scores (adding "+" or "-" to an integer score) was allowed in this study as it increases reliability to a greater extent than using an integer scale of the same length (Penny, Johnson et al. 2000; Penny, Johnson et al. 2000). Overall, the reliability of the Universal Rubric meets or exceeds that of relevant published comparisons (Table 4.3) indicating that the rubric is an acceptably effective psychometric tool. Further, the little information that currently available on the consistency of graduate teaching assistants indicates that there is little correlation in grades among instructors (Kelly & Takao, 2002). Therefore, tools or pedagogical strategies which improve reliability are desirable. Reliability generally increases as scores are summed across multiple criteria (e.g. Total Score has a consistently higher reliability than vs. any single criterion) (Cicchetti, 1991; Marsh, Herbert W & Bazeley, 1999). Consequently, practitioners are encouraged to use as many criteria as are relevant for assessing student performance.

Impact of assignment alignment on criteria reliability

As demonstrated earlier, instructor exclusion of criteria in assignment instructions can strongly impact whether or not students attempt to address criteria and consequently affect the reliability of a criterion. Alignment of rubric criteria and course assignments are shown in Table 4.4. The choice by instructors to emphasize a subset of the criteria and excluding other criteria appeared to have affected some reliabilities (e.g. Methods: Controls). In contrast, other criteria seem to be naturally incorporated into student thinking. For example, the concept that hypotheses should have scientific merit (some hypotheses are more interesting or worthwhile to pursue than others) was not explicitly mentioned in any assignment (Hypotheses: Scientific Merit, Table 4.4), yet there was very little variability in the reliability of this criterion across the three courses and it had a reasonably high reliability score ($g = 0.70$).

Table 4.4 *Inclusion of Rubric Criteria in Course Assignments*

		Criterion incorporated into the assignment for the course indicated?		
Criteria	Code	101	102	301
Results: Statistics	R: St	Yes		Yes
Discussion: Conclusions based on data selected	D: C	Yes	Yes	Yes, but highly implicit
Hypotheses: Scientific merit	H: S			
Hypotheses: Testable and consider alternatives	H: T	Partial: "clear with rationale"	Yes, verbatim	Yes, clearly stated
Results: Data presentation	R: P	Yes	Yes	Yes
Results: Data selection	R: S	Determined by instructor	Yes	Yes, but implicit
Discussion: Alternative explanations	D: A		Yes	
Introduction: Accuracy and relevance	I: A	Yes	Yes, verbatim	
Discussion: Significance of research	D: S	Yes	Yes	
Discussion: Limitations of design	D: L	Yes		
Methods: Controls	M: C	n/a		
Methods: Experimental design	M: E	Determined by instructor	Yes, but implicit	
Introduction: Context	I : C	Yes	Yes, verbatim	
Writing Quality	WQ	Yes	Yes	Yes
Primary Lit	PL	Yes, 2 required	Bonus only	Partially, citations required, but primary lit not specified.

Note. Criteria are rank ordered from least variable to most variable based on spread between minimum and maximum reliability per course reported in Table 4.2. Blank cells indicate that the criterion was not explicitly mentioned in the assignment. For BIOL 101 and 102, alignment designations were derived directly from the grading rubric handed out to students in the class. For BIOL 301, no written assignment was given to the students. Alignment of the assignment with the rubric was generated by 301 teaching assistants reviewing the list of criteria shortly after the assignment occurred and identifying those they felt were communicated to the students. Codes are provided to facilitate comparison with data presented in Figure 4.3.

Thus, this analysis would seem to indicate that instructors should not presume that all scientific reasoning skills are equally easy or difficult for students to develop.

Some aspects of experimental design such as methodological controls, incorporation of statistics and discussion of limitations and implications do not appear to come naturally to students and require explicit pedagogical support.

In contrast, writing quality appears to be widely valued by instructors as well as practicing scientists (Yore, Hand, & Florence, 2004) and was included in all three assignments, yet had a much lower average reliability (0.49) and a large spread (minimum reliability 0.35 and maximum reliability 0.71). Graduate student raters apparently find it easier to assess the merit of scientific hypotheses, than to assess writing quality (the variation in scores assigned for this criterion was more than twice the average (SD = 0.44 compared to 0.2, see Table 4.2). The criterion of Writing Quality developed for the Universal Rubric appears to be subject to greater interpretive latitude than other criteria despite being inspired by the South Carolina Department of Education English Language Arts Rubric (2006). While revision of the criterion may improve reliability, it is also possible that as it is a more holistic criterion (raters must consider the writing quality of the entire work at once) high levels of reliability may simply more difficult to achieve. With the exception of Writing Quality, explicit inclusion of criteria in assignments appears to improve reliability however. To test this conclusion, a post hoc analysis was performed. Reliability scores for individual criteria for each course were drawn from Table 4.2 and overlaid on the inclusion information provided in Table 4.4. Each criterion was categorized for each course assignment as being included, partially included or excluded and its reliability score (*g*) averaged accordingly (Table 4.5).

Table 4.5. *Correspondence Between the Inclusion of Criteria in an Assignment and Criterion Reliability*

	Rubric criterion included in assignment?		
	Yes	Implicitly	No
Average reliability score (*g*)	0.63	0.68	0.55
Standard deviation	0.12	0.25	0.27
n =	23	7	14

Note. Reliability scores of individual criteria in each course were categorized according to the degree of inclusion in that assignment. Sample sizes are the number of reliability scores in that category. *Methods: Control* was not included because the BIOL 101 raters intentionally omitted it).

Approximately half of the time, criteria were explicitly included in the assignment instructions across the three courses (n = 23). In some of these instances, the assignment used criteria wording from the Universal Rubric verbatim. In seven other instances, criteria were implicit or partially included in the assignment. For example, for the criterion "Hypotheses are clearly stated, testable and consider plausible alternative explanations" (Table 3.3) was rated as partially included in the BIOL 101 assignment because reviewers were asked to evaluate if "the hypothesis [was] clearly stated for the unknown cross?' and "[were] observations given here as rationale for the hypothesis?" In 14 instances Universal Rubric criteria were not mentioned in the assignment in any way. The variation in reliability scores clearly increases as criteria are left of out assignment instructions (doubling of the standard deviation from included to excluded, Table 4.5). Thus, it is recommended that criteria be included explicitly in assignment instructions if that concept will comprise a portion of a student's grade. If instructors choose to leave out particular criteria from assignment instructions, then that information is necessary for any curriculum-wide comparison of scores to be properly interpreted. Student performance is sensitive to context and poor performance is only meaningful if students were explicitly instructed to attempt a criterion.

In short, no single criterion should be used alone to indicate the quality of a student's scientific reasoning ability, but when grouped, the collective score gives a reliable indication of student performance. Reliability generally increases as criteria are explicitly included in assignment instructions. Some criteria are more natural than others and student performance addresses the criterion of scientific merit even though it was not explicitly mentioned in the assignment instructions.

Effect of biological subject matter on the reliability of the Universal Rubric

The Rubric was intended to be universal meaning applicable to all experimental research projects in which students were likely to engaged while completing their bachelor's degrees in biology. This need for the Rubric to be reliable regardless of the biological subject matter was the main motivation for testing reliability over three separate courses. Results to date support the conclusion that the Rubric functions independent of subject matter. Criterion reliabilities (g) vary as a function of the inclusion or exclusion of criteria, or other factors, but do not appear to vary as a function of the course. Specifically, reliability maxima are

evenly distributed among the courses (101 and 102 each have 4 maxima, 301 has 6 maxima, Table 4.2). Reliability of the Total Score in particular is consistent (indeed identical) regardless of course. Thus, the Rubric's reliability appears to be independent of subject matter.

Summary of results for Study 2: for the reliability of the Universal Rubric

The Universal Rubric was found to be a reliable tool ($g = 0.85$) for measuring students' overall performance in the design, implementation and interpretation of scientific research. Its reliability was also independent of biology content area, with no notable differences occurring among the three separate courses. Total Scores had notably higher reliabilities than any individual criterion on average.

Comparison of student performance on individual criteria was contextually sensitive however. For many of the criteria, failure to explicitly include the criterion in the assignment resulted in poor performance on that criterion. Reliability of a few criteria could not be completely explained by inclusion or exclusion of the criterion in the assignment however, so confidence in the reliability of student scores was highest when points for multiple criteria were summed. Data on student achievement must therefore be interpreted within the context of alignment between assignment and curriculum goals.

Study 4: Student achievement of scientific reasoning skills in laboratory reports (cross-sectional sample)

When student performance from a cross-sectional sample of laboratory reports was viewed across all three courses, there was a decided trend of improvement from 101 to 102 and a notable decline in 301 scores for most criteria (Figure 4.3). Average scores for 12 of the 15 criteria were significantly different from one course to the next (ANOVA $p = 0.001$). The primary explanation for the decline in 301 scores was that the 301 papers reported in the cross-sectional study did *not* undergo peer review. These results suggest that peer review had a noticeable impact on students' performance. For the courses where peer review did occur, students in 102 had significantly higher scores than students in 101 for 7 of the 12 significant criteria (Figure 4.3). Of the five criteria in which 101 students had higher scores, three were explicitly included in the 101 assignment, but not in the 102

assignment (Results: Statistics, Discussion: Limitations, Primary Literature). There are several potential explanations for the increase in scores from BIOL 101 to 102.

The first and most obvious explanation would be that scores increase as experience with peer review and scientific writing and reasoning increase. As these scores represented a cross-sectional rather than longitudinal sample however and data on the number of prior peer review experiences were not available for these students, this conclusion remains speculative. As both samples of student papers were collected in the Fall of 2004 (a year when sequential progression through the courses was not enforced), it was possible that 102 sample included many first semester freshman.

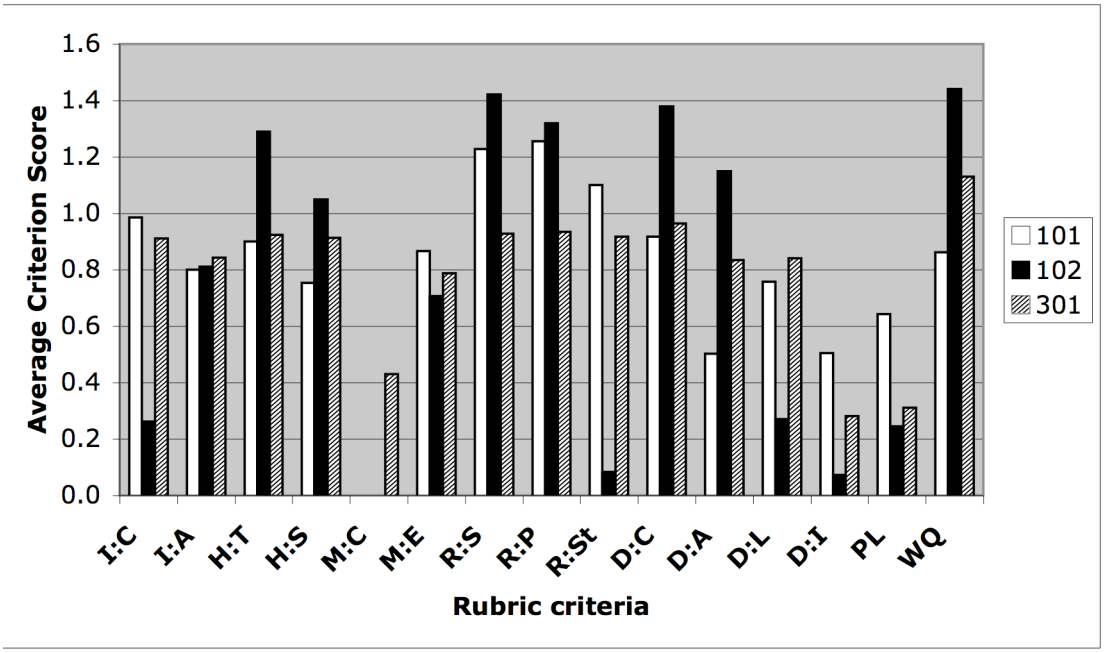


Figure 4.3. Student performance across a cross-section of biology courses.
All scores within a criterion significantly different at $p < 0.001$ level except for I:A, M:C and M:E. The maximum score possible per criterion is 3.0. Refer to Table 4.2 for Rubric criteria codes. Sample sizes were BIOL 101 (n = 49 papers); BIOL 102 (n = 45papers); BIOL 301 (n = 48 papers).

Another potential explanation for the difference in scores derives from differences in how the peer review experience was constructed for BIOL 101 vs. 102. The peer review experience in BIOL 101 was much more heavily scaffolded. Approximately 37 yes/no and high/med/low multiple choice queries comprised the criteria. It provided only a few opportunities for open-ended written. The peer review points earned by students were based on how well their multiple choice

answers aligned with each other and those of the instructors, rather than on the content of open-ended text responses as in 102. Essentially, students earned points for completing the online peer review process regardless of the quality of the written feedback they provided.

Students in BIOL 101 were encouraged to take the process seriously and to provide substantive feedback, but no systematic mechanism held them accountable for the quality of their feedback. The faculty instructor did spot-checks, or investigated if a writer complained, but in a course of several hundred students, evaluation of the quality of reviews did not involve many students. The BIOL 102 peer review experience focused on less than a dozen criteria and required open-ended responses be provided for all those criteria and that at least ten pieces of useful feedback be provided per review. Graduate teaching assistants randomly graded the quality of the feedback for one review per student to ensure accountability. BIOL 102 students also conducted a laboratory exercise (complete with handout) on how to define and provide useful feedback. Therefore, it is possible that the greater scores earned by the 102 papers are a result of higher quality peer feedback in that course due to the relative emphasis placed on the quality of the feedback.

The most likely explanation for the lower performance of the 301 students is that the 301 papers did not undergo peer review and so lacked peer feedback or subsequent revision. Students in this BIOL 301 laboratory sample (Fall 2005) had earned an average of 90.6 ± 32.8 total credit hours indicating more than three years of academic experience and had an average institutional (USC) GPA of 3.14 ± 0.62 on a four-point scale. In contrast, the students in BIOL 101 and 102 were predominately freshman (64%-76.5%) and had lower institutional GPAs (3.00 ± 0.85 for 101 and 2.71 ± 0.87 for 102). Thus, the 301 students possessed greater academic experience and had stronger academic records. Consequently, it is unlikely that their lower scores were the result of lesser academic experience at the university level or lesser academic success in other courses. BIOL 301 student appear to be more experienced and academically competent lending support to the idea that the difference in scores is the result of the lack of the peer review experience. So the effect of peer review appears to fade over time if the process is not continued.

The lower GPA of BIOL 102 students compared to BIOL 101 students further suggests that higher scores on scientific reasoning in 102 were caused by

differences in the peer review process rather than by differences in student demographics.

Additionally, it should be noted that student achievement on individual criteria appeared to be affected by alignment between the assignment and the Rubric in the same way that reliability was. Namely, students tended to perform poorly on criteria which were not included in the assignment. For example, the BIOL 102 assignment did not ask students to perform any statistical tests, nor require them to use any primary literature and students performed quite poorly on those items. It should also be noted that overall performance is low for all courses. On average, students scored at the novice level (1 point out of a maximum of 3) regardless of criterion, course or level. This is appropriate for the introductory biology students and provides ample room for higher level courses to further develop students' scientific reasoning skills.

Summary of results for Study 4: Quality of laboratory reports as a result of peer review (cross-sectional sample)

This cross-sectional sample consequently suggested that peer review improved student performance to a greater extent that generalized academic experience or ability. Students in courses which engaged in peer review tended to produce higher quality laboratory reports than students in a course which did not engage in peer review despite the fact that the students who engaged in peer review were less academically experienced and had a lower average GPA. Specifically, the highest scores for 7 of 12 criteria occurred in the BIOL 102 lab reports despite the fact that BIOL 102 students had lower average GPA and fewer credit hours than BIOL 301 lab students. BIOL 101 students outperformed BIOL 301 students in an additional 5 criteria. The stronger performance of BIOL 102 students may be due at least in part to the fact that the BIOL 102 peer review process had the greatest emphasis on students providing substantial and meaningful feedback as reviewers were actually graded on the quality of the feedback they provided. BIOL 102 students also had more experience with the peer review process on average than did BIOL 101 students which may also have bolstered performance.

Lastly, student performance was improved when Rubric criteria were strongly incorporated into the assignment. On average, students performed at the novice level. Such performance is appropriate for those who were enrolled in

Introductory Biology at the time. Without peer review, students in the 300 level biology laboratory also performed at the novice level. No information was available here for 300 level student performance on laboratory reports with peer review.

Study 5:
Student achievement of scientific reasoning skills in laboratory reports (longitudinal sample)

Longitudinal data on 17 students were generated by combining the rubric study with an independent rating of additional papers produced subsequent to the rubric study. The independent rater had internal reliability checks whereby the duplicate copies of the same paper were assigned different identity codes and inserted into the scoring stack as if they were independent papers. For 60% of the papers, the independent rater's Total Scores on redundant papers were less than 1 point different (on a 45 pt scale). The average difference in Total Score across all redundant papers was 1.53 points. As trained rater Total Scores had a standard deviation of 2.0 (see Table 4.14) the independent rater's scores were considered equally reliable. For papers that were part of the rubric study, the total scores across all three trained raters were averaged and that average value used as the score for this longitudinal study.

In contrast to the cross-sectional data, when the scores earned by a particular student were plotted chronologically, there was a no significant difference among scores earned in the different classes. The reader should recall that inclusion or exclusion of a criterion from an assignment may impact student performance (and hence score). This lack of significance is true however, regardless of whether all 15 criteria are considered, or just the six criteria that were equally emphasized across all three assignments (Figure 4.4). There does appear to be a positive trend of increasing score from 1st semester of introductory biology to 301, but this perception should be guarded against for three reasons.

Firstly, given the lack of significance, the trend may not exist at all. Secondly, the positive trend is not evident at the level of individual students. Only five of 17 students made large gains from introductory biology to 301. Their gains were sufficiently large however to obfuscate the fact that 12 of 17 made no gain or declined when the average is calculated (Table 4.6). Thirdly, beyond statistical

significance among means, it should be noted that gains must be larger than 2.0 points to be considered meaningful. This cutoff was selected because the average standard deviation among three raters on a single paper ranged from 1.94 for BIOL 301 to 2.13 for BIOL 102 (BIOL 101 had an average standard deviation of 2.06).

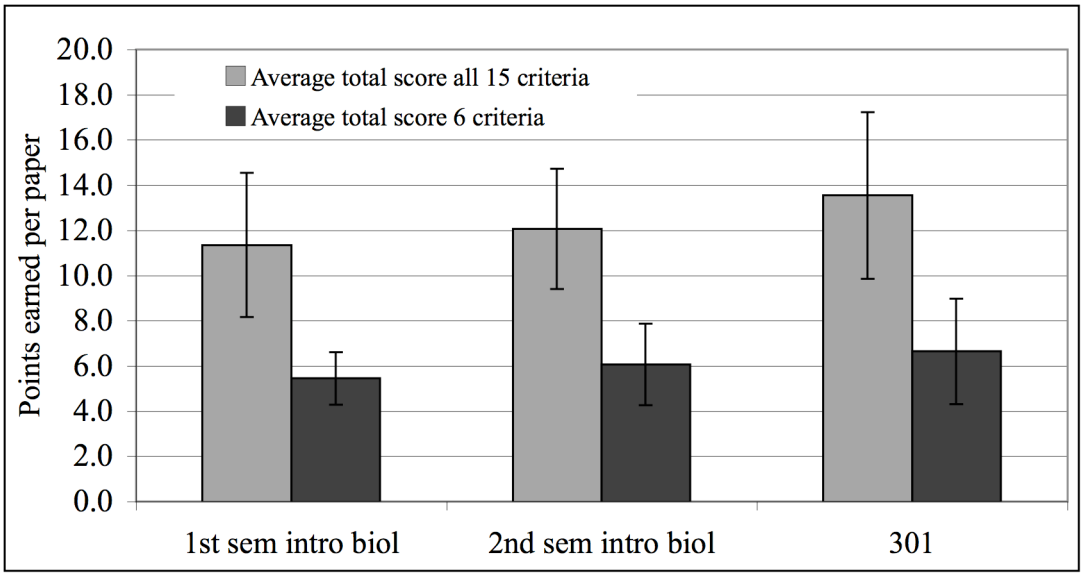


Figure 4.4. Average scores earned by laboratory reports across multiple courses from longitudinal sample (n = 17 students). As some students in this sample took BIOL 102 prior to BIOL 101, results are reported in chronological order rather than by course. There is no significant difference over time within a set of criteria. Darker bars are from the subset of six criteria that were emphasized equally across all three assignments (refer to Table 4.5).

The trend from first to second semester of introductory biology may be more robust because four students made improvements of greater than 2.0 points from the first to second semesters of introductory biology while five had no change (Table 4.6). None showed a decline. Twelve of the 17 students showed no change or a decline in score from introductory biology to 301, but five had notable gains (29%) sufficient to increase the average. A larger sample size may either reinforce the general positive trend until it is clear or provide explanatory insight for the lack of improvement. Additionally, the majority of the 301 papers did not undergo peer review, so significant gains may be realized if a peer-reviewed assignment is selected for sampling at the 300 level.

Thus with a larger sample size such as was available for the cross-sectional sample, statistically significant change may be observed. Unfortunately, due to the

unconstrained nature of the influences and challenges faced by college students, it is quite common for longitudinal studies in higher education to suffer attrition rates of 43% to 96% (Haswell 2000). Concerted efforts will be necessary to provide a larger sample size in the future.

Table 4.6. *Longitudinal Performance of Individual Students Using Laboratory Report Total Scores.*

Student	1st semester Intro Biology	2nd semester Intro Biology	BIOL 301
A	7.0	15.7	12.4
B	11.8		15.1
C	11.2		11.2
D	14.7		13.8
E	8.4	9.7	17.3
F	9.3	10.0	6.3
G	14.3	15.3	13.7
H	8.4	8.3	22.0
I	10.5	9.7	9.0
J	14.5		17.9
K	5.7	14.0	15.5
L	7.7	12.7	12.7
M	15.6		8.2
N	13.9		18.4
O	11.7	14.2	12.7
P	12.1		15.4
Q	15.1		11.2
Average	**11.3**	**12.1**	**13.7**
SD	**3.2**	**2.7**	**4.0**

Note. Second semester papers were not available for some students. Gains must be greater than 2.0 points on this scale to be considered meaningful. Laboratory reports were scored using all 15 criteria regardless of assignment inclusion or exclusion. Four of nine (44%) students produced meaningful gains from 1st to 2nd semester in introductory biology with the rest showing no change. Five of 17 (29%) showed gains from introductory biology to 301, four showed declines and seven were neutral.

No comparable longitudinal studies of science writing were found in the literature, but a few longitudinal studies of undergraduate composition were available. When student essays for placement into freshman and junior year English composition courses were compared, significant "changes toward competent, working-world performance" were found (n = 64, ANOVA $p \leq 0.02$) and the mean number of words per sentence and mean clause length increased (Haswell 2000, p. 307). Another longitudinal study which collected writing samples over entire

undergraduate careers indicated that while students appear to learn from writing, even after four years students may not have received sufficient support in their coursework to gain analysis and synthesis skills or write in sophisticated or complex ways (Sternglass 1993).

Summary of results for Study 5: Quality of laboratory reports as a result of peer review (longitudinal sample)

Study 5 did not identify large changes in scientific reasoning ability. Review of the literature suggests that this is not surprising given that the three writing samples were all generated within a three-semester period (Fall 2004 to Fall 2005) and the sample size was quite small. Additionally, some of the endpoint (BIOL 301) essays did not undergo peer review.

Study 3: *Reliability of the* Scientific Reasoning Test *in this undergraduate biology population*

While student performance on laboratory reports is the most direct and rich source of data for evaluating the effect of peer review on student inquiry abilities. Collection of longitudinal portfolios and scoring of lengthy reports is a time and resource consuming process that only allowed evaluation of a subset of majors. Coarser-grained measures therefore serve a useful function as they allow sampling of entire cohorts of majors. Additionally, if coarser-grained measures show an effect of peer review on student reasoning ability than that effect will likely be richer and deeper with more fine-grained measures. The coarser-grained measure selected for use here was the *Scientific Reasoning Test* (*SRT*) (Lawson, 1978) developed for use in higher education large enrollment biology courses. While found to be reliable and informative in such settings in other institutions (Lawson, Anton E, 1979, 1980, 1983; Lawson, Anton E, Alkhoury, Benford, Clark, & Falconer, 2000; Lawson, Anton E, Banks, & Logvin, 2007), reliability was also assessed directly in this study population.

Typical factors affecting reliability of a psychometric test are: the instrument, the population, the setting, the raters (if applicable) as well as all the interactions between these factors and the unavoidable "other sources of error." The *SRT* was administered in two different terms and six different biology courses including

introductory biology and upper division courses for a combined sample size of 851. In Spring 2005, the test was administered pre-post in BIOL 102 (n= 303 students who took the test) and the Kuder-Richardson 20 (KR20) pre-test score was 0.83. In Fall 2005 it was administered again at the beginning of the term to 548 students in five different biology courses ranging from introductory biology to a 500 level upper division course. The corresponding KR20 score (n = 548 biology majors) was 0.85 in that administration. These reliability values meet or exceed those published recently for this test (see Table 2.3, Lawson, Anton E, 1978, 1983; Lawson, Anton E et al., 2000; Lawson, A.E, Baker, DiDonato, Verdi, & Johnson, 1993; Lawson, Anton E et al., 2007; Norman, 1997). Thus, the instrument is reliable in the population of biology majors at this institution (Cudek 1980).

The *SRT* was developed using Piaget's concepts of concrete and abstract reasoning. These reasoning patterns are predicted to develop once students reach the stage of logical operations (usually around seven or eight years of age) (Karplus, 1977). The advancement of students from concrete to abstract thinking is not thought to be a linear or unidirectional progression however and secondary and tertiary students may exhibit either or both reasoning patterns under various circumstances (Karplus 1977). Published average scores indicate that there is not an automatic increase in *SRT* with an increase in student chronological age. Westbrook and Rogers (1994) administered the *SRT* pre-post to 56 ninth graders (average age = 15.3 years) to test the effect of various instructional strategies. There were no significant differences (though general positive trends did exist) between pre and posttest scores for any group (ave $\pm$ stdev = 4.21 $\pm$ 2.07 to 5.15 $\pm$ 1.98, no reliability value reported). In our administration, freshman had an average score $\pm$ standard deviation of 4.89 $\pm$ 2.02 and 78% of students were between the ages of 19 and 20 thus confirming that scores should not be expected to increase simply due to chronological age. Other administrations of the *SRT* reported above were not scored using the two-tiered system and used either a 22 or 26 item version of the test so similar comparisons could not be made.

Summary of results for Study 3: Reliability of the Scientific Reasoning Test. The *Scientific Reasoning Test* was found to be as or more reliable (KR20 = 0.85) in this population of undergraduate biology students as in other published studies.

Study 7:

Relationship between Scientific Reasoning Test *scores and peer review experience*

Administration and sample reduction decisions

The *Scientific Reasoning Test (SRT)* was administered to a cross-section of biology majors in 100 to 500 level courses in the Fall of 2005 in order to capture students with a range of peer review experiences. From the initial sample of 1048, non-biology majors were removed leaving 573 students. The *SRT* was administered at the beginning of the semester with students self-reporting the number prior peer review experiences. Of the 573 biology majors, 123 did not report the number of prior peer review experiences leaving a sample of 451 (Table 4.7). As Spring 2005 was the first semester that peer review was implemented in any course beyond the introductory biology (101 and 102) courses, it was uncertain how many students in 302L or 530 might have had three consecutive peer review experiences.

Table 4.7 *Distribution of Biology Majors' Prior Peer Review Experiences as a Function of Course Enrollment*

	# prior peer review experiences reported					
Course	Zero	One	Two	Three	Four	Total
101	*24.4*	2.2	0.7	-	-	27.3
102	12.0	*5.3*	2.0	-	-	19.3
301	12.2	4.0	*8.0*	1.1	-	25.3
302	7.1	5.1	*6.0*	*0.7*	0.2	19.1
530	5.5	2.2	1.1	0.2	0.0	9.1
Total	61.2	18.8	17.7	2.0	0.2	100.0

Note. Values reported are the percentage of students reporting that number of cumulative peer review experiences (n = 451 students). BIOL 101 and 102 are introductory biology. BIOL 301 and 302 were the lab components of sophomore level classes. BIOL 530 was Histology. Numbers in italics are the expected number of peer review experiences for a student progressing through the curriculum in the traditional fashion.

As the data in Table 4.7 make clear, more than half of the biology majors in the sample did not progress through the curriculum in the intended way. They had

either too many peer review experiences for their course level indicating that they had retaken a course, or too few indicating that they were transfer student who had taken some proportion of the curriculum elsewhere. The maximum number of legitimate peer review experiences that a student could have accumulated at this point is three (BIOL 101, 102 and 301L, marked in Table 4.7 with italics) with three unique experiences being a possibility only for students enrolled in 302L or 530. Only 44.3% of the students in this sample are successful immersed participants in the USC Biology Curriculum (values on the diagonal in italics in Table 4.7). Course enrollment is therefore clearly not an effective proxy for peer review experience. In-depth review of the remaining 55.7% of the students commenced. Some students retaking courses had not actually completed the peer review experience in past enrollments, while others had. Past research has shown that student performance is affected by a variety of factors that are magnified when students retake classes. For example, the lower self-efficacy or motivation associated with having failed a class once could reduce performance (Lawson, et al., 2007; Mistler-Jackson & Songer, 2000; van Berkel & Schmidt) in subsequent enrollments. Repeated exposure to the same task (writing assignment) can affect performance was well (Anderson, Fisher et al. 2002). Therefore, the decision was made to remove students who had failed and were retaking a class thereby reducing the sample size to 389 (Table 4.8).

Table 4.8 *Distribution of Students' Prior Peer Review Experience Once Students Who Repeated Courses are Removed.*

Course	**# reported prior peer review experiences**					
	Zero	One	Two	Three	Four	Total
101	***27.8***	-	-	-	-	27.8
102	13.1	***5.7***	-	-	-	18.8
301L	12.9	4.4	***9.0***	-	-	26.2
302L	7.7	5.4	***5.4***	-	-	18.5
530	5.9	1.5	1.3	-	-	8.7
Total	67.4	17.0	15.7	-	-	100

Note. Values are percentage of the total sample (n = 389 students). Values in bold italics represent the expected progression through the curriculum. Students with alternative combinations of course and peer review experience (values not in italics) have taken some portion of their relevant biology coursework at other institutions (transfer students).

Students who had fewer than the expected number of peer review experiences due to transfer credits comprise slightly more than half the total sample reported in Table 4.8 heavily skewing the sample towards no prior peer review experiences (67.4% of the total sample). This skew does allow however a nice opportunity to distinguish between the effect of academic maturity (credit hours) and peer review as notable proportions of the sample have high numbers of credit hours with little peer review experience.

The reader should please note that this variation in credit hours versus number of peer review experiences serves as the basis of comparison in this cross-sectional sample. As these *SRT* scores were earned at the beginning of the fall semester, there are students present in the sample who are incoming freshman (and hence have no peer review experience) as well as transfer students with no peer review experience but many credit hours as well as students who have always attended USC but recently changed into the biology major. This variation in experience by academic maturity among a cross-sectional sample of biology majors was the most valid and informative comparison that could be generated in these circumstances. The experience of attending college itself is expected to improve students' reasoning abilities, so the effect of peer review experience on scores must be evaluated by comparing it to the natural and expected gain due to increased academic maturity.

Student performance on the Scientific Reasoning Test

Student performance on the *SRT* shows a notable relationship to general academic maturity (defined as total credit hours earned) (Figure 4.5). Credit hours were translated into academic class (freshman, sophomore, etc.) using the standard conversion of 30 hrs per year. Student performance did vary significantly with total credit hours at the $p = 0.011$ level (see Table 4.9 for ANOVA results). An increase in *Scientific Reasoning Test* score with increasing academic experience is not surprising; it would be extremely disheartening if students' reasoning ability did not improve over years of university level coursework. The question is how strong is the effect of peer review compared to that of academic maturity or institution.

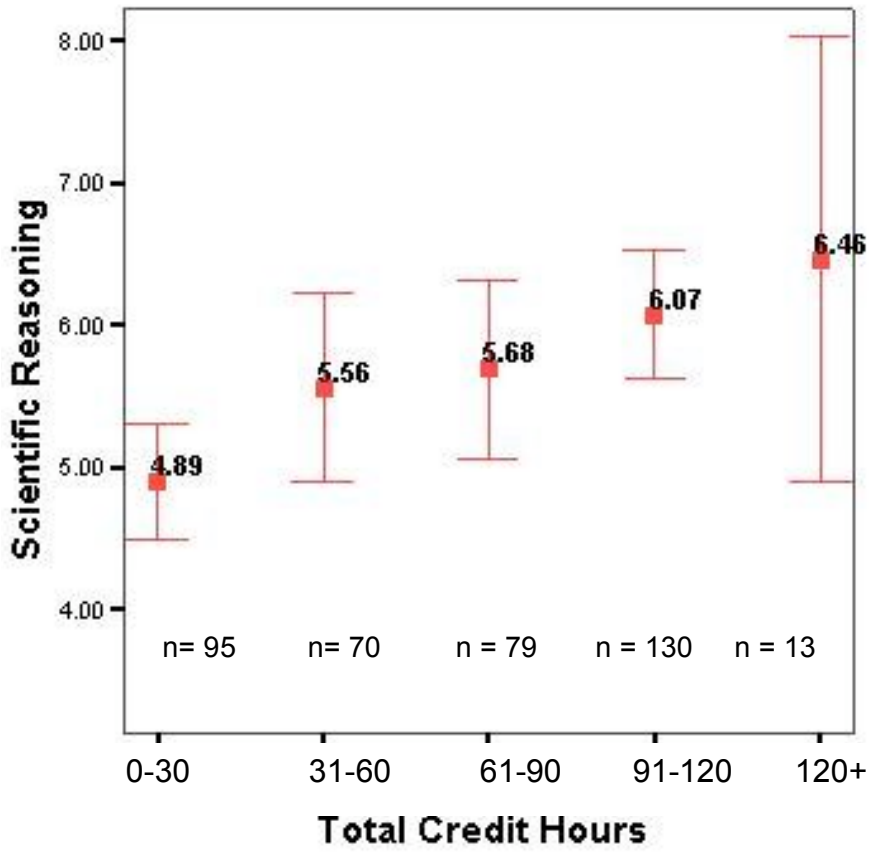


Figure 4.5 Relationship between academic maturity and students' *Scientific Reasoning Test* scores. Total credit hours are cumulative and include collegiate transfer credits, and USC credit. Scores are significantly different among groups ($p = 0.011$, see Table 4.9 for ANOVA results). Sample sizes are the number of students per group. Error bars are 95% confidence intervals around the means.

Table 4.9. *ANOVA Results When* Scientific Reasoning Test *Scores are Sorted by Total Credit Hours*

	Sum of Squares	df	Mean Square	F	Significance
Between Groups	86.005	4	21.501	3.295	.011
Within Groups	2492.915	382	6.526		
Total	2578.920	386			

Note. These results correspond to Figure 4.5

As a large proportion of the students in this sample were transfer students, a comparison was also made between performance on the *SRT* and the number of USC credit hours earned (Figure 4.6). The source of the credit hours was a concern as students with greater peer review experiences would also have greater USC credit

hours than transfer students with similar backgrounds. Significant differences in performance as a function of USC credit hours was also found at the 0.005 level (Figure 4.6). Notably, the pattern is not the same as for total credit hours with the students with >90 USC credit hours having a lower score then the preceding class. Explanations for the drop-off for students who have >90 credit hours may include delayed administration or administrator effects. Only 85 students in the entire sample of biology majors had > 90 USC credit hours and 67 of those were enrolled in BIOL 530. The *SRT* was administered to BIOL 530 several weeks into the semester as opposed to the first week as was the case with the other courses. Additionally, this researcher was not able to be present in the BIOL 530 administrations, as she had been in the other administrations, and does not have any objective information as to the seriousness or effort that 530 students were asked to invest in the test. So reduced effort due either to upcoming examinations or context in which the *SRT* was presented may have affected the effort students enrolled in BIOL 530 as a cohort. As these students comprise 78% of the students in that category, reduced effort in the 530 class is a plausible explanation for the lower scores.

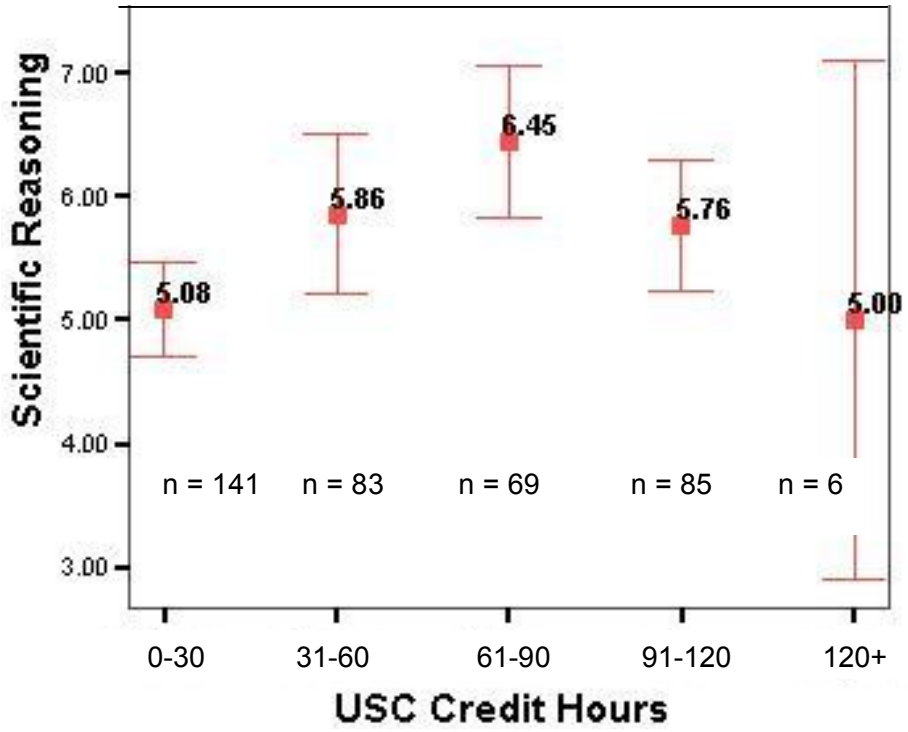


Figure 4.6. Relationship between students' scores on the *Scientific Reasoning Test* and time spent in the USC curriculum (USC credit hours). Scores are significantly different among groups (p = 0.005 see Table 4.10). Sample sizes are the number of students per group. Error bars are 95% confidence intervals around the means.

Table 4.10. *ANOVA Results When Transfer Credits are Excluded and* Scientific Reasoning Test *Scores are Sorted by University of South Carolina Credit Hours*

	Sum of Squares	df	Mean Square	F	Significance
Between Groups	97.349	4	24.337	3.295	.005
Within Groups	2442.773	379	6.445		
Total	2540.122	383			

Note. These results correspond to Figure 4.6

In contrast, when student scores on the *SRT* were sorted by number of peer review experiences, scores rise consistently (Figure 4.7). The relationship between peer review and scientific reasoning scores was significantly stronger than that of either total credit hours or USC credit hours ($p < 0.000$). In addition, the maximum value achieved for any group mean in this analysis (6.82) was for students with two peer review experiences and the largest gain (1.6 points) occurred from zero to two peer review experiences.

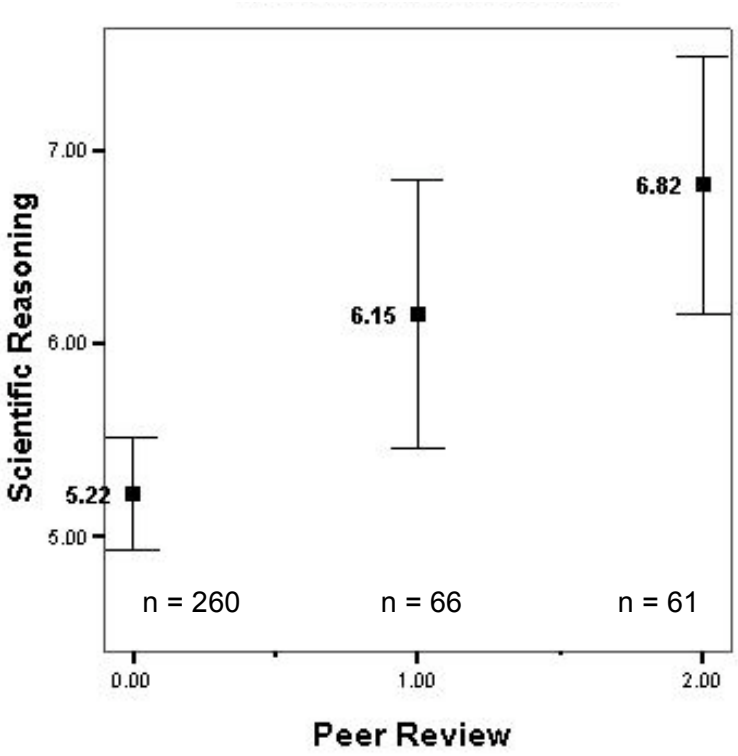


Figure 4.7. Relationship between students' scores on the *Scientific Reasoning Test* and the number of peer review experiences in which they have engaged. Scores are significantly different among groups ($p = 0.000$ see Table 4.11). Sample sizes are the number of students per group. Error bars are 95% confidence intervals around the means.

Table 4.11. *ANOVA Results When* Scientific Reasoning Test *Scores are Sorted by the Number of Peer Review Experiences in Which a Student Has Engaged*

	Sum of Squares	df	Mean Square	F	Significance
Between Groups	148.851	2	74.425	11.697	.000
Within Groups	2456.100	386	6.363		
Total	2604.951	388			

It should be noted that the maximum score possible on the *SRT* is 12, so there was no concern of a ceiling effect. If additional peer review experiences will further increase scientific reasoning, the *SRT* is a plausible means of capturing those changes. GPA was not considered as a factor here because while scientific reasoning score does vary significantly with GPA, GPA is also an educational outcome, not an independent factor. Plotting two outcomes against each other provides little information except that good students score better on tests than do poor students.

Summary of results for Study 7: Relationships between peer review and Scientific Reasoning Test scores.

In conclusion, even using a relatively insensitive and contextually removed tool such as the *SRT* multiple-choice test, the effect of peer review is apparent. The most statistically significant gains and the greatest overall scores occurred when biology majors were categorized by the number of peer review experiences in which they had engaged. Specifically, two peer review experiences produced larger gains and scores than did three to four years of university coursework regardless of the institution at which that coursework occurred. It should be noted that there is considerable room for improvement in test scores however (two peer review experiences producing an average score of 6.82 on a 12 pt scale).

Study 6:
Reliability of scores given by graduate teaching assistants in natural grading conditions

Besides its utility as a measurement instrument, the Universal Rubric has potential to benefit students and instructors in the classroom as well. Students learn best when expectations are made clear and are consistent over time and rubrics can

specifically aid student learning in this way (McNeill, Bellamy et al. 1999). More to the point, however, science graduate students often receive little support for their teaching and little training on pedagogical issues such as grading in particular (Boyer Commission, 2001; Davis & Fiske, 2001; Luft, Kurdziel, Roehrig, Turner, & Wertsch, 2004). Use of a standardized Rubric would both provide consistency of expectations for students across multiple courses within a curriculum as well as save graduate teaching assistants the work of developing their own grading schema. Given the common lack of attention to graduate students as instructors, one of the additional questions asked by this study included, "What is the natural reliability of grades produced by the graduate teaching assistants? Are there any factors (besides training) which seem to improve grading consistency?"

Consequently, a parallel study to the first reliability study was conducted with an additional eight graduate students who did not participate in the first reliability study, nor receive explicit training on the rubric. Using the same student papers as from the first study, these untrained graduate teaching assistants represented a sample of the natural conditions under which grading of laboratory reports occurs. They were provided with a list of the Rubric criteria and the point scale and asked to score papers as if they were laboratory reports in the representative courses. Each of the graduate students involved in this comparison study had previous teaching experience in the relevant course, however. Thus, while not receiving any explicit training on the use of the rubric as occurred for the reliability study, the raters were familiar with the assignments and any rubric criteria which were already incorporated into the course assignments in which they had experience.

The untrained, natural raters had demographic similarities to those in the first reliability study including inexperienced and experienced teaching assistants in each group (see Table 3.6 for a list of rater characteristics in each study, note differences in 301 – Natural rater group only had two members, neither of them inexperienced). The primary difference between the two types of raters was that raters in study 2 (reliability of the Universal Rubric) received the Universal Rubric and Scoring Guide which contained examples of student work at each performance level as well as five hours of training using multiple exemplar papers and discussion until raters came to consensus on the meaning and distribution of criterion scores. The natural, untrained raters received support similar to that provided to graduate teaching

assistants when they actually taught in the courses. Specifically, 10 minutes of verbal instructions as the goals and means of the task and a list of criteria.

Table 4.12. *Effect of a Few Hours Training on the Reliability of Scores Given by Graduate Teaching Assistants*

Course	**Trained Raters**			**Natural Raters**		
	1 rater	2 raters	3 raters	1 rater	2 raters	3 raters
101	0.66	0.79	0.85	0.51	0.68	0.76
102	0.66	0.79	0.85	0.57	0.73	0.80
301	0.66	0.79	0.85	0.68	0.81	--[1]

Note. Papers differed by course (n = 142 papers total), but within a course, trained and natural raters scored identical papers. [1]No third rater available.

Natural grading conditions in this study compared to other published results

In general, graduate students under natural conditions had produced similar (though lower) average and maximum reliability scores as trained raters (Table 4.12). These reliabilities compare quite favorably with the only other published reliability of graduate teaching assistants found in the literature, as well as with published reliability scores in general (compare to reliabilities in Table 4.3). In the only published student reporting reliability of science graduate students as raters, Kelly and Takao (2002) compared the point values assigned for research papers in a university oceanography class and found significant differences in the mean scores awarded by each teaching assistant (ANOVA $p \leq 0.022$, i.e. no correlation among teaching assistants). In addition, when the rank orders of the student papers produced by the graduate teaching assistants were compared with those produced by trained raters using a rubric there was little correlation ($r = 0.12$, Kelly and Takao 2002). The natural reliability of teaching assistants in this study thus appears to be notably higher as reliability scores of $g = 0.76$ and 0.80 indicate that most (76-80%) of the variation in student score was actually due to differences in the quality of student work rather than inconsistencies among raters.

A likely explanation for this finding is that teaching assistants under natural conditions actually received more pedagogical training and support for consistency in grading then did the teaching assistants in Kelly and Takao's study. While not

receiving any explicit training as part of this study, the graduate students who participated as natural raters had all taught in the introductory biology course at least once at some point in the past (see Table 3.6). Graduate students are always assigned to introductory biology for their first teaching assignment because there they receive support and pedagogical training from the faculty laboratory coordinator. Introductory biology teaching assistants are required to attend a weekly meeting typically lasting two to three hours during which they receive support and training for that week's teaching duties. Faculty laboratory coordinators monitor grade distributions and meet with teaching assistants who seem to have exceptionally high or low grading schemes and graduate teaching assistants are exposed to Universal Rubric criterion whenever those criteria are incorporated as part of the course assignments. Thus, our natural raters may have greater experience with applying criteria to laboratory reports than did the teaching assistants reported in Kelly and Takao's study. As the level of support described here for teaching assistants appears to be greater than that reported for many other institutions. For example, in many cases training in how to teach is not even considered in discussions of the quality of doctoral programs (e.g. Mervis 2000; Carnegie Initiative on the Doctorate 2001) or when support provided to graduate teaching assistants is investigated, most are found to work autonomously with little pedagogical support or training (Luft, Kurdziel et al. 2004).

Reliability of individual criteria under natural grading conditions

Looking at individual criteria, criterion reliability maxima were again distributed across courses (Table 4.13) showing no dependence of reliability on subject matter. Methods: Controls and Results: Statistics for BIOL 102 also posed challenges for Natural Raters due to universally low student performance (same situation as described for Trained Raters section on low criteria reliabilities). One notable difference was that BIOL 101 Natural Raters were able to successfully apply the Methods: Controls criterion and generated a reliability score of $g = 0.74$. No description of the successful interpretive framework used by Natural Raters was available. Methods: Controls was re-written based on Trained Rater feedback to address the difficulties in interpretation suffered by five of the six groups of raters. It should also be noted that Hypotheses: Scientific Merit which was one of the more successful criteria for trained raters, had a lower reliability under natural conditions

(average reliability of $g = 0.42$). This result suggested that the explanatory descriptions of student performance at the various scoring levels was necessary for this criterion, or else it was prone to excessive variation in interpretation.

Table 4.13 *Reliability (g) Scores for Individual Criteria under Natural Conditions*

	Course			Overall
Criteria	**101**	**102**	**301**[1]	**Ave**
Introduction				
Context	0.48	0.67	**0.70**	0.62
Accuracy and relevance	0.50	**0.64**	0.55	0.56
Hypotheses				
Testable	0.57	**0.60**	0.49	0.55
Scientific Merit	**0.45**	0.38	0.42	0.42
Methods				
Controls	**0.74**	0.00[2]	0.00[2]	0.25
Experimental Design	0.67	**0.84**	0.60	0.70
Results				
Data Selection	0.25	**0.61**	0.41	0.42
Data Presentation	0.31	**0.72**	0.61	0.55
Statistics	0.27	0.06	**0.52**	0.28
Discussion				
Conclusions based on data	0.48	0.38	**0.49**	0.45
Alternative explanations refuted	0.54	**0.66**	0.38	0.53
Limitations	0.62	**0.62**	0.39	0.54
Implications / significance	0.47	0.50	**0.55**	0.51
Primary Literature	**0.83**	0.57	0.76	0.72
Writing Quality	0.62	**0.67**	0.65	0.65
Total Score	0.76	0.80	**0.81**	0.79

Note. Student papers and samples sizes are identical to that of Trained raters (Table 4.2). Maximum reliabilities per criterion are highlighted in **bold**. [1]BIOL 301 reports two-rater reliability scores. [2]See section on low criteria reliabilities, the same situation described for trained raters applies for natural raters for BIOL 102 and 301.

When comparing trained and natural raters, in general natural raters had lower reliability scores (Figure 4.8). Reliabilities were similar for some criteria (e.g. Introduction: Context and Introduction: Accuracy, Methods: Experimental Design,

Primary Literature and of course the Total Score while other varied noticeably (e.g. Results: Statistics). In only two instances did natural raters generate reliabilities higher than trained raters: Writing Quality and Methods: Controls. It should be noted that these were two criteria that seemed to pose difficulty for trained raters. Natural Rater reliabilities for Writing Quality were higher (averaging to 0.65) and less variable than those produced by trained raters (Figure 4.8).

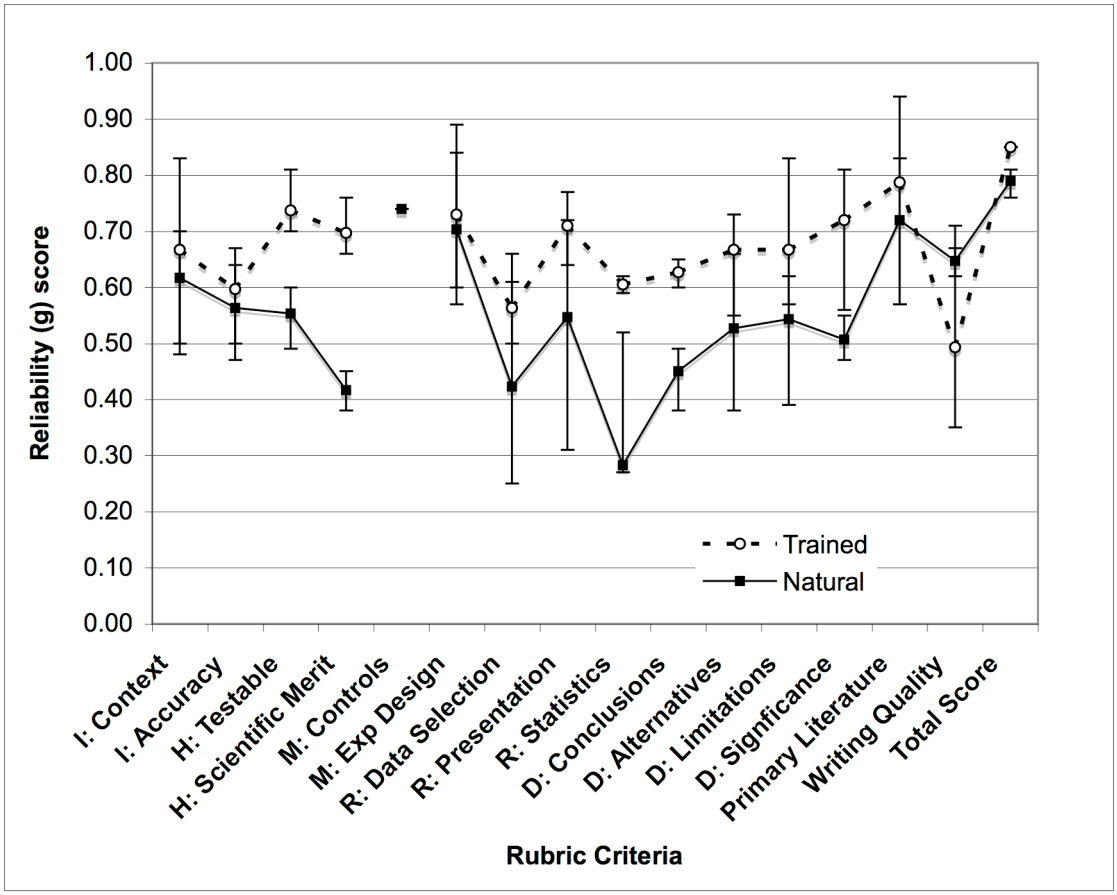


Figure 4.8. Comparison of the reliability scores of Trained vs. Natural raters for individual rubric criteria. Data points are the *average* three-rater reliability across all three courses (n = 142 papers) except for M: Controls which is single data point from BIOL 101: Natural Raters only (reliability = 0.74, n=49 papers). The top of each bar indicates the *maximum* reliability achieved by that type of rater. Lower bar indicates the *minimum* reliability (bars are not standard error bars).

Comparison of the stringency of natural vs. trained raters

One evident difference does exist between trained and natural raters in the number of points each tends to award per criterion. Trained raters were much more stringent on average than natural raters (Table 4.14, Figure 4.9). Out of the maximum total score that could be earned (45 points), natural raters awarded an average of 9.1 more points than trained raters, nearly doubling or more than doubling

scores depending on the course (Table 4.14). Natural raters were also more variable having an average standard deviation that was more than twice that of trained raters and double the range in score (Table 4.14). The higher overall scores for the natural raters may be at least partially caused by these graduate students using a more grade-like mentality. In other words, when grading, the average student is expected to earn approximately 70-75% of the possible points and the range of scores between excellent students and extremely poor students is only approximately 40% (e.g. the span between an "F" and an "A" is usually 60% to 100%).

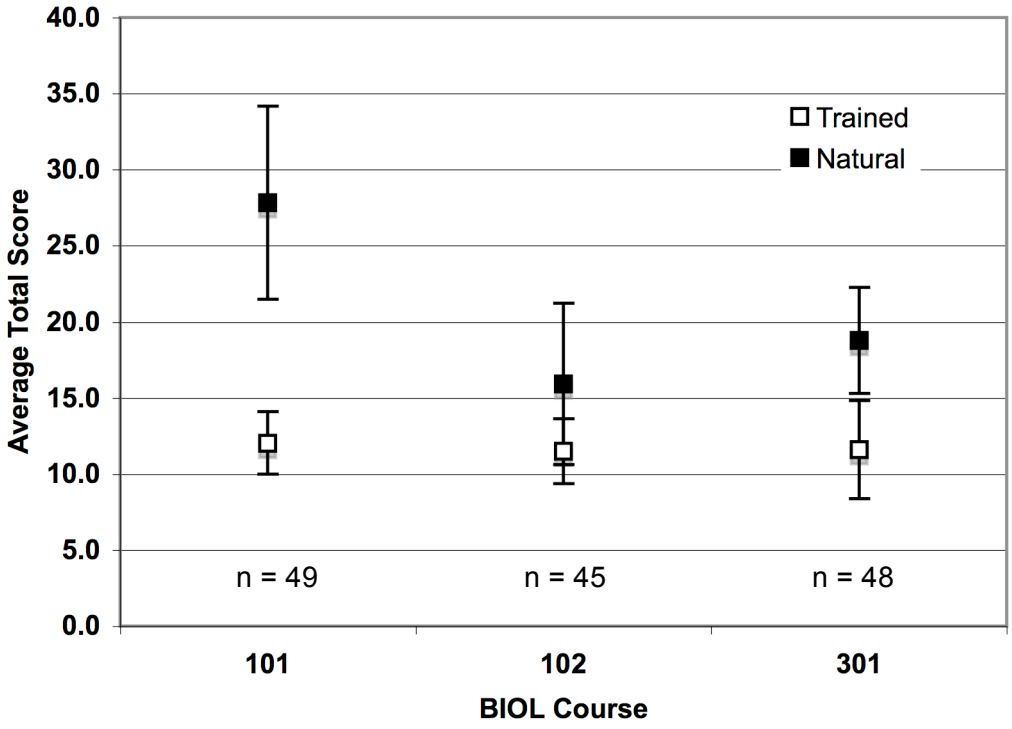


Figure 4.9. Comparison of the stringency of Natural vs. Trained raters (average total score ± standard deviation). Sample sizes are the number of papers scored per course. Three raters per group with the exception of the 301 Natural raters group which only had two raters. Maximum total score possible was 45.

By comparison, scores generated by the Rubric study are absolute scores based on criteria rather than relative scores (e.g. grades). For natural raters grades are usually assigned relative to other students in the same course rather than being based on absolute criteria. Thus, without explicit training, it may have been challenging to for natural raters to change their perspective and use the absolute scale required by the rubric. With scoring, novice students who performed well, might still only earn 30% of the available points which differs a great deal from the percentage they would earn as a grade in the class.

Table 4.14. *Variability of Scores Awarded by Trained vs. Natural Raters.*

Course	Ave total score ± Ave Std Dev		Average Range of Total Scores		Ave Score per criterion ± Ave std dev	
	Trained	Natural	Trained	Natural	Trained	Natural
101	12.0 ± 2.1	27.8 ± 6.4	3.9	12.0	0.8 ± 0.2	1.9 ± 0.5
102	11.5 ± 2.1	15.9 ± 5.3	4.0	10.0	0.8 ± 0.2	1.1 ± 0.4
301	11.6 ± 1.9	18.8 ± 3.5[1]	3.7	4.9[1]	0.8 ± 0.2	1.3 ± 0.3[1]
Ave.	**11.7 ± 2.0**	**20.8 ± 5.0**	**3.9**	**9.0**	**0.8 ± 0.2**	**1.4 ± 0.4**

Note. The maximum score possible was 45. Range was calculated by subtracting the smallest total score awarded by an individual rater from the largest to indicate the degree of variation per student among the three raters. Similarly, the average standard deviations reported are the standard deviation in total score among the three raters per student averaged over all the papers in that course for that type of rater. [1]Only two raters in this group.

The greater variability in natural rater scores and consequential lesser reliability are likely realities of the research-oriented university classroom. Most university science departments are not able to provide pedagogical training for graduate students, especially calibrated training on how to grade, despite their ubiquitously role as undergraduate science laboratory instructors (Boyer Commission, 2001; Carnegie Initiative on the Doctorate, 2001; Luft et al., 2004). This study did not address the reliability of science graduate students grading in the absence of standardized criteria, so it is possible that the use of the list of criteria alone improves reliability.

Summary of results for Study 6: Reliability and stringency of graduate teaching assistants in natural conditions.

Without explicit training on grading with the rubric, graduate teaching assistants in this program (which provides more pedagogical support than many) are more lenient and slightly more variable. Their reliabilities are only slightly less than those for trained raters however (g = 0.76 to 0.81). The generalized pedagogical training in introductory biology appears to have provided these graduate students

with the ability to be reasonably reliable in their assessments of student performance. An additional few hours of training did further improve that consistency. The graduate raters for this project were self-selected volunteers. No assessment was made of their teaching abilities in comparison to the graduate student population at large.

Study 8:
Graduate teaching assistants' perceptions of the Universal Rubric

Graduate student teaching assistants' perceptions of the Rubric were surveyed anonymously immediately after the completion of the Rubric training and scoring sessions. Because graduate teaching assistants' perceptions of the utility of a tool are likely to impact the effectiveness of that tool and because the feedback was gathered as an exit survey for the training, these perceptions are presented here rather than in Chapter 5.

Most raters found that the concreteness and specificity of the rubric made scoring easier than grading without a rubric.

> *It highlights several categories that are expected in scientific writing and allows for fairly easy and unbiased assessment of whether students are competent in these areas across their academic years.*

> *Straight forward; Very well organized/formatted document - manageable & efficient*

They also often felt that training was useful and that it would be beneficial for science departments to provide such training to their teaching assistants (TAs).

> *TA orientation should have at least an hour dedicated to working with and calibrating with the rubric. A must if scientific writing is to be a major objective of the department.*
> *Absolutely [training] should be given to new TAs. Specific instructions will help them grade more consistently - as in how to handle specific errors, specific misconceptions, etc.*

Graduate student raters overwhelmingly indicated that the use of exemplar papers was a key point in the training experience. For example,

The practice lab reports were very beneficial. Until we looked at what you guys (Sue & Briana) scored [on the exemplars] we weren't too sure of what applied for criteria for example.

Yes! Bad papers were very easy to score but superficially good ones were a real pain and it was surprising to see what scores a "good paper" would get, therefore I trusted the tool even more.

If departments choose to provide some training to graduate students, the use of a rubric and exemplar papers are therefore recommended as minimum components of that training. When asked if they would incorporate elements of the rubric into their own assignments in the future, most graduate students replied positively. Specific comments either indicated that they already did use such criteria or listed specific criteria on which they thought the students should focus. Overall comments wished for more incorporation of rubric elements into departmental courses.

Believe it or not, this scoring experience really makes me wish I TA'ed a writing intensive course! I would love the opportunity to help my students develop into expert writers over the semester and would definitely use this tool to do so.

Suggestions for improving the rubric focused mostly on adding additional criteria for various elements the graduate students thought were missing or for giving greater detail in the rubric about how to handle specific scoring situations. Notably there were no suggestions to shorten the rubric.

Summary of results for Study 8: Graduate teaching assistants' perceptions of the Universal Rubric.

Graduate teaching assistants who volunteered to be the trained raters in the Universal Rubric reliability study found the five hour training to be useful enough that they recommended that all graduate teaching assistants receive similar training. Aspects of the training mentioned as being particularly useful included the scoring of common exemplar papers similar to those which were to be scored in the future and

discussion of discrepant scores to calibrate rater's interpretations of the rubric criteria and expected student performance at various levels.

Summary of Achievement Results

The incorporation of peer review was effective for improving students' scientific reasoning skills and scientific writing. Students were effective and capable reviewers at even the introductory level. Use of peer feedback improved student laboratory reports. Laboratory reports were a rich source of data when investigated with the Universal Rubric for Laboratory Reports. Application of the rubric to longitudinal and cross-sectional portfolios of laboratory reports measured the progression of students in acquiring scientific inquiry skills and highlighted gaps and mis-alignments between assignments and curriculum goals. Repeated exposure to peer review accelerated gains in scientific reasoning beyond that achieved by academic maturity alone. University science departments are thus encouraged to incorporate peer review as an effective pedagogical strategy that benefits students without increasing the grading load on instructors. To assist the reader, the results of this study are concisely summarized in Table 4.15.

Table 4.15. *Summary of Achievement Data Results*

- Undergraduates (even freshman) were effective and consistent peer reviewers whose feedback produced meaningful improvements in final paper quality.
- Peer review of science writing in science classrooms accelerated the development of scientific reasoning skills ($p = 0.000$).
- The Universal Rubric for Laboratory Reports was reliable independent of biological subject mater and improved the consistency of scores generated by graduate teaching assistants.
- A few hours of training on the use of the rubric improves the consistency of graduate teaching assistants even further. Graduate teaching assistants suggested that such training become part of regular teaching assistant training.
- Greater incorporation of rubric criteria into assignments improves student performance. Some criteria required explicit instruction or students did not attempt them (e.g. use of controls in experimental design, use of statistics, use of primary literature).
- Application of a Universal Rubric to assignments in multiple courses is a valuable tool for detecting gaps in the curriculum as well as identifying curricular strengths.
- Greater emphasis on the quality of open-ended written feedback significantly improved student performance ($p = 0.001$) to a larger extent than academic experience or grade point average.

CHAPTER 5
RESULTS OF THE SURVEY OF STUDENT PERCEPTIONS

Overview

Student achievement results were presented in Chapter 4. This chapter primarily reports the results of an online survey of undergraduate student perceptions of peer review and its impact on their scientific reasoning skills. As learners' perceptions of the relevance of an activity to their personal life and future success strongly affects their motivation and performance, information on student perceptions of the purpose and impact of peer review provide insight into the student achievement data. Failure to perceive peer review as a worthwhile activity could noticeably detract from student performance of peer review tasks. If student achievement is less than anticipated, it is important to determine the cause of the poor performance so that pedagogical revisions can be targeted at the actual cause. Consequently, the online survey was developed to assess student perceptions of the purpose of peer review in the classroom and the relationship between the classroom activities and real-world scientific competencies. In addition, the survey probed student perception of the effectiveness of the instructional supports for the process in case further potential improvements were identified.

Overview and brief summary of Survey structure

Perceptions of relevance can have significant impacts on motivation and learning (Bendixen & Hartley, 2003; Mistler-Jackson & Songer, 2000; Osborne, Erduran, & Simon, 2004; Van Berkel & Schmidt, 2000). Consequently, an understanding of student perceptions of the peer review process would provide additional insights to improve classroom implementation. Further, one of the goals of this curriculum included students developing an understanding of the role of peer review in the science community. While a number of studies have suggested students perceive peer review as a positive educational experience (Haaga, 1993; Pelaez, 2002; Stefani, 1994), no extensive quantitative survey has been published despite recommendations by previous authors that such information would be useful (Hanrahan and Isaacs 2001). A Survey was therefore constructed to elicit students' perceptions of the purpose, process and impact of peer review on their learning and

development as scientists as well as their perceptions of the role of peer review within the scientific community.

Students' anonymous opinions regarding the purpose and impact of peer review were solicited from introductory biology students over the course of two semesters, corresponding with their first and second engagements in peer review. Their responses were overwhelmingly positive. As described in Chapter 3, the Survey had a high response rate (85.5%) and contained four subsections: 1) statements concerning students understanding of the *purpose* of peer review (items A-F in Table 5.1); 2) statements concerning their understanding of the *process* and mechanics of peer review (items G-O, Table 5.1); 3) statements concerning the *impact* of peer review on students' papers and future courses (Q-AC, Table 5.1); and 4) open-ended questions about the rationale for peer review in the class, the role of peer review in professional scientists' work and suggestions for change. Components 1, 2 and 3 were statements to which students responded on a Likert scale of 1 to 6 with 1 being "strongly disagree" and 6 being "strongly agree." A number of items (X-AC, Table 5.1) were also added for the spring administration. The added items probed for greater detail concerning what aspects of learning were affected by the process of peer review. Two items addressing any effects of multiple peer review experience were also added for the spring administration. Identical versions of the survey were administered to both BIOL 101 and 102 in the spring.

The distribution of student responses to each item was reviewed. The responses: *slightly agree, agree* or *strongly agree* were deemed to be positive and those percentages were summed for each item and reported as the *% of positive responses*. Positive response rates were tabulated separately for each course. Given the small standard deviations among *% positive responses* in the different courses, the positive response rates were averaged over all three courses for each item and reported in Table 5.1.

Table 5.1. *Average Percentage of Students' Positive Responses Regarding the Impact of Peer Review Across Three Introductory Biology Courses.*

(#) Survey Component	Survey Item	Total # responses	% Positive Responses ave ± SD
(1) I understand the Purpose of:			
	Peer Review in this class (A)	1006	93 ± 3
	Peer Review for Scientists (B)	1003	94 ± 0
	Of Calibration Papers (C)	1004	89 ± 5
	Of Receiving Feedback (D)	1003	93 ± 3
	Of Giving Feedback (E)	998	94 ± 2
	Of Self-assessment (F)	1004	91 ± 3
	Average	**1003**	**92 ± 3**
(2) My teaching assistant provided:			
	Rationale for in-class use (G)	1001	87 ± 8
	Future usefulness (H)	999	83 ± 8
	Use of Peer Review by Scientists (I)	1000	84 ± 8
	How to use peer review system (J)	993	92 ± 6
	Training for *CPR* was adequate (K)	999	84 ± 10
	I was motivated to do Peer Review (P)	1001	65 ± 5
	Average	**997**	**83 ± 9**
	Handout readable (L)	993	91 ± 6
	Criteria readable (M)	1000	88 ± 2
	CPR website user friendly (N)	997	86 ± 1
	Time required manageable (O)	996	87 ± 2
	Average	**996**	**88 ± 2**
(3) Peer Review improved my:			
	Laboratory report (Q)	440[1]	79[1]
	In-class understanding (R)	993	73 ± 5
	Work in other courses (S)	997	75 ± 7
	Writing skills (T)	995	67 ± 5
	Editing skills (U)	998	81 ± 5
	Critical thinking skills (V)	998	71 ± 5
	Research skills (W)	997	69 ± 5
	Average	**917**	**73 ± 5**
Because:			
	Calibrations were useful (X)	558[2]	83 ± 2
	Self-assessment was helpful (Y)	558	81 ± 6
	Feedback received was helpful (Z)	559	80 ± 5
	Feedback quality was satisfactory (AA)	557	69 ± 3
	Giving feedback made me think (AB)	557	86 ± 6
	I gave quality feedback to others (AC)	557	95 ± 2
	Average	**558**	**83 ± 9**
Multiple Peer Review Experiences			
Peer Review less difficult the 2nd time (AD)		303[3]	83
Peer Review more useful the 2nd time (AE)		303	64

Note. Survey items are abbreviated here (see Appendixes 7 and 8 for further detail). Courses surveyed were BIOL 101 Fall 2006 and BIOL 101 and 102 in Spring 2007 (total n = 1026 students). [1]Item Q was asked BIOL 101, Fall 2006 only. [2]Items X-AE were asked in Spring 2007 only. [3]312 students in BIOL 102 reported that they participated in peer review in BIOL 101 in Fall 2006. Of these, 303 responded to items AD and AE.

Sample independence

The total number of respondents over all three courses was 1026, but the number of responses varied slightly per item as not all respondents completed all items (Table 5.1). Variation in sample size among items was never more than 3% of the relevant number of respondents however. Overall, the trends were quite similar for all three courses. Only approximately 26 of the students (2.5% of the total sample) enrolled in BIOL 101 in the Spring 07 semester, identified themselves as having been enrolled in BIOL 101 in Fall 2006. Three hundred and three (303) of the 376 students (80%) who responded to the BIOL 102 Survey identified themselves as having been enrolled in BIOL 101 and in peer review the previous semester. As the peer review experiences were distinct each semester and only 38% of the BIOL 102 students remembered having taken the Survey the previous semester repeat administration of the survey was therefore not considered to be an issue of concern. Sample sizes for items Q and X through AC are approximately half those reported in other sections because those items were only included for a single semester.

All quotes reported in this chapter were collected anonymously from students via open-ended text boxes in the Fall 2006 administration of the Survey. Therefore attributions are not provided for individual quotes.

Study 9:
Undergraduate perceptions of peer review in the classroom

Student perceptions of the <u>purpose</u> of peer review in the classroom:

Contrary to the anecdotal reports received from students who came to the researcher's or other instructor's offices seeking help, the majority of students reported that peer review was beneficial and worthwhile whether viewed on a course basis or cumulative basis (Figure 5.1, Table 5.1). In particular, students reported that they understood the purpose of peer review, both within and outside of the classroom (positive response average for this section over all three classes = 92%, average n = 1003, Table 5.1). Students generally considered the process of peer review to both improve their coursework and their general critical thinking skills. For example:

After filling [the online survey] out, I realized that peer review had helped me more than I thought. My researching skills have improved as well as my thinking skills. I actually paid close attention to the advice the other students gave me and it was very helpful it correcting my paper. I believe that peer review thoroughly works well and that it should be used more often.

When students were asked directly "why [they] thought we asked them to do peer review in this class," more than half (55.6%, n = 444) of students' responses were best categorized as perceiving peer review as a mechanism to improve their laboratory reports. They specifically identified peer review as improving their writing and editing skills.

I think you asked me to use peer review in order to develop my writing and editing skills. Peer review was used to help each other give useful tips in writing our lab reports.

We did peer review in this class to allow us to see the mistakes in our first draft and be able to make changes before we handed them in.

Interestingly, 11% of students who believed the major purpose of peer review was to improve their laboratory reports felt that learning from other people's perspectives was the primary mechanism by which the improvement happened.

You asked us to use peer review in this class because you wanted us to get a sense of what other people were writing so we could add to our papers and also get a better understanding of how to write a lab report.

To view how other[s] interpreted the same experiment, to widen our knowledge of the experiment, and to observe others' opinions.

Nearly twenty percent (19.6%, n = 444) of students believed that instructors were asking them to do peer review because peer review was a useful science skill in and of itself rather then just a means to improve one's grade on a laboratory report.

[You asked us to do peer review in class] to help us to begin to understand the process of peer review that allows scientific studies to

be vetted to prevent shoddy work or bias to slip through, and to develop the skills needed for peer review.

Most people in the class are planning to become scientists or engineers in a scientific field. It will be useful later, because we will eventually have to do peer reviews in our career paths.

We were asked to use peer review in order to help us understand the usefulness of having your peers review your work and how it has helped the scientific community.

If we are going to grow up to be research scientists, we will need to know how to give and take peer review.

This perception of peer review as being a useful skill relevant to a student's future in 20% of the class may actually be quite notable as only 35% of the students enrolled in the course (and taking the Survey) were declared biology majors (196 of 562 students). Close to 43% of students enrolled were declared Pharmacy or Exercise Science students whose curricula would not include any further biology courses. If the *assumption* is made that only biology majors would perceive the research skills taught in introductory biology to be relevant for a future career as a scientist, then a large proportion of majors took this broader view of the purpose of peer review in the class. As the Surveys were anonymous however, there was no conclusive way to determine if biology majors in particular perceived peer review as a broadly useful scientific skill. The remaining quarter (24%) of students (n = 444, Fall 2006) was composed of various miscellaneous beliefs. Five percent (5%) of students believed that the purpose of peer review in the classroom was to provide opportunities to learn how to edit writing while another four percent (4%) thought it was to increase their understanding of the assignment. Negative comments were expressed by 2%, miscellaneous comments by 6% and 7% of students did not respond to this query.

For the Spring 2007 administration, these open-ended responses were coded into six categories. The Survey was revised so that students were asked to "select the top three reasons why we asked you to use peer review in the classroom. " Rank order and percentages of the selections were similar between BIOL 101 and 102 for

the top three choices. Students clearly believed that peer review was selected as a pedagogical tool because of its role in the scientific community (76.5%, n = 558 Table 5.2) as well as to improve laboratory reports (77.1%, n = 558). The third most popular rationale (63.6%) was "to learn to critique scientific work" further indicating that students viewed peer review as a functional skill rather than a purely in-class process (Table 5.2).

Table 5.2 *Top Three Reasons Why Students Believe They Were Asked to Use Peer Review in the Classroom.*

Reason	101 (n = 187)	102 (n = 371)	**Combined (n = 558)**
To receive feedback to improve our laboratory reports.	77.0%	77.1%	**77.1%**
To learn the importance of the peer review process in science.	73.8%	77.9%	**76.5%**
To learn to critique scientific work.	59.4%	65.8%	**63.6%**
To improve our ability to communicate through writing.	40.1%	37.2%	**38.2%**
To increase comprehension of the laboratory assignment.	40.6%	28.0%	**32.2%**
To correct grammar and similar mistakes.	8.6%	12.4%	**11.1%**
Other	0.5%	0.8%	**0.7%**

Note. Students were asked to select their top three choices, so percentages do not sum to 100%. Sample size is the number of students who submitted responses in Spring 2007.

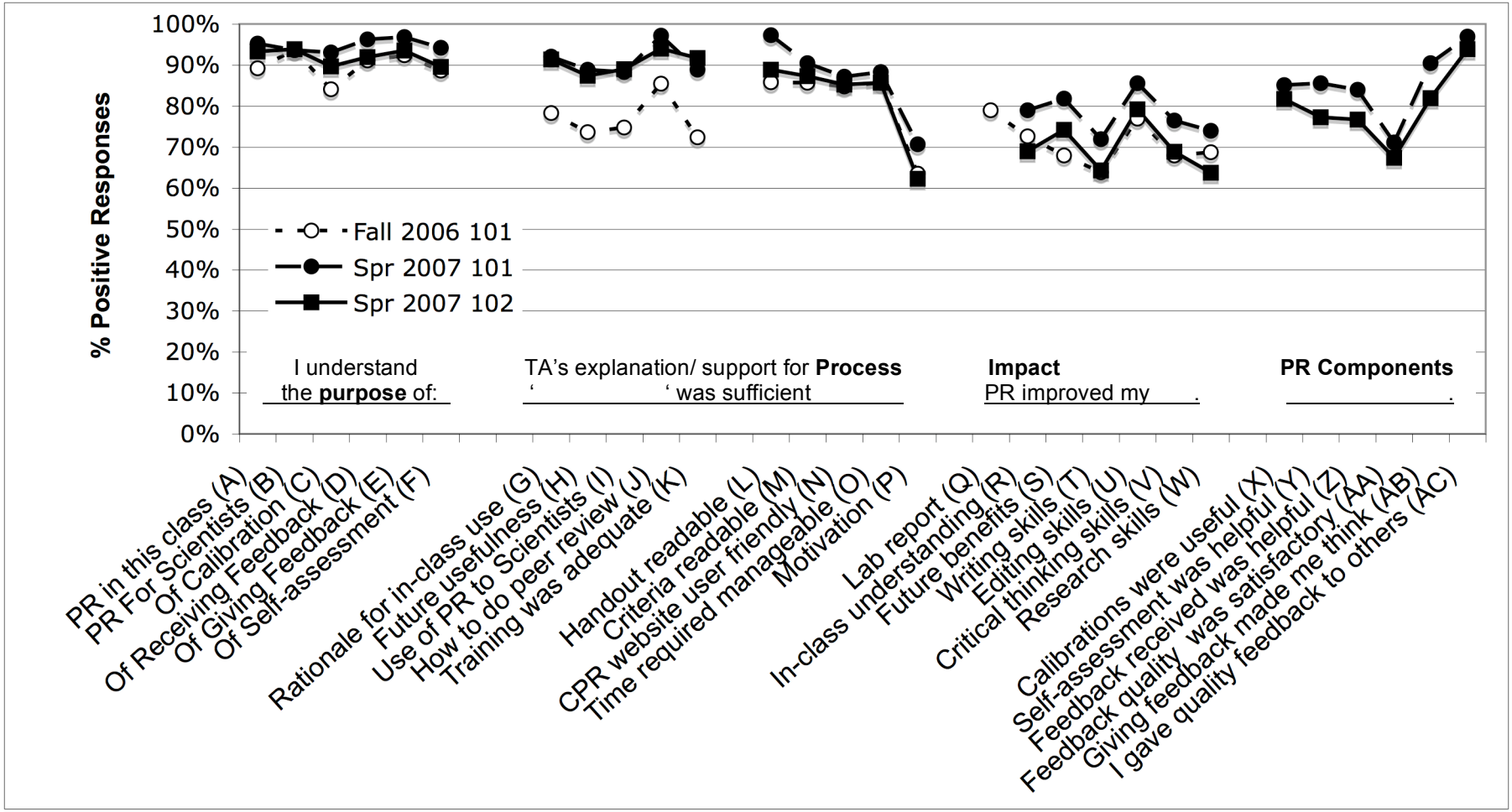


Figure 5.1. Student perceptions of the role and impact of peer review by Introductory Biology course and term (n = 1026 students). Percentage positive response is the cumulative % of students who responded *slightly agree, agree,* or *strongly agree.* PR = peer review. Full Surveys in Appendixes 7 and 8.

Student perceptions of the process of peer review in the classroom:

Students considered the support provided to them for engaging in the process of peer review to be effective (average percentage of positive responses was 85%, n = 998). For example, student comments included statements such as: "My TA's do a wonderful job explaining how to do CPR procedures, why we should find peer reviews important, and how to scientists use peer reviews in real life." Student perceptions of teaching assistant explanations improved from the Fall to the Spring semesters becoming more positive and less variable (Figure 5.1). This was the only noticeable difference between the three courses. While there were slight wording changes to improve the clarity of these items (G-K, Table 5.1) between administrations, it is more likely that the improvement was due to the teaching assistants being more experienced in the spring semester. Only 4 of 15 teaching assistants in the spring semester were new to *Calibrated Peer Review (CPR)* compared to 9 of 15 teaching assistants in Fall 2006 semester. In general, students reported that they received sufficient explanation and support from their teaching assistants and that the handouts and other instructional materials were useful. When asked "what changes would you recommend to improve peer review in this class?" 23.4% of students (n = 444, Fall 2006) said that no changes were necessary and an additional 9.5% provided positive comments in the "additional comments" section.

> *I enjoyed the peer review system because it really helped me to revise my paper and fix its weak points. I believe this is a helpful tool (especially for freshman).*
>
> *I like the way [peer review] is set up in this class because it is all annonymous, [sic] so I wouldn't change anything about it. Plus, it is very simple to use and give good descriptions for each step.*
>
> *I thought it was helpful, so I probably would not change much.*

The largest proportion of students' suggestions for change (31.1%) actually requested increasing student involvement in peer review (n = 444 students, Fall 2006). The largest single category within this group (comprising 12.4% of the total 444 respondents) wanted more thorough training or more calibration papers.

To improve the peer review systems, I would recommend that the idea behind the assignment be taught thoroughly so that students do their reviews and learn from it rather than strictyly [sic] going through the motions and not caring anything away from it.

Changes that I would recommend to improve peer review in this class would include providing a more [sic] clearer examples about primary literature, and how it should be incorporated into the lab assignment.

Another portion of this group (8.3% of the total) wanted to improve the quality/increase the level of detail of the peer feedback they received; 6.3% wanted to make peer review a face-to-face process in class (often to improve accountability) and 4.1% wanted more opportunities to do reviews or to receive reviews: "I would like to have more than three opinions on the papers that I write."

Considerably fewer students wanted less involvement with peer review. Slightly more than five percent (5.6% of 444) wanted to reduce the time spent of peer review ("It takes an innane [sic] amount of time to peer review an entire paper") and 4.3% thought that peer review was not necessary. An additional 4.3% wanted changes to the grading system. The remaining third of students suggested changes which are not within our control: changes to the *CPR* website (10.4%) or gave comments which were too few to form categories or were off topic (10.1%) or gave no response at all (10.8%);

One notable low point in the quantitative Survey was the degree to which students felt motivated to engage in the assignment. Despite their previously articulated understanding of the value of peer review, only slightly more than half (63%, n = 1001) said that they were motivated to do the assignment. As no comparative measure of their motivation to accomplish any other assignment was made, this percentage could be quite high relatively speaking (given anecdotally perceived levels of general student motivation), but lacks sufficient context to be more fully interpreted. This result was consistent over all three courses (Figure 5.1).

The other notable area of complaint (which was also common in anecdotal reports) was that peers did not provide high quality feedback. Investigation into this issue indicates that only slightly less than one-third of students actually (31%, n = 557) felt that the quality of feedback that they received was unsatisfactory (Item AA,

Table 5.1). It should also be noted that while 31% of students felt they received poor quality feedback, 95% of students reported that they provided high quality feedback to others (Item AC, Table 5.1) however. As previously discussed, introductory biology students provided useful feedback, but did not do so 100% of the time. Six of 10 feedback items per reviewer were likely useful at best for any given writer (3.7 average plus one standard deviation of 2.6). Approximately one third of feedback provided by peers was not considered helpful, thereby supporting students' perceptions that the quality of feedback could be improved. Given the discrepancy between 31% reporting that they received unsatisfactory feedback, but only 5% reporting having given lesser quality feedback, some students apparently were not cognizant or accurate in their assessment of the quality of the feedback that they personally provided to others. While there was clearly a perception that the quality of the feedback could improve, 80% of students did feel that, "Other students' feedback was helpful to me in revising my laboratory reports" (Item Z, Table 5.1). The distinction between these two likely lies with the word "satisfied." While 80% of student felt they received some useful feedback, 31% perhaps desired a larger quantity of useful feedback.

Student perceptions of the effect of peer review

Students generally perceived peer review as benefiting both their laboratory report and their generalized writing and critical thinking abilities. Specifically, three-quarters of students were quite positive about the direct effect of peer review on their writing, editing, research and critical thinking skills, (73%, n = 917, Table 5.1). Improvement in their laboratory report and their editing skills received the highest positive response (79% and 81% respectively). In-class understanding, their work in other courses and generalized writing skills were also positively affected for 67-75% of the students. Most notably, 69%-71% of students felt that peer review directly improved their critical thinking and research skills (Items V and W, Table 5.1) and provided comments such as:

> *I think that we were asked to use peer review in this class so that our critical thinking skills would be enhanced within the scientific community. It was also useful in helping us develop better grades on the assignment. Reviewing our own work and the work of others*

allowed us to see the mistakes that we made, and mistakes that we should not make.

Peer review was used in this class to help people gain a better understanding of their own writing and their classmates' writing. Students had to utilize their writing, editing, critical thinking, and research skills; therefore, benefiting greatly from this exercise.

In the scienticic [sic] community it is very important to have others review your paper, and this is why it is instilled in us at such the beginign to gte others [sic] input to make your research more accurate and precise.

Peer review helped us understand the process scientists have to go through when publishing a report. It also helped us teach each other and develop our researching skills.

Thus, the majority of students perceived peer review as having a positive impact on both their immediate work as well as broader impacts on their scientific and writing skills. Consequently, it can be concluded that students perceive peer review as a valuable and worthwhile portion of the curriculum.

Student perceptions of <u>why</u> peer review was helpful

Students reported the various components of the peer review process to be roughly equivalent in their usefulness. The exemplar papers (Item X), self-assessment (Item Y) and peer feedback (Item Z) were all rated as beneficial by approximately 80-83% of the students (Table. 5.1).

A small (but notable) percentage of students (7.5%) wrote open-ended responses in Fall 2006 indicating that the process of *giving* feedback to others or viewing others' work was helpful to them in their own writing. This effect was reported both in addition to and instead of, receiving peer feedback from others. Examples of student comments evidencing this opinion include:

When reviewing other peers work, we would also be more inclined to think about ours, which would in return help out our own paper.

This was in order to learn by teaching. By reading and grading other papers, one can easily see what needs done in their own paper.

To be able to give feedback to our classmates which then might give us a better understanding of what's right or wrong in our own papers.

Consequently, these open-ended responses were condensed into a Likert scale item for the Spring 2007 administration (item AB, Table 5.1). While only 7.5% of students volunteered that opinion in the fall, when systematically surveyed 86% of the 557 students agreed with the statement that "[p]roviding feedback to other students helped me in making revisions to my own laboratory report." These results support and expand the one other evident report on the impact of *giving* feedback where 33 graduate students rated the value of reviewing other people's papers as 7.9 $\pm$ 2.3 on a scale of 10 pts (Haaga 1993). The qualitative data reported here shed new light on the mechanism behind this effect however. Student comments such as those reported above mention two important facets in this effect. Firstly, giving feedback appeared to stimulate reflection and self-evaluation as evidenced in comments such as: "when reviewing other peers work, we would also be more inclined to think about ours." Secondly, exposure to peers' work caused students to compare and contrast among works. This led to evaluation and self-evaluation as evidenced by responses such as "by commenting on other students papers and my own I was able to compare and further understand the process of a lab report" and "the activity taught you what to look for in a paper and how to apply it to your own."

The self-reflection caused by reviewing other people's papers was the likely mechanism by which *giving* feedback would be perceived as improving the reviewer's own paper. Self-reflection often leads to metacognition which is an awareness of one's own learning process, or more specifically, the ability to reflect upon, understand, and control one's learning (Schraw and Dennison 1994). Students who specifically mentioned the process of *giving* feedback as being beneficial to their own work have clearly grasped the metacognitive aspects of the process. As indicated by the quotes above, the process of giving feedback caused students to engage in self-evaluation and stimulated metacognition.

Metacognition is a central component of meaningful learning (Wandersee, Mintzes et al. 1994; Bendixen and Hartley 2003) and an important professional

competency for professionals and experts in both science and teaching (Halonen et al., 2003; McAlpine & Weston, 2000; Roche & Marsh, 2000; Sluijsmans, Dochy, & Moerkerke, 1999; Yore, Hand, & Florence, 2004; Zembal-Saul, Blumenfeld, & Krajcik, 2000). Baird and White (1996) define meaningful learning as "informed, purposeful activity to the extent that learners exert control over their approach, progress and outcomes" and indicate four necessary conditions: 1) multiple time periods devoted to the activity, 2) opportunity to reflect (reflection valued as an explicit activity), 3) guidance or feedback which encourages reflection and 4) support in the form of a culture of collaboration. All four components were present in the peer review process. For example, the time from writing the draft to peer review to revision and final version encompasses several weeks of class time. The elements of feedback and self-assessment are specific steps in the *CPR* process. In addition to the reflection which was apparently caused by the act of giving feedback to others, reflection is hard to separate from self-assessment. Peer review thus appears to present a particularly powerful pedagogical tool because of its focus on the higher order skills of comparison, evaluation and its ability to generate reflective thinking.

Student perceptions of the effect of multiple peer review experiences

While the proportion of students who felt that peer review was a positive experience did not vary noticeably among semesters, faculty involved hypothesized that as students gained experience with peer review, the mechanics of the process might become easier allowing students to focus more effort on the purpose and quality of peer feedback. This shift in cognitive load from mechanics to substance might thereby improve the quality of the feedback and the usefulness of the whole process. Therefore, items specifically asking students to comment on the impact of subsequent peer review experiences compared to the first were added to the spring version of the Survey.

In the Spring 2007 administration of the Survey, students were specifically asked how their prior experience compared to the current one. In BIOL 102, most (80.5% or 303 of 376) students responding to the Survey indicated that they had participated in peer review the previous semester. The remaining 20% were likely transfer students who brought in credit for BIOL 101. The BIOL 101 course in Spring 2007 also reported a few students who said they had engaged in peer review

the previous semester (12.6%, n = 206 respondents). When asked specifically about peer review in biology classes however, this percentage fell to 8% (queries #35 and 37 both report 14 to 16 of 203 respondents with past peer review experiences in biology). While it could be expected that students who failed the course the previous term might respond differently, the majority of students repeating BIOL 101 reported the process was less difficult (65%) and more useful (61%) the second time around. In BIOL 102, a larger majority (83%, n = 303) reported that peer review was less difficult the second time and 64% felt that it was more useful (Items AD and AE, Table 5.1). Students were then asked to elaborate on how the 2nd experience differed from the first in an open-ended response.

The majority of students in BIOL 102 (265 of 408) provided some type of open-ended response to the query: "If you have used peer review in earlier classes, please explain the differences between your recent and previous experience." The largest proportion of comments (47%) however either did not clearly distinguish among semesters or reported on logistical differences between the semesters without indicating how those differences impacted the difficulty or usefulness of the experience (e.g. "I did not post my graphs correctly. I was docked points for this." or "The paper was more of a challenge to write.")

Within the relevant comments, positive statements outweighed negative, usually by at least a 2:1 ratio. The majority of relevant responses (71%) indicated that 2nd experience was *easier* because of increased familiarity and understanding of the mechanics of peer review or the *CPR* website (e.g. "The first experience I didn't really know how to use it, but I quickly learned how to work the system. The second experience was a lot easier because I was familiar with the system."). Three-quarters (75%) of the students commented on the *usefulness* of the subsequent peer review experience indicating that the quality of the peer feedback had improved, the student's understanding of the purpose or process of peer review had improved and/or the focus on feedback quality was helpful. For example,

> *This year, we had to comment on every question, whereas last year we only commented on a few questions. It was good to comment on all of them because it was easier to explain your answer.*

The peer review this semester was more structured and the responses came out more helpful because they were more detailed.

I understand about peer review a lot more this semester. I also received better feedback this semester.

In earlier classes I was somewhat confused as to what the reason for using peer review was but taking it now I realize that peer review helps me write better papers and helps me with researching information.

My previous experience with peer review was that I hated it and I thought it was ignorant. Since my recent experience, I realize the importance of peer review.

Better teaching assistant explanations were also cited as improving the experience in the spring semester. The minority of negative comments regarding usefulness cited poor reviewer feedback as the major source of frustration (25% of total comments specifically mentioning usefulness, 5% of the total number of responses received). Two students did report that the 2nd semester experience was less useful because all gains to be made from the process had already occurred in the first semester: "While I understand how to work peer review better due to already using it, I had already found the major faults in my writing style in the previous class as well, and found fewer points of improvement due to this."

Students who said that the experience was more difficult the second time commonly identified the additional requirement of graph uploading as the reason or cited discrepancies in teaching assistant instructions as the major source of difficulty. BIOL 101 assignments use data in tabular form which can be imported directly into the *CPR* website. For BIOL 102, data types require graphs. The *CPR* website does not allow the uploading of images due to server space restrictions. So graphs must be uploaded to the departmental server and linked to student papers by embedding html code within the student's laboratory report. Thus, students are not reporting greater difficulty or frustration with the actual process of peer review, but with a technical work-around step in the process required by the software. Poor communication by teaching assistants is also a problem external to the process of peer review. Thus, the actual process of peer review appears to become easier as

students gain experience with the major reported sources of increased difficulty being technical or instructor-based in nature.

Thus, the benefits of peer review likely increase as students gain experience. Students clearly indicate that the basic mechanics of the peer review process are easier in subsequent experiences due to increased familiarity with procedures and expectations. Students also clearly reported that the quality of the feedback in the second semester was better than in the first peer review experience. Given the differences in the structure and emphasis of the 101 vs. 102 assignments described earlier however this sample cannot not distinguish between improvements in feedback quality due directly to greater student experience from gains in quality due to the 102 assignment's focus on feedback quality.

Summary of results for Study 9: Undergraduate perceptions of peer review in the classroom

Students perceived peer review as a worthwhile activity both because of its positive effect on their classroom work and because they viewed it as a personally relevant skill for developing scientists. Namely, approximately three quarters of students surveyed reported that peer review improved their laboratory report and in-class understanding of the experiment as well as content. They also believed that peer review improved their writing, editing and thinking skills and would benefit them in the future in other courses. More than three quarters of students also specifically reported that peer review improved their research and critical thinking skills and many elaborated on the benefits of peer review to their scientific reasoning skills in their open-ended written responses. Notably, 86% of students reported that the act of *giving* feedback was helpful. Written comments detailed that this beneficial effects was because reviewing other student's work required comparison and evaluation and thereby stimulated self-reflection. Thus, peer review appeared to stimulate metacognition and meaningful learning. Peer review was considered to be an effective pedagogical tool by the students.

For students who had engaged in multiple peer review experiences, frustrations with peer review seemed to decline with repeated exposure. Students attributed the decline in frustration to having gained familiarity with the mechanisms and procedures and because their attention was shifted to providing more substantial and useful feedback ("It was the same, except they were more strict on whether or not

you give genuine feedback to other papers.") Repeated exposure to evaluative tasks was also perceived by students as improving their critical thinking skills, particularly as they pertained to scientific reasoning and detecting poor quality scientific work. This section is perhaps best summarized by one student's comment:

> *Peer review was used in this class to help people gain a better understanding of their own writing and their classmates' writing. Students had to utilize their writing, editing, critical thinking, and research skills; therefore, benefitting [sic] greatly from this exercise.*

Study 10:
Undergraduate perceptions of the role of peer review in the scientific community

One of the major reasons for choosing peer review as a pedagogical tool was its corresponding use in the scientific community. Consequently, student understanding of that connection was probed by both quantitative Survey items and open-ended responses. Ninety-four (94%, n = 1003) of students reported that they understood the role of peer review in the scientific community and 84% (n = 1000) indicated that their teaching assistant's explanation of the role of peer review in science was effective. Many of the open-ended responses on the purpose of peer review in the classroom reported in the previous section already indicated that students understood the real-world significance of peer review by citing it as scientific skill that they needed to learn in order to be functioning scientists.

When asked specifically how they thought scientists used peer review in their own work, students' open-ended responses from Fall 2006 (n = 444) were divided approximately equally among the following categories. Students believed that scientists use peer review:

1) to improve work/correct mistakes in general (21.8%),

> *Real scientist use peer review as a source of criticism of their papers. Every time their work is published in a journal, it is there for the whole scientific community to criticize.*

I would think that multiple scientists check over each others work to make sure there are no errors because otherwise they have problems with what they are trying to accomplish.

2) to ensure accuracy of results and receive approval on methods, findings, conclusions (22.5%),

Real scientists probably use peer review as a method to make sure all there work is correct and understandable. Also, they count on having fellow scientists to tell them the truth about their work (whether it is valid or not, etc).

Peer review will help scientists to find and overcome bias and mistakes they cannot see themselves, to improve studies and ensure that the conclusions drawn are valid.

3) to receive feedback and gain new insights/perspectives from others (20.5%).

I think scientist look at each other's work and research to learn and better th[eir] work. One person might have ideas or theories that could leave [sic] to new discoveries and more knowledge. Scientists tend to build on each other's work to forward their proccesses [sic] of finding out unsolved questions.

They let other scientist read what they have done and get feedback, which lets them know what could be done better the next time. With peer review, there are many more ideas that will be used in development of the paper and possibly lead to new discoverys [sic] by using the feedback.

Less common reasons as to why students thought scientists use peer review were: to improve writing/readability (11.7%), as a requirement for publishing (4.1%), and to encourage replication of their work (4.1%). Other reasons (mostly statements that peer review is just how science is done without explanation or rationale) comprised 5.9% of the total sample and 8.1% of students were unresponsive to this item.

As with student perceptions of the use of peer review in the classroom, these open-ended responses were coded and transformed into an item that asked students to "select the top three reasons why you believe scientists use peer review." When this item was administered in the Spring of 2007, the results were similar to the open-ended responses. Receiving feedback/perspective from others and having one's work evaluated/validated were the most common concrete reasons students selected for why scientists use peer review (Table 5.3).

Table 5.3. *Top Three Reasons Why Students Believe Scientists use Peer Review*

Reason	BIOL101 (n = 187)	BIOL102 (n = 371)	**Combined (n = 558)**
To receive feedback for new opinions, perspectives, and insights.	82.4%	77.9%	**79.4%**
To allow others to evaluate the accuracy of their work.	74.3%	75.7%	**75.2%**
To allow others to evaluate the credibility of their work.	62.6%	70.1%	**67.6%**
To improve the quality of their writing.	32.1%	27.0%	**28.7%**
To correct mistakes in their writing.	28.3%	21.8%	**24.0%**
To allow others to try to replicate their results.	13.9%	14.0%	**12.9%**
As a requirement for publication.	5.9%	12.4%	**11.3%**
Other	0.0%	0.8%	**0.5%**

Note. Students were asked to select their top three choices, so percentages do not sum to 100%. Sample size is the number of students who submitted responses to this query in Spring 2007.

Comparisons between the classroom and the scientific community

Interestingly, students often attributed the same values to peer review in the scientific community as in the classroom.

> *I believe real scientists use peer review for many of the same reasons that our lab class did, but on a deeper level. Scientists probably have other scientists review their work, not just for grammar and content, but perhaps another scientist has more current/updated information that could be added. Regardless, it is a good way for peers in their own field of work to critique reports.*

Real scientists use peer review because of many reasons. Other scientists might know more facts or details about another scientist's paper. Some may know how to rephrase paragraphs for better understanding. Some may know a lot about the subject of the paper and be able to help critic it. There are many reasons why real scientists use peer review for their work, but the most definite answer is to make their papers better.

When compared across similar venues (open-ended responses or quantitative Survey items), students often perceived the functions and values of peer review to be similar in the classroom and for practicing scientists (Table 5.4).

Table 5.4 *Comparison of Students' Perceptions of the Functions of Peer Review in the Classroom and in the Scientific Community*

Function	**Open-ended responses Fall 2006 (n = 444)**		**Select the Top 3 Items Spring 2007 (n = 558)**	
	classroom	scientific	classroom	scientific
To gain feedback (perspective/ insights) from peers to improve the quality of work	42.6	42.3	77.1	79.4
To allow public evaluation/ critique of the quality of scientific work		22.5	63.6[1]	75.2 / 67.6[2]
To learn peer review as a valuable future skill	19.6	n/a	76.5	n/a
To improve the quality of written communication.	11.9	11.7	38.2	28.7 / 24.0
To correct grammatical or similar mistakes in the writing			11.1	24.0
To improve own work by giving feedback or to improve own editing skills	12.0		n/a	n/a
Issues specific to only one context	3.6	8.2	32.2	
Other/no response	10.3	14.0	0.7	0.5

Note. Samples sizes are the number of respondents. Numbers are the percentages of respondents who expressed this opinion. Note "select top 3" items sum to 300% rather than 100% as students selected 3 choices. Some categories in the open-ended responses were collapsed for clarity and better correspondence with quantitative Survey items. [1]This item focused on students "learning how to critique scientific work" rather than the act of actually critiquing it. [2]The first number refers to

evaluation of the accuracy of scientific work and the second to the evaluation of the credibility of scientific work.

Students viewed peer review as improving the quality of a person's work, whether a laboratory report or a publishable manuscript as well as simply improving the quality of a person's writing. Students believed that scientists benefit from the perspectives of others just as they reported that they themselves benefited. They also viewed peer review as serving an important function of quality control in the scientific community and viewed the classroom peer review process as teaching the same evaluative skills as used by practicing scientists.

Differences existed in that some students cited the act of giving feedback as improving their own work through reflection, but this concept was not mentioned for scientists at large.

It should be noted that the frequencies of the open-ended responses should not be construed as a definitive basis for comparison (and thus no comparative statistics were performed). Many student responses contained multiple concepts and many students likely held conceptions that they simply did not articulate in response to the single query item. Responses were categorized by the primary thrust of the comment even if other concepts were mentioned. So the existence of a body of comments on a topic should be taken as evidence that it is important enough to a notable number of students for them to mention it, but it should *not* be assumed that the concept was absent in an inverse proportion of students. Evidence of this multiple views per student can be found in the differing proportions that exist when students were asked to discuss just the primary reason (open-ended query) for peer review vs. when they were asked to select the top three functions (quantitative Survey item). For example, "to increase comprehension of the laboratory assignment" was the reason given only 3.6% of the time in open-ended responses, but selected as a top three reason 32.2% of the time in the quantitative Survey items (Table 5.2).

Summary of results for Study 10: Students' perceptions of the role of peer review in the scientific community

Students largely believed that peer review provided many of the same benefits to practicing scientists as it did to them. Students reported they believed

that scientists use peer review to improve the quality of their work as reviewers help them to find conceptual and factual mistakes, share insights and provide new perspectives. They also responded that a major function of peer review was to ensure the accuracy and credibility of research findings; that peer review functioned as a gate keeper for quality assurance. Lastly, small percentages of students believed that functions of peer review included improving quality and readability of scientists' writing, to stimulate others to research in the same area and as a plainly pragmatic requirement for publication.

A major finding of the previous study was that students found peer review to be equally beneficial to the reviewer as the writer reporting that the act of evaluating someone else's paper caused beneficial self-reflection and self-evaluation. Surprisingly, this major benefit of peer review was absent from students' perceptions of the role of peer review in the practicing scientific community. Students appeared to view themselves as being on a learning curve and cited peer review as an opportunity to improve their critical thinking skills. Perhaps because of this perspective, they assumed that practicing scientists have already culminated in the development of their critical thinking skills and no such corresponding cognitive stimulation of the reviewer would occur.

Summary of students' perceptions of peer review

Students reported that peer review was beneficial in the classroom both for the immediate benefit of improving their lab reports as well for helping them to improve their critical thinking, research and writing skills more broadly. They found both the processes of giving and receiving feedback to be educative and reported that this experience of peer review would benefit them in future classes as well. Students also believed that peer review was a valuable process in the scientific community and helped to maintain the integrity of the scientific process. They reported that engaging in peer review in these classes would help further their development as scientific researchers.

CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

Summary of the study context and problem statement

Engagement in authentic scientific practices such as scientific reasoning and scientific writing are common goals for science curricula, particularly in higher education. The curriculum goals for the Department of Biological Sciences at the University of South Carolina detail these and other desired skills to be developed in biology major students (see Appendix 1). Departmental curriculum review indicated, however, that students had insufficient opportunities to develop these skills to the desired levels necessitating some form of curriculum reform. A rich and varied literature exists detailing the benefits of engaging students in authentic scientific research as part of their coursework and the necessity of providing individualised formative feedback in order for meaningful learning to occur. Such instructional methods can be challenging to enact however in large courses at research-active universities given the limited time available to accomplish the multiple missions of external funding, research, publication and teaching. Thus, this research investigated peer review as a potential mechanism for accomplishing both goals simultaneously without undue burden on the instructor. Peer review is a required competency of practicing scientists, as well as a potential means of increasing student learning, reasoning and writing skills.

This chapter provides a brief review of the theoretical support for peer review as a pedagogy generated in Chapter 2 as well as the results reported in Chapters 4 and 5. The findings are discussed in each subsection. The chapter concludes with an overall summary and compilation of recommendations for how to best implement peer review and provide the greatest opportunities for student growth.

Results of literature review and significance of the study

Past educational research literature and the current social climate within the scientific community suggest that the use of peer review to develop students' scientific reasoning skills may simultaneously overcome the challenges of limited time but desire for substantial development of reasoning skills. A few studies have demonstrated peer review to be an effective pedagogical strategy for learning science

content (Birk & Kurtz, 1999; Crouch & Mazur, 2001; Pelaez, 2002). More importantly, while no direct investigation of the effect of peer review on *scientific reasoning* has been made, peer review contains many elements demonstrated to be effective pedagogical strategies for development of such skills: peer-peer collaboration (Committee on Undergraduate Biology Education, 2003; Crouch & Mazur, 2001; Schwartz, Lederman, & Crawford, 2004), sustained practice (Ericsson and Charness, 1994), formative feedback (Schunn and Anderson 1999) particularly in higher education (Yorke, 2003), multiple contrasting examples (Bransford, Brown et al. 2000), and extensive writing (Connally & Vilardi, 1989; Hand, Hohenshell, & Prain, 2004; Keys, 1999). Meta-analyses have further concluded that all these strategies notably improve student achievement, especially in combination (Schroeder, Scott, Tolson, Huang, & Lee, 2007).

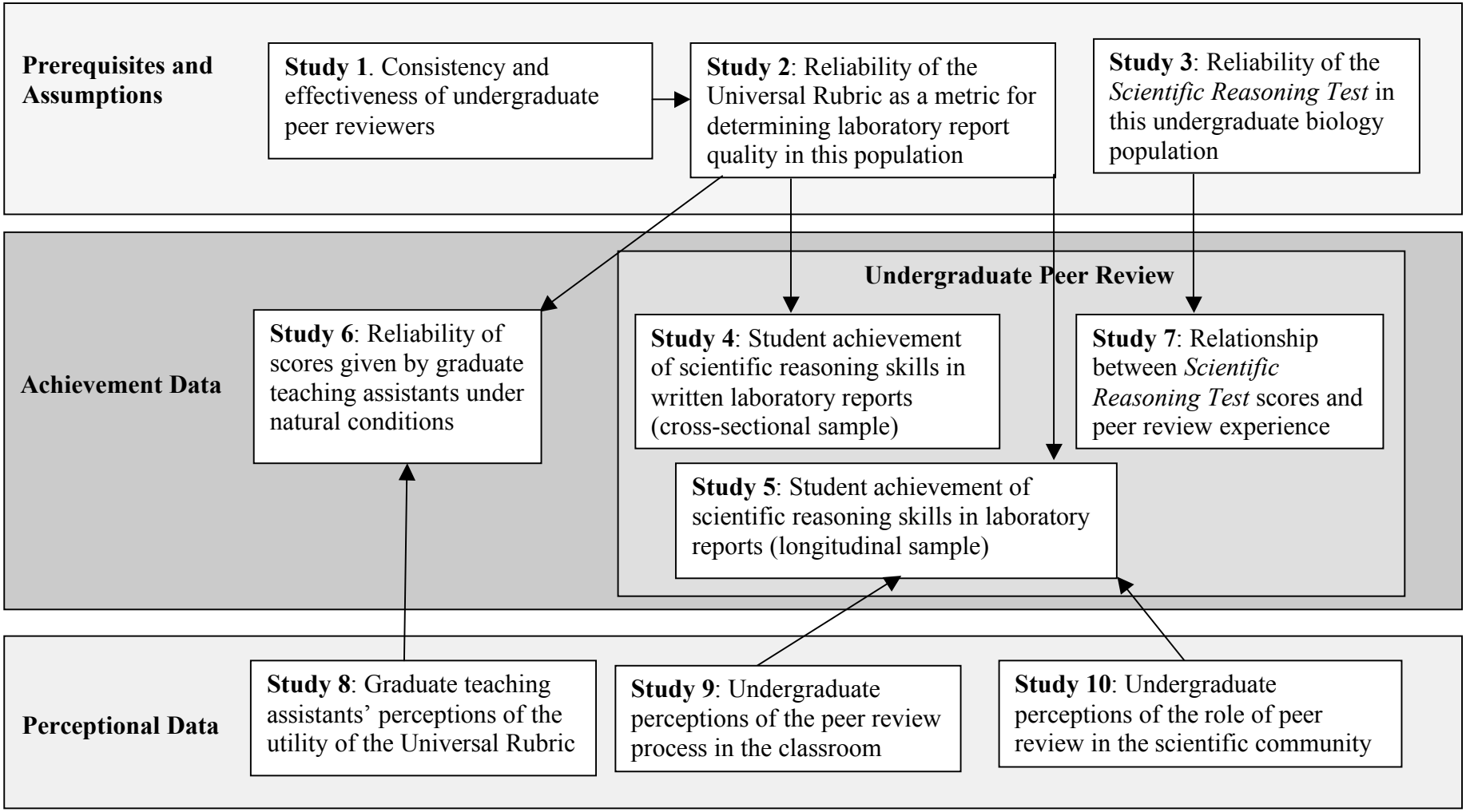
Peer review may also accelerate the development of scientific reasoning and writing compared to other methods due to the self-reflection and self-evaluation it causes as awareness of one's own learning process has been shown to improve learning (Duit & Confrey, 1996; Posner, Strike, Hewson, & Gertzog, 1982; Zohar, 1996). Further, peer review also provides a several-fold increase in the number of opportunities students have to practice evaluative skills while providing three times the formative feedback. These increases in time-on-task and formative feedback are further accelerants for the development of scientific reasoning skills (Ericsson and Charness 1994). Peer review, while little studied in the classroom, therefore shows great promise as a highly effective pedagogical strategy for improving student scientific reasoning skills. Investigation of its impacts will thereby provide useful and novel findings as well as hopefully stimulate innovation in higher education.

Summary of the components of the study

The focus of this study was to evaluate the effectiveness of peer review as a mechanism for accelerating students' scientific reasoning and writing abilities without significantly increasing the time burden on faculty. In addition, the tools and data sources developed for this study also provided longitudinal and cross-sectional windows into the effectiveness of the biology curriculum over the course of students' undergraduate careers. The effect of peer review on scientific reasoning was assessed using three major data sources: student performance on written lab

reports, student performance on a *Scientific Reasoning Test* (Lawson, 1978; Lawson et al., 2000) and a Survey of student perceptions of the roles of peer review in the classroom and the practicing scientific community. Measuring the development of students' scientific reasoning skills is challenging. As scientific reasoning develops over at least a period of several years, measurement tools independent of assignment and course are necessary to track students' longitudinal progress. No suitable metric was found in the research literature so an instrument was developed (Universal Rubric for Laboratory Reports).

Review of published research evaluating scientific reasoning in students yielded support for most of the 15 criteria comprising the Universal Rubric (refer to Table 2.1 for details) with the remaining criteria receiving support from professional peer review priorities. Four criteria had support from both published rubrics or lists of criteria as well as professional peer review (*Introduction provides appropriate context, Experimental Design, Primary Literature* and *Writing Quality*). Several criteria were explicitly mentioned in science education heuristics, but not in professional peer review (*Hypotheses are Testable, Hypotheses have Scientific merit*, *Data selection and presentation, Statistics accurate and appropriate, Conclusions based on data*, *Limitations appropriately discussed*). In addition, review of research on professional peer review indicated that the criteria of *significance* and *methodology* were consensus priorities (refer to Table 2.2). It should be noted that the professional peer review criteria cited here are not a comprehensive representation of the values held by professional referees, but merely the relevant common threads across multiple journals. Marsh and Ball (1989) determined 21 different criteria to have been employed by the professional referees in their study (n = 415 reviewers), but found that variation in referees recommendations as to whether a manuscript should be published or not converged on just four of those 21 criteria (*significance, appropriate to journal's readership base, quality of methodology* and *writing quality*) two of were relevant for undergraduate laboratory reports (*significance* and *methodology*). The criteria developed for the Universal Rubric for Laboratory Reports thus are supported by research in the field of science education and the scientific community at large. The Universal Rubric for Laboratory Reports was reliability tested using biology graduate students as raters and three separate course assignments. An overview of the research design and the relationships among data sources are provided in a reproduction of Figure 1.2 and Table 6.1.



Reproduction of Figure 1.2. Overview of research questions and relationships between studies.

Table 6.1. *Brief Summary of Data Sources and Methodological Details for Each Study.*

Study	BIOL Course, term	Data type	Sample size
1: Consistency and effectiveness of undergraduate peer reviewers	102, Fall 2004	Number of students who complete peer review process Time per review, numerical ratings of draft papers Changes to laboratory reports as a result of peer review	n = 308 students n = 335 reviews of 119 papers n = 22 students' draft and final papers
2: Reliability of the Universal Rubric as a metric for determining laboratory report quality in this population	101, Fall 2004 102, Fall 2004 301, Fall 2005	Laboratory reports scored by 3 *trained* raters for each course (n = 9 raters total). Raters were biology graduate teaching assistants who received 5 hours of training as part of the study.	101 n = 49 papers (genetics) 102 n = 45 papers (evolution) 301 n = 48 papers (ecology)
3: Reliability of *Scientific Reasoning Test (SRT)* in this population	Fall 2005: 101, 102, 301L, 302L and 530, Spring 2005: 102	Fall 2005 courses: *SRT* scores from enrolled biology majors Spring 2005: *SRT* scores from all students enrolled	Fall 2005 n = 548 students Spring 2005 n = 303 students
4: Student scientific reasoning skills in laboratory reports (cross-sectional)	Same as Study 2	Same as Study 2 using the average of the trained rater scores per student	Same as Study 2
5: Student scientific reasoning skills in laboratory reports (longitudinal)	101, 102 and 301L, Fall 2004 to Spring 2007	Laboratory reports from various terms to form longitudinal portfolios for individual students. Includes papers from Study 2 where possible (using the average of the trained rater scores). Papers from additional terms scored by an independent rater.	n = 17 students
6: Reliability of graduate teaching assistants under natural conditions	Same as Study 2	Same papers as Study 2 using similar *natural* raters not explicitly trained as part of this investigation (n = 8 raters total)	Same as Study 2
7: Relationship between *SRT* scores and peer review experience	Fall 2005 sample from Study 3	Students who reported the their previous peer review experiences and who had not failed the class in a previous semester	Subset of Study 3 Fall 2005 data n = 389 students
8: Graduate teaching assistants' perceptions of the Universal Rubric	n/a	Trained raters (biology graduate students) from Study 2	n = 9 raters
9: Student perceptions of the peer review process in the classroom	101 Fall 06 and Spring 07; 102 Spring 2007	All enrolled students who responded to anonymous online survey offered near the end of each semester	n = 1026 students
10: Student perceptions of peer review in the scientific community	Same as Study 9	Same as Study 9	Same as Study 9

Summary of results of each study and discussion

This section contains a brief recapitulation of the major findings from each study reported in Chapters 4 and 5 followed by a discussion of the implications of these results. Corresponding recommendations for how peer review could best be implemented at other institutions are provided.

Consistency and effectiveness of undergraduate peer reviewers (Study 1)

Past studies have reported student concerns regarding the ability of peers to provide productive feedback (Hanrahan and Isaacs 2001). The results of this study and others indicate that those concerns are unfounded. Investigation of introductory biology students' peer review experiences demonstrated that they were capable of engaging in peer review, produced useful feedback for their peers and that the process was a reasonable time commitment for an introductory level course assignment (average of 32.4 $\pm$ 14.3 minutes per review including time to read the paper). Peer reviewers were reasonably consistent (average standard deviation in scores among reviewers of a single paper equivalent to 15% of the total score) and to provided an average of 3.7 $\pm$ 2.6 pieces of useful feedback per review. Each student was thus provided with an average of ten useful pieces of feedback across the three reviewers. Peer feedback was identified as useful both by an external rater and because it produced increases in laboratory report quality. Therefore, even freshman were productive peer reviewers and instructors should not let concerns about ability deter themselves or their students from peer review. It should be noted that there is room for improvement however, in that peers may learn to provide a greater number of useful comments per as they gain experience.

Similarly, Cho, Schunn and Charney (2006) found that their undergraduate peer reviewers produced an average of approximately 3 directive (i.e. useful) idea units per writer. Their students possessed an average of 3.4 years of college however compared with introductory biology students who were three-quarters freshman. It therefore appears that this rate of helpful comments is indicative of beginning peer reviewers rather than academic age. Again, the effectiveness of peer reviewers is therefore likely to increase as students gain experience. Additionally, Cho and colleagues demonstrated that when students were blinded to the source of feedback, they rated peer feedback as equally helpful compared to instructor feedback (no

significant difference based on expertise source ANOVA F (1, 45) = .86, p = 0.36) or criterion (F (2, 90) = .97, p = .38), and no interaction between expertise source and criterion score, F (2, 90) = .69, p = .51) (Cho, Schunn et al. 2006). Thus, students' concern that peers are not effective reviewers appears to be unfounded. The final determinant of the usefulness of peer feedback is its effect on student writing however.

Qualitative investigation of a subset of students (n=22 writers) indicated that when writers incorporated peer feedback into their final laboratory reports on evolution those reports improved in quality. For each individual piece of peer feedback incorporated, final paper score increased by three percent (3%). The average overall gain in score as a result of peer feedback was 28% of the points earned on the rough draft. Peer feedback primarily caused gains in both scientific reasoning (here the consideration of alternative explanations) and content knowledge regarding the mechanisms of evolution. As these results were generated by peer reviewers in introductory biology (mostly freshman), it is plausible therefore that both students' capabilities as reviewers and the benefits of peer feedback would improve with greater peer review experience.

Reliability of the Universal Rubric for determining laboratory report quality in this population (Study 2)

A Universal Rubric for Lab Reports was developed for the purpose of assessing student abilities over time and across multiple biology courses, though it may also have utility in other scientific disciplines. The rubric has 15 criteria organized around the standard format of scientific papers. The reliability of the rubric as a measurement tool was assessed using generalizability analysis (g) and three unique raters for each of three separate assignments generated in three distinct biology courses. Total scores generated by the rubric each had a reliability score of g = 0.85 in these three independent tests (n = 45 to 49 student papers per test, see Table 4.2) indicating that 85% of the variation in scores was due to variation in the quality of student papers and only 15% of the variation was due to rater error or interaction factors. Thus, as reliability did not vary based on assignment, the Rubric appeared to be independent of biological subject area as well as a reliable overall measure of student scientific reasoning abilities as defined by the Rubric criteria.

The reliability of individual criteria varied from $g = 0.16$ to 0.94, though not in any predictable pattern by subject matter. It is therefore recommended that instructors include multiple criteria per assignment and not heavily weight any single criterion score. As indicated above, total scores using multiple criteria were uniformly reliable however at the $g = 0.85$ level. The variation in the reliability of some individual criteria did appear to be based however on the degree to which those tasks were included in the assignment. For example, the use of *methodological controls* or the reporting of methodology at all, the *discussion of the limitations* of the research, or the use of *statistics* all appeared to require explicit delineation in the assignment or else student performance was absent to notably low. In contrast, one criterion appeared to be innate (e.g. that *hypotheses must have scientific merit*) in that reliable scores were produced for this criterion across all three courses even though that criterion was absent from all three assignments.

This variation in performance by rubric criterion may suggest variation in the ease with which students acquire various scientific process and reasoning skills. Some skills may be easier for students to learn and some criteria (such as *hypotheses must have scientific merit*) appear to be obvious to students while other skills such as the inclusion of *controls* in experimental design, the use of *statistics* and consideration of *limitations* of the research appear to require more explicit and focused instruction. It is recommended that instructors identify the curricular goals of interest and the criteria by which they will measure student performance *prior* to the development of the assignment and that all performance criteria of interest to the instructor be explicitly included in the written assignment provided to students. Further, how well instructional supports align with curriculum goals must be considered as a context for interpreting student performance scores. In other words, if assignments do not ask students to perform various scientific skills, students are neither likely to develop those skills over time nor score well on those criteria when assessed at the end of their program. These findings further suggest that communication and coordination among faculty to ensure that curriculum goals are included in course assignments and that expectations for student performance increase at appropriate junctures would make a notable difference in student performance and the achievement of departmental curriculum goals. Thus, student achievement trends, the details of assignments within courses and programmatic curricular assessment were more closely linked than previously appreciated.

Student achievement of scientific reasoning skills in laboratory reports as a result of peer review: Cross-sectional (Study 4) and Longitudinal views (Study 5)

Student performance on written lab reports was assessed across multiple biology courses using the Universal Rubric for Lab Reports. Student performance varied by criterion type and assignment emphases as described above. Performance was higher when assignments focused on peers providing substantive and useful feedback, when reviewers were held accountable for the quality of their feedback and when assignments were more closely aligned with the Rubric criteria. Further, performance declined significantly when peer review did not occur, even though the students in the non-peer review class (BIOL 301L Fall 2005) had greater academic experience (91 vs. < 30 credit hours on average) and higher grade point averages (3.14 vs. 2.71 USC GPA). The distinction among student performance in these different classes was significant for 12 of the 15 criteria (ANOVA $p < 0.001$, n= 142 students total) with introductory biology laboratory reports which had undergone peer review consistently outscoring those collected from a sophomore level (301) course. Thus, peer review elevated the quality of introductory biology laboratory reports to a greater degree than did several years of academic experience (refer to Figure 4.3 for more detail).

Longitudinal views using portfolios of individual student performance over time show no significant trend in total score (n = 17 students). In-depth-analysis indicated highly variable trajectories in student performance suggesting that seventeen students were an insufficient sample for making definitive conclusions regarding longitudinal performance.

Reliability of scores given by graduate teaching assistants under natural conditions (Study 6)

Raters who participated in the reliability study were biology graduate teaching assistants who had received five hours of explicit training on how to use the Universal Rubric for Lab Reports. A second parallel test was conducted using the same student papers but a different set of *natural* science graduate teaching assistants who did *not* receive the five hours training as part of the reliability study. It should be noted that as part of the development process for the Rubric, its criteria were piloted in introductory biology courses for some number of semesters prior to the reliability study. Thus, *all* raters had some experience with the rubric as they had all

taught at least one semester in introductory biology at some point in the past, but the natural raters lacked the explicit 5 hrs of training on the rubric immediately prior to scoring. These natural raters were provided with the same level of support that teaching assistants typically receive when teaching laboratory sections. To the author's knowledge, no other rigorous, controlled evaluation of the grading consistency of graduate teaching assistants has ever been made, despite their ubiquity as instructors in higher education.

Natural raters (e.g. teaching assistants) were slightly less consistent than raters who had received five hours of training (total score reliability of $g = 0.76$ to 0.80 for groups of three natural raters compared to $g = 0.85$ for comparable groups of three trained raters), but their reliability scores were still well within or above reliabilities found in the published literature for comparable rubrics (see Table 4.3). Five hours of training did noticeably reduce the variation in reliability as well as elevate reliability scores across individual criteria (see Figure 4.8) so it is recommended that graduate students receive at least one explicit training session on scoring laboratory reports. It is unlikely that most educational institutions will be able to provide three raters per student paper however. The corresponding expected reliability of a single graduate teaching assistant in this situation was calculated to be $g = 0.65$ to 0.66 across the three courses investigated. This means that the majority (65-66%) of the variation in student scores would be attributable to variations in the quality of student work. This result compares favourably with published reliabilities of trained raters (refer again to Table 4.3 for greater detail) and notably exceeds the reliability of graduate teaching assistants reported by Kelly and Takao (2002). Thus, while ideally 100% of the variation in grades assigned to students would be due to variation in the quality of student work, this result is not achievable even in a research setting with multiple raters. Thus, it is strongly advocated that pedagogical support the provided to graduate teaching assistants in this program be continued as the existing use of rubrics has produced a level of reliability akin to that produced in research settings.

Natural teaching assistants were twice as lenient as trained raters however producing average total scores nearly twice as high (20.8 ± 5.0 points per paper compared to 11.7 ± 2.0 for trained raters (refer to Table 4.14). This leniency appeared to originate in the disparate expectations of grading vs. scoring. Natural teaching assistants were likely thinking from a grading perspective rather than a

scoring perspective. When grading, expectations of student performance are scaled to a relative level appropriate for the course. In contrast, the trained raters were using an absolute scale for which novice students tend to score in the bottom 30%. This discrepancy is therefore appropriate. The rubric thus appears to improve consistency in both scoring and grading by teaching assistants and is recommended for both pedagogical and research use in biological classes. It should be noted however, that this comparison demonstrates that grades are not an appropriate proxy for longitudinal scores. Grades are scaled relative to individual course expectations whereas scores must be assigned on an absolute scale in order to note progress over time.

The departmental policy of requiring graduate students to begin their teaching experience in the introductory biology course with pedagogical support and rubric-based assignments appears to have notably elevated the performance of biology teaching assistants. Namely, departmental teaching assistants produced reliability scores comparable to those published in the literature using trained raters and well above those published for professional peer review (compare 3 rater g = 0.76 to 0.85 to Table 4.3). The only other comparable assessment indicated no correlation among the scores generated by teaching assistants or between teaching assistants and/or the instructors and/or trained raters (Kelly and Takao 2002). Specifically**,** Kelly and Takao (2002) compared the scores given by three graduate teaching assistants grading oceanography laboratory reports using a rubric and found significant differences among the total scores given by each teaching assistant (ANOVA, F ratio = 4.6; $p \leq 0.022$). There was also little correspondence in relative rankings when total scores given by the graduate teaching assistants, the faculty instructor for the course and two trained raters were compared (Kelly and Takao 2002). The two trained raters were highly correlated with each other ($r = 0.80$), but no correspondence existed between their relative rankings of merit and those of the instructor or graduate teaching assistants (Kelly and Takao 2002). The comparably high level of reliability produced by our graduate teaching assistants, regardless of training, was therefore quite notable.

As this benchmark study took place in a comparably sized university with a high quality graduate program (University of California Santa Barbara), this author suggests that the difference in reliability between these two populations of graduate teaching assistants was likely due to the embedded training provided in the

Introductory Biology courses at the University of South Carolina. All biology graduate teaching assist assistants at USC are first assigned to teach in Introductory Biology as their first teaching experience. Therefore, all raters used in these studies had past generalized training and experience in the use of the Universal Rubric criteria as well as generalized pedagogical support focused on fairness and consistency when assigning grades. Thus, this research suggests that the Universal Rubric, when combined with training on its use, improves consistency in scoring to a notable degree. This research does not provide any information on the effect of the Rubric in the absence of training as even the natural raters had significant past experience with using this Rubric. As five hours of training did produce visible improvements in reliability (Figure 4.8, Table 4.12), it is recommended that new adoptions of the Rubric begin with a similar training using exemplars and discussion of discrepancies in interpretation.

Graduate teaching assistants perceptions of the usefulness of the Universal Rubric and the corresponding training on its use (Study 8)

A brief exit survey was given to the nine biology graduate students who participated in the five-hour training on the Universal Rubric as part of Study 2. They reported that the Rubric facilitated scoring by clarifying expectations and benchmarks for the different performance levels. Graduate students recommended that training on the use of the Universal Rubric should be provided to all teaching assistants in the biology department. Graduate students suggested that any such training should include the use of exemplar papers followed by discussion of discrepant scores until all teaching assistants reach consensus as to how the criteria should be applied to student work.

Reliability of the Scientific Reasoning Test (SRT) *in this population (Study 3) and the relationship between performance on the* SRT *and the extent of students' peer review experiences (Study 7)*

The *Scientific Reasoning Test* was found to be more reliable in this population (KR20 = 0.83 to 0.85) than was reported for other undergraduate biology populations whose reliability scores ranged from $\alpha = 0.55$ (Lawson, Baker et al. 1993) to KR20 = 0.79 (Lawson, Banks et al. 2007) (see Table 2.3 for more details).

Additionally, the group mean scores were all mid-range (5-6 points on a scale of 12) indicating that the *SRT* targeted an appropriate level of difficulty for this population.

The *Scientific Reasoning Test* uses mostly non-biological contexts for its questions and scenarios. As performance usually declines when students are asked to apply reasoning strategies learned in the classroom to new contexts (Zimmerman 2000), the scientific reasoning strategies learned in the biology classroom were unlikely to be fully transferred to situations tested by the *Scientific Reasoning Test.* The *SRT* therefore serves as conservative measure of gains in student reasoning due to peer review experience.

A cross-section of biology majors from five different courses (freshman to senior year, n = 389 students) was tested with the *Scientific Reasoning Test* as a means of distinguishing between the effect of peer review over multiple courses and the effect of increasing academic experiences (Study 7). Student scores varied significantly when sorted by academic maturity (total credit hours) (ANOVA, p = 0.011 n = 387). When sorted by number of peer review experiences however, the average scores of students with no peer review, one or two experiences were more significantly different (p = 0.000) than when sorted by credit hours (details of ANOVA results in Tables 4.9 to 4.12). Additionally, the largest gains among groups were found when students were categorized by peer review experiences than by credit hours. The largest group average overall was produced by students with two peer review experiences (refer to Figures 4.5 to 4.7). In sum, engaging in peer review in two different (freshman) courses produced a higher average score than did 120 credit hours of collegiate coursework or 90 credit hours of coursework at this institution in particular. Peer review thus seemed to accelerate the development of students' scientific reasoning abilities.

Student perceptions of peer review in the classroom (Study 9)

Lastly, student perceptions of peer review were assessed with an anonymous survey (n = 1,026 students). Students were overwhelmingly positive about the use of peer review in the classroom with 83% on average reporting that it positively impacted their laboratory reports, editing, writing, critical thinking and research skills (Table 5.1) and these positive perceptions were consistent for different three introductory biology courses surveyed over two terms (Figure 5.1). Notably, 86% students reported that that act of *giving* feedback was equally useful for improving

critical thinking skills as the act of receiving feedback. Written comments indicated that the act of reviewing others' work stimulated self-reflection, self-evaluation and an awareness of one's own learning process.

Students expressed some concern about the ability of peer to be effective reviewers though 80% reported that the feedback received was helpful and 69% reported it was satisfactory (see Table 5.1). Only 5% of students admitted to giving poor quality feedback however indicating a disjunction in students' perceptions. To address this concern, instructors are urged to share the results of research on peer review with students. Students see only the few papers they review and the few reviews they receive. Providing them with research results will allow them a course-wide perspective that peers, especially in aggregate, are reliable and provide useful feedback (Study 1 reported here as well as Cho, Schunn, & Wilson, 2006). Notably, when students are blinded to the source of the feedback, they often perceive peer feedback as comparable in quality to that provided by instructors (Cho, Schunn et al. 2006). Thus, the only concern consistently expressed by students engaging in peer review is repudiated by a course-wide perspective and corresponding research data. If such data are provided to students, it is anticipated that student concerns would dissipate.

Further, as students gain experience with peer review, they maintain or increase their positive perspective (Figure 5.1). The majority of respondents from the BIOL 102 sample reported that they had participated in peer review the previous semester (n = 303) and most (83%) reported that peer review was less difficult the second time and 64% said it was more useful. Thus, this finding further supports the notion that repeated exposure to peer review may show accelerating benefits. As they gain experience, students can focus more of their cognitive energy on the substance of the task rather than the procedural details. This increased focus should facilitate the improvement that is likely to be seen in their evaluative skills that will correspondingly increase the quality of the feedback they provide.

A few other studies exist which have captured students' perceptions of the peer review process and they generally agree with the findings reported above. Stefani (1994) reported that 100% of first year undergraduates said that peer review of biochemistry laboratory reports made them "think more" and 85% said it made them "learn more" (n = 120 students) but provided no further information as to how or why peer review caused these changes. Hanrahan and Issacs (2001, p. 57) surveyed

233 third year university students with a single open-ended query "give the pros and cons of peer review and self-assessment." Their students reported the following similar results: peer review was productive, improved their papers, and helped to develop critical thinking skills. In addition, Hanrahan and Issacs' (2001) students found the process time-consuming and desired higher quality feedback from their peers. In contrast to our results, their students felt empathy for the time instructors spent grading papers and found the exposure of their work to their peers to be motivating. It is unclear if their peer review process was anonymous which might have been the difference that caused this increased motivation. Hanrahan and Issacs (2001) do not provide any data on how prevalent each perception was in the student population, so it is unclear if the benefits and challenges they report were experienced by many or a few students.

Thus, this work enriched this field of knowledge in four ways. Firstly, it surveyed student perceptions of peer review from a larger sample size than the largest published study to date on (four times that of Hanrahan and Issacs). Secondly, this work contributed some much needed detail on the mechanisms by which students believed peer review benefited them. Thirdly, these data determined that the majority of the student population believed peer review was beneficial and negative experiences were in the minority. Fourthly, this work provides information on the effect of multiple peer review experiences which has not been previously discussed in the literature at all.

Student perceptions of the role of peer review in the scientific community (Study 10)

Students also made connections between the use of peer review in the classroom and its role in the scientific community. Students believed scientists experienced many of the same benefits from peer review that they themselves did. They were cognizant of the quality control role that peer review plays in maintaining the integrity of scientific work thereby indicating an awareness of the process that distinguishes scholarly publications from popular literature. Students also characterized reviewing as a valuable scientific skill they wished to acquire in their development as scientists. Students thus perceived peer review as an effective pedagogical strategy for improving scientific reasoning and writing skills in the classroom as well as a valuable scientific skill in and of itself.

Summary of conclusions

These finding suggest that peer review was effective for improving students' scientific reasoning skills and scientific writing. Repeated experience with peer review accelerated gains in scientific reasoning beyond that achieved by academic maturity alone. Students were effective and capable reviewers from the introductory level onwards dispelling concerns that peer review is too challenging for freshman. Use of peer feedback alone improved student laboratory reports indicating that student writing can be improved in the absence of time-intensive instructor feedback. These findings do not suggest that there is no need for instructor feedback, merely that student feedback is also productive and should be used to increase the overall amount of formative feedback provided to students.

Laboratory reports were further determined to be a rich source of data on student progress over time. The Universal Rubric for Laboratory Reports was demonstrated to be a reliable common metric. Application of the Rubric to multiple course assignments highlighted gaps and mis-alignments between assignment expectations, desired student performance and curriculum goals. When graduate student teaching assistants were provided training on the use of the Rubric in teaching and grading, the reliability of scores assigned to student work were comparable to those for published research in the science education field and above those produced by professional peer review. Graduate teaching assistants recommended that training on the Rubric be provided to all incoming biology graduate students.

Undergraduate students perceived peer review as a worthwhile activity. They believed peer review improved their writing and critical thinking skills and they perceived it as a valuable future skill they would need in their development as scientists. To assist the reader, the results of this study are summarized in Table 6.2 and recommendations for improving classroom enactment follow in Table 6.3.

Table 6.2. *Summary of Research Findings From This Study*

- Undergraduates (even freshman) were effective and consistent peer reviewers whose feedback produced meaningful improvements in final paper quality.
- Peer review increased scientific reasoning and writing skills to a greater degree than did academic maturity. Specifically, freshman laboratory reports which underwent peer review scored higher on 12 of 15 criteria than laboratory reports written by students with an average of 91 credit hours and higher GPAs which were not peer reviewed.
- Greater incorporation of rubric criteria into assignments improved student performance. Some criteria required explicit inclusion in the assignment instructions or students did not address them at all (e.g. use of controls in experimental design, use of statistics, use of primary literature) while one criterion (e.g. that hypotheses needed to have scientific merit) was addressed whether or not it was mentioned in the assignment.
- The reliability of the Universal Rubric for Laboratory Reports was notable ($g = 0.85$) and independent of biological subject matter in the three courses tested.
- Scores generated by trained or natural biology graduate teaching assistants using the Rubric were as reliable as those reported in science education research literature. It should be noted that even the natural raters in this study had at least a full semester of pedagogical training in introductory biology that included exposure to the Rubric.
- A few hours of explicit training on the use of the rubric did slightly improve the consistency of graduate teaching assistants over natural conditions.
- Graduate students who received a few hours of explicit training on the use of the Rubric recommended that such training be provided to all teaching assistants.
- Application of a Universal Rubric to assignments in multiple courses detected mis-alignments and gaps between curricular goals, course assignments, and Rubric criteria. These gaps affected student performance in those areas.
- Undergraduates were positive about peer review and reported that it benefited them in multiple ways (writing, reasoning, thinking, researching).
- Undergraduates perceived peer review as a valuable stand-alone skill and a natural part of their development as scientists.

University science departments are thus encouraged to incorporate peer review as an effective pedagogical strategy for improving student scientific reasoning and science writing. The incorporation of peer review is particularly recommended whenever instructor time is too limited for students to receive feedback on their writing. Peer review should also be incorporated however even in situations where instructors have sufficient time to provide extensive written formative feedback because the quadrupling of practice time that students spend engaged in evaluation and self-reflection is valuable and does not occur when

students only receive instructor feedback. The characteristics of the peer review process do alter its effectiveness however. When incorporating peer review, instructors should observe the following recommendations.

Study limitations and recommendations

This research occurred at an institution that had incorporated peer review for several years prior to the collection of data and that experience likely strengthened these findings. Namely, initial incorporation of peer review or the Rubric might not produce gains as large as those reported here. Instructor experience (both faculty and graduate students) in how to best implement and present the peer review process is expected to affect the impact of peer review. Specifically, degree of experience with peer review and the Rubric are anticipated to have the greatest impacts in three areas: 1) reliability of graduate student scores, 2) impact of peer feedback on writing and 3) student perceptions of and satisfaction with peer review as a pedagogy. To improve the reliability of the scores produced by the Rubric as rapidly as possible, it is recommended that instructors score exemplars and discuss how they will interpret and apply the Rubric criteria to student work. The process of building consensus on a few example papers is believed to significantly expedite the instructor's development as consistent scorers. To increase the impact of peer feedback, instructors are encouraged to design assignments so reviewers are accountable for the quality of the feedback they provide as well as provide them with instructional supports on what makes feedback useful (directive, constructive suggestions for change, not praise or criticism based). The best way to improve student perceptions of the value of peer review is to directly and frequently discuss the rationale for incorporating peer review into coursework as well as its role in the scientific community.

Additionally, instructors and program evaluators are cautioned to view the Universal Rubric as a tool rather than an answer. Post-hoc application of the Rubric is likely to be unproductive. There must be intentional and conscious alignment between curriculum goals, course design, assignment details and Rubric criteria in order for students to reasonably develop the desired skills over time and for laboratory report scores to consequently show meaningful improvement. Without such intentional coordination, the Rubric scores will mostly return information on

mis-alignments among these factors. Within a course, instructors are specifically encouraged to select Rubric criteria that are directly relevant to their instructional goals *prior* to the development of the assignment. Rubric criteria must be a natural fit for the assignment or the assignment must be designed to address those criteria. Instructional practices must also consistently valued and support those criteria (i.e. students need opportunities to practice the desired skills and instructors should role-model effective scientific reasoning).

Table 6.3. *Summary of Recommendations for Implementing Peer Review.*

- Be explicit in discussing with students the role of peer review in the scientific community as well as its benefits in the classroom.
- Share research results with students demonstrating that peers are effective reviewers and that peers can provide useful feedback that improves paper quality if incorporated.
- Design assignments to encourage students to provide high quality written feedback to each other. Means of doing this include explicitly defining and discussing what comprises useful feedback and using accountability measures such as randomly checking review quality (such checks are much less time consuming that reading draft papers).
- Design assignments so that assignment criteria and peer review criteria both align with instructional goals. Ideally, instructional goals span multiple courses and expectations for student performance are consistently aligned and developed throughout those educational experiences.
- Use a rubric as a means of defining assignment criteria to students. Use of a rubric deepens student understanding of the intent of criteria and helps them to provide better feedback to peers.
- Have relevant instructors build consensus on the interpretation of rubric criteria to facilitate scoring consistency within and across courses.
- Try to borrow from rubrics developed by others, especially if they have been reliability tested in relevant contexts and contain criteria derived from the scientific community. The Universal Rubric for Laboratory Reports is recommended when relevant to program or instructional goals.

Additionally, it should be noted that none of the measures used here provide a comprehensive examination of students' scientific reasoning ability. These measures are biased towards students who are effective at written communication and may miss examples of gains in reasoning skills for students who have difficulty translating their thinking onto paper. As effective written communication is an

explicit goal of the studied curriculum however, this emphasis on writing is appropriate, but does cause these estimates of student ability to be conservative.

In sum, incorporation of peer review can cause significant gains in student scientific reasoning and writing abilities especially if enacted in the manner described above. The primary criterion for producing an effective peer review process is to build the process using the same the motivations for peer review as exist in the scientific community: to produce useful formative feedback on the validity of one's scientific work in order to elevate the quality of science. This focus on improving the quality of students' scientific thought and writing through authentic practice will concurrently improve students' learning of science at the university level.

LITERATURE CITED

Abd-El-Khalick, F., & Lederman, N. G. (2000). Improving science teachers' conceptions of nature of science: A critical review of the literature. *International Journal of Science Education, 22*(7), 665-701.

Admiraal, W., Wubbels, T., & Pilot, A. (1999). College teaching in legal education: Teaching method, students' time on task, and achievement. *Research in Higher Education, 40*(6), 687-704.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*(4), 523-550.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching, 39*(10), 952-978.

Anderson, G. (2002). *Fundamentals of educational research* (2nd ed.). Philadelphia: Routledge Falmer.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks CA: Corwin Press.

Baird, J. R., & White, R. T. (1996). Metacognitive strategies in the classroom. In D. F. Treagust, R. Duit & B. J. Fraser (Eds.), *Improving teaching and learning in science and mathematics* (pp. 190-200). New York: Teachers College Press.

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizabilty of domain-independent performance assessments. *Journal of Educational Research, 89*(4), 197-205.

Basey, J. M., Mendelow, T. N., & Ramos, C. N. (2000). Current trends of community college lab curricula in biology: An analysis of inquiry, technology, and content. *Journal of Biological Education, 34*(2), 80-86.

Baxter, G., Shavelson, R., Goldman, S., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1-17.

Bendixen, L. D., & Hartley, K. (2003). Successful learning with hypermedia: The role of epistemological beliefs and metacognitive awareness. *Journal of Educational Computing Research, 28*(1), 15-30.

Bianchini, J. A., Whitney, D. J., Breton, T. D., & Hilton-Brown, B. A. (2001). Toward inclusive science education: University scientists' views of students, instructional practices, and the nature of science. *Science Education, 86*, 42-78.

Birk, J. P., & Kurtz, M. J. (1999). Effect of experience of retention and elimination of misconceptions about molecular structure and bonding. *Journal of Chemical Education, 76*(1), 124-128.

Bloxham, S., & West, A. (2004). Understanding the rules of the game: Marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education, 29*(6), 721 - 733.

Boyer Commission. (1998). *Reinventing undergraduate education: A blueprint for America's research universities*. Washington DC: Carnegie Foundation for the Advancement of Teaching.

Boyer Commission. (2001). *Reinventing undergraduate education. Three years after the Boyer report: Results from a survey of research universities.* Washington DC: National (Boyer) Commission on Educating Undergraduates in the Research University.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington DC: National Academy Press.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge UK: Cambridge University Press.

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issue and Practice, 11*, 27-34.

Cabrera, A. F., Colbeck, C. L., & Terenzini, P. T. (2001). Developing performance indicators for assessing classroom teaching practices and student learning. *Research in Higher Education, 42*(3), 327-352,.

Campbell, B., Kaunda, L., Allie, S., Buffler, A., & Lubben, F. (2000). The communication of laboratory investigations by university entrants. *Journal of Research in Science Teaching, 37*(8), 839-853.

Carnegie Initiative on the Doctorate. (2001). *Overview of doctoral education studies and reports: 1990 - present*. Stanford, CA: The Carnegie Foundation for the Advancement of Teaching.

Chinn, P. W. U., & Hilgers, T. L. (2000). From corrector to collaborator: The range of instructor roles in writing-based natural and applied science classes. *Journal of Research in Science Teaching, 37*(1), 3-25.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260-294.

Cho, K., Schunn, C. D., & Wilson, a. R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*, 119-135.

Committee on Undergraduate Biology Education. (2003). *Bio 2010: Transforming undergraduate education for future research biologists*. Washington DC: National Academy Press.

Connally, P., & Vilardi, T. (1989). *Writing to learn mathematics and science*. New York: Teachers College Press.

Crick, J. E., & Brennan, R. L. (1984). *General purpose analysis of variance system (GENOVA)* (Version 2.2) [Fortran]. Iowa City: American College Testing Program.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Cudek, R. (1980). A comparative study of indices for internal consistency. *Journal of Educational Measurement, 17*(2), 117-130.

Davis, G., & Fiske, P. (2001). *The 2000 national doctoral program survey.* University of Missouri-Columbia: National Association of Graduate-Professional Students.

Dhillon, A. S. (1998). Individual differences within problem-solving strategies used in physics. *Science Education, 82*, 379-405.

Driver, R., & Scott, P. H. (1996). Curriculum development as research: A constructivist approach to science curriculum development and teaching. In D. F. Treagust, R. Duit & B. J. Fraser (Eds.), *Improving teaching and learning in science and mathematics* (pp. 94-108). New York: Teachers College Press.

Duit, R., & Confrey, J. (1996). Reorganizing the curriculum and teaching to improve learning in science and mathematics. In D. F. Treagust, R. Duit & B. J. Fraser (Eds.), *Improving Teaching and Learning in Science and Mathematics* (pp. 79-93). New York NY: Teachers College Press.

Dunbar, K. (1997). How scientists think: Online creativity and conceptual change in science. In T. Ward, S. Smith & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change* (pp. 461-492). Washington DC: American Psychological Association Press.

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology, 21*(1), 49-58.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725-747.

Feldon, D. (2007). The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review. 19*(2), 91-110

Florence, M. K., & Yore, L. D. (2004). Learning to write like a scientist: Coauthoring as an enculturation task. *Journal of Research in Science Teaching, 41*(6), 637 - 668.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92*(3), 404-423.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research: An International Journal, 1*, 7-33.

Fraser, B. J. (2002). Learning environments research: Yesterday, today and tomorrow. In S. C. Goh & M. S. Khine (Eds.), *Studies in educational learning environments: An international perspective.* Singapore: World Scientific Publishing.

Gaff, J. G. (2002). The disconnect between graduate education and the realities of faculty work: A review of recent research. *Liberal Education, 88*(3), 6-13.

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching, 33*(7), 773-798.

Goodman, B. E., Eisenhart, M., DeHaan, R. L., Kemm, R. E., Rodenbaugh, D. W., & Pelaez, N. (2007). Scientific principles of education research: Experimental Biology 2007. *Advances in Physiology Education, 31*, 374-376.

Golde, C. M. (2001). *Findings of the survey of doctoral education and career preparation: A report to the Preparing Future Faculty program.* Madison: University of Wisconsin.

Grant, P. R., & Grant, B. R. (2002). Unpredictable evolution in a 30-year study of Darwin's Finches. *Science, 296*(5568), 707-711.

Haaga, D. A. F. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*(1), 28-32.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509-1528.

Halonen, J. S., Bosack, T., Clay, S., McCarthy, M., Dunn, D. S., Hill, G. W., IV, et al. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology, 30*(3), 196-208.

Hand, B., Hohenshell, L., & Prain, V. (2004). Exploring students' responses to conceptual questions when engaged with planned writing experiences: A study with year 10 science students. *Journal of Research in Science Teaching, 41*(2), 186-210.

Hand, B., Prain, V., & Wallace, C. (2002). Influences of writing tasks on students' answers to recall and higher-level test questions. *Research in Science Education, 32*(1), 19-34.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70.

Haswell, R. H. (2000). Documenting Improvement in College Writing: A Longitudinal Approach. *Written Communication, 17*(3), 307-352.

Hoadley, C. M., & Linn, M. C. (2000). Teaching science through online, peer discussions: Speakeasy in the knowledge integration environment. *International Journal of Science Education, 22*(8), 839 - 857.

Hogan, K., Nastasi, B. K., & Pressley, M. (2000). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition & Instruction, 17*(4), 379-432.

Jensen, M. S., & Finley, F. N. (1996). Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *Journal of Research in Science Teaching, 33*(8), 879 - 900.

Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching, 35*(1), 89-103.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-29.

Karplus, R. (1977). Science teaching and the development of reasoning. *Journal of Research in Science Teaching, 14*(2), 169-175.

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education, 86*(3), 314 - 342.

Keys, C. W. (1994). The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth-grade students. *Journal of Research in Science Teaching, 31*(9), 1003-1022.

Keys, C. W. (1995). An interpretive study of students' use of scientific reasoning during a collaborative report writing intervention in ninth grade general science. *Science Education, 79*(4), 415-435.

Keys, C. W. (1999a). Language as an indicator of meaning generation: An analysis of middle school students' written discourse about scientific investigations. *Journal of Research in Science Teaching, 36*(9), 1044-1061.

Keys, C. W. (1999b). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education, 83*, 115-130.

Keys, C. W. (2000). Investigating the thinking processes of eighth grade writers during the composition of a scientific laboratory report. *Journal of Research in Science Teaching, 37*(7), 676-690.

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching, 36*(10), 1065-1084.

Klahr, D., & Dunbar, K. (1988). Dual search space during scientific reasoning. *Cognitive Science, 12*, 1-48.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3), 203-270.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., et al. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121-138.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education, 91*(5), 710-726.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kuhs, T., Johnson, R. L., Agruso, S. A., & Monrad, D. M. (2001). Guides to scoring student work: Checklists and rubrics. In *Put to the test: Tools and techniques for classroom assessment* (pp. 54-75). Portsmouth, NH: Heinemann.

Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational Researcher, 32*(8), 21-25.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*, 11-24.

Lawson, A. E. (1979). Relationships among performances on group-administered items of formal reasoning. *Perceptual and Motor Skills, 48*, 71-78.

Lawson, A. E. (1980). The relationship among levels of intellectual development, cognitive style and grades in a college biology course. *Science Education, 64*, 95-102.

Lawson, A. E. (1983). Predicting science achievement: The role of developmental level, disembedding ability, mental capacity, prior knowledge and beliefs. *Journal of Research in Science Teaching, 20*(2), 117-129.

Lawson, A. E. (1992). What do tests of formal reasoning actually measure? *Journal of Research in Science Teaching, 29*(9), 965-984.

Lawson, A. E., Alkhoury, S., Benford, R., Clark, B. R., & Falconer, K. A. (2000). What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching, 37*(9), 996-1018.

Lawson, A. E., Baker, W. P., DiDonato, L., Verdi, M. P., & Johnson, M. A. (1993). The role of physical analogies of molecular interactions and hypothetico-deductive reasoning in conceptual change. *Journal of Research in Science Teaching, 30*(9), 1073-1086.

Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-efficacy, reasoning ability and achievement in college biology. *Journal of Research in Science Teaching, 44*(5), 706-724.

Lerner, N. (2007). Laboratory lessons for writing and science. *Written Communication, 24*(3), 191-222.

Linn, M. C. (1997). Finding patterns in international assessments. *Science, 277*(5333), 1743.

Luft, J. A., Kurdziel, J. P., Roehrig, G. H., Turner, J., & Wertsch, J. (2004). Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching, 41*(3), 211-233.

Marcoulides, G. A., & Simkin, M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business, 70*(4), 220-224.

Marsh, H. W., & Ball, S. (1981). Interjudgmental reliability of reviews for the Journal of Educational Psychology. *Journal of Educational Psychology, 73*(6), 872-880.

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education, 57*(2), 151-169.

Marsh, H. W., & Bazeley, P. (1999). Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. *Multivariate behavioral research, 34*(1), 1-30.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17*(2), 13-17.

McAlpine, L., & Weston, C. (2000). Reflection: Issues related to improving professors' teaching and students' learning. *Instructional Science, 28*(5), 363 - 385.

McNeill, B., Bellamy, L., & Burrows, V. (1999). A quality based assessment process for student work products. *Journal of Engineering Education, 88*(4), 485-500.

Mervis, J. (2000). Graduate Educators Struggle to Grade Themselves. *Science, 287*(5453), 568 - 570.

Mistler-Jackson, M., & Songer, N. B. (2000). Student motivation and internet technology: Are students empowered to learn science? *Journal of Research in Science Teaching, 37*(5), 459-479.

National Research Council Committee on the Foundations of Assessment. (2001). *Knowing what students know: The science and design of educational assessment.* Washington DC: National Academy Press.

National Science Foundation. (2008*). NSF 08-1 Grant Proposal Guide. Chapter III: NSF proposal processing and review*. http://www.nsf.gov/publications January 2008. Retrieved Nov 2, 2007

Nichols, P., Pokorny, R., Jones, G., Gott, S. P., & Alley, W. E. (1992). *Evaluation of an avionics troubleshooting tutoring system. Technical Report*. Brooks AFB, TX: Air Force Human Effectiveness Directorate.

Norman, O. (1997). Investigating the nature of formal reasoning in chemistry: Testing Lawson's multiple hypothesis theory. *Journal of Research in Science Teaching, 34*(10), 1067-1081.

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

O'Neill, D. K., & Polman, J. L. (2004). Why educate little scientists? Examining the potential of practice-based scientific literacy. *Journal of Research in Science Teaching, 41*(3), 234-266.

Ong, J. (2007). Automated performance assessment and feedback for free-play simulation-based training. *Performance Improvement, 46*(10), 24-31.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Pelaez, N. J. (2002). Problem-based writing with peer review improves academic performance in physiology. *Advances in Physiology Education, 26*(3), 174-184.

Pelaez, N. J., & Gonzalez, B. L. (2002). Sharing science: Characteristics of effective scientist-teacher interactions. *Advances in Physiology Education, 26*(3), 158-167.

Penny, J., Johnson, R. L., & Gordon, B. (2000a). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143-164.

Penny, J., Johnson, R. L., & Gordon, B. (2000b). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education, 68*(3), 269-287.

Petty, R. E., Fleming, M. A., & Fabrigar, L. R. (1999). The review process at PSPB: Correlates of inter-reviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin, 25*(2), 188-203.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*(1), 51 - 63.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Powell, K. (2003). Science education: spare me the lecture. *Nature, 425*, 234-236.

Ravitz, J. (2002). Cilt2000: Using technology to support ongoing formative assessment in the classroom. *Journal of Science Education and Technology, 11*(3), 293-296.

Reiser, B. J., Smith, B. K., Tabak, I., Steinmuller, F., Sandoval, W. A., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress.* Mahway, NJ: Erlbaum.

Reiser, B. J., Tabak, I., Sandoval, W. A., & Steinmuller, F. (2003). *The Galapagos Finches* (Version 26j14.09): Northwestern University.

Rijlaarsdam, G., Couzijn, M., Janssen, T., Braaksma, M., & Kieft, M. (2006). Writing experiment manuals in science education: The impact of writing, genre, and audience. *International Journal of Science Education, 28*(2-3), 203-233.

Rivard, L. P. (2004). Are language-based activities in science effective for all students, including low achievers? *Science Education, 88*, 420 - 442.

Rivard, L. P., & Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education, 84*, 566-593.

Roche, Lawrence A., & Marsh, Herbert W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science, 28*(5), 439 - 468.

Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., Daniels, S. R., et al. (2006). Effect of blinded peer review on abstract acceptance. *Journal of the American Medical Association, 295*(14), 1675-1680.

Russel, A. (2007, June). *Showcasing successful practices and advancing [*Calibrated Peer Review*] CPR research*. Paper presented at the National Science Foundation Calibrated Peer Review Symposium. Texas A&M, College Station.

Schraagen, J. M. C. (1990). How experts solve a novel problem within their domain of expertise (No. IZF 1990 B-14). Soesterberg, The Netherlands: Netherlands Organization for Applied Scientific Research.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460-475.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T. Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436-1460.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cogitive Science, 23*(3), 337-370.

Schwartz, R. S., Lederman, N. G., & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education, 88,* 610-645.

Schwarz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs may make a right ... If they argue together! *Cognition and Instruction, 18*(4), 461-494.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory* (1st ed.). London: Sage Publications, Inc.

Simmons, P. E., Emory, A., Carter, T., Coker, T., Finnegan, B., Crockett, D., et al. (1999). Beginning teachers: Beliefs and classroom actions. *Journal of Research in Science Teaching, 36*(8), 930-954.

Simonton, D. K. (2004). *Creativity in science: Chance, logic, genius and zeitgeist*: Cambridge University Press.

Sluijsmans, D. M. A., Dochy, F. J. R. C., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer and co-assessment. *Learning Environments Research, 1*, 293-319.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*(4), 753–766.

South Carolina Department of Education. (2005) English Language Development Assessment (ELDA) Extended Constructed Response Writing Rubric http://ed.sc.gov/agency/offices/assessment/programs/elda/elda.html. Retrieved April, 5, 2005.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-76.

Sternberg, R. J., & Gordeeva, T. (1996). The anatomy of impact: What makes an article influential? *Psychological Science, 7*(2), 69-75.

Sternglass, M. S. (1993). Writing development as seen through longitudinal research: A case study exemplar. *Written Communication, 10*(2), 235-261.

Timmerman, B., Strickland, D., & Carstensen, S. (2008). Curricular reform and inquiry teaching in biology: where are our efforts most fruitfully invested? *Integrative and Comparative Biology, 48*(2), 1-15.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169.

Trevisan, M. S., Davis, D. C., Calkins, D. E., & Gentili, K. L. (1999). Designing sound scoring criteria for assessing student performance. *Journal of Engineering Education, 88*(1), 79-85.

Trowbridge, J. E., & Wandersee, J. H. (1994). Identifying critical junctures in learning in a college course on evolution. *Journal of Research in Science Teaching, 31*(5), 459-474.

van Berkel, H. J. M., & Schmidt, H. G. (2000). Motivation to commit oneself as a determinant of achievement in problem-based learning. *Higher Education, 40*(2), 231-242.

Volkmann, M. J., & Zgagacz, M. (2004). Learning to teach physics through inquiry: The lived experience of a graduate teaching assistant. *Journal of Research in Science Teaching, 41*(6), 584-602.

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. In D. Gabel (Ed.), *Handbook of research on science teaching* (pp. 177-210). New York: Macmillan.

Ward, S.L. & Overton, W.F. (1990). Semantic familiarity, relevance, and the development of deductive reasoning. *Developmental Psychology, 26*(3), 488-493

Westbrook, S. L., & Rogers, L. N. (1994). Examining the development of scientific reasoning in ninth-grade physical science students. *Journal of Research in Science Teaching, 31*(1), 65-76.

Wilson, T. D., DePaulo, B. M., Mook, D. G., & Klaaren, K. J. (1993). Scientists' evaluation of research: The biasing effects of the importance of the topic. *Psychological Science, 4*(5), 322–325.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*(1), 47-58.

Yore, L. D., Florence, M. K., Pearson, T. W., & Weaver, A. J. (2006). Written discourse in scientific communities: A conversation with two scientists about their views of science, use of language, role of writing in doing science, and compatibility between their epistemic views and language. *International Journal of Science Education, 28*(2&3), 109-141.

Yore, L. D., Hand, B. M., & Florence, M. K. (2004). Scientists' views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching, 41*(4), 338-369.

Yore, L. D., Hand, B. M, & Prain, V. (2002). Scientists as writers. *Science Education, 86*, 672-692.

Yorke, M. (2003). Formative assessment in higher education: Moves toward theory and enhancement of pedagogic practice. *Higher Education, 45*, 477-501.

Zembal-Saul, C., Blumenfeld, P., & Krajcik, J. (2000). Influence of guided cycles of planning, teaching, and reflection on prospective elementary teachers' science content representations. *Journal of Research in Science Teaching, 37*(4), 318-339.

Ziman, J. (1998). What is science? In E. D. Klemke, R. Hollinger & D. W. Rudge (Eds.), *Introductory readings in the philosophy of science* (pp. 48-53). Amherst NY: Prometheus Books.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99-149.

Zohar, A. (1996). Transfer and retention of reasoning strategies taught in biological contexts. *Research in Science and Technological Education, 14*(2), 205-219.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*(1), 35-62.

Appendix 1

Goals of the USC Biology Undergraduate Curriculum

By graduation students will:

1. **Demonstrate a Base of Knowledge**
    a. Possess a conceptual framework that identifies the relationships between the major domains in the field of biology.

2. **Demonstrate understanding and use of scientific reasoning and process (students should be able to "think like a scientist")**
    a. Identify assumptions
    b. Create and evaluate hypotheses
    c. Create abstract models of data
    d. Design experiments relevant to the questions and models
    e. Analyze qualitative and quantitative data
    f. Assess validity of work, identify gaps in knowledge
    g. Evaluate the results of the analyses and experiments and decide on next step
    h. Learn from "mistakes" (identify unintended results as opportunities for discovery)
    i. Learn new concepts and integrate them with current knowledge

3. **Demonstrate information literacy and technological fluency**
    a. Locate and evaluate information needed to make decisions, solve problems, design experiments, understand scientific data
    b. Work effectively with common technologies in biology
    c. Read primary literature and evaluate validity (on an appropriate level)
    d. Evaluate and use biological databases (literature and public datasets)

4. **Effectively communicate within a scientific context**
    a. Be able to simplify and explain scientific concepts and results of experiments to a non-biologist (requires sufficient understanding to avoid jargon, half-answers, etc.)
    b. Display and explain scientific results clearly and persuasively to peers both verbally and in writing (includes the ability to graph data appropriately and accurately).

5. **Demonstrate independent and collaborative learning skills**
    a. Be able to learn independently and then share that knowledge with others
    b. Work collaboratively to leverage greater learning than if working alone

6. **Articulate the Nature of Science and its Interface with other Disciplines**
    a. Appreciate the role of creativity in science
    b. Understand the recursive nature of science (how new results continually modify and push forward previous knowledge)
    c. Be able to explain the role of peer review in science as a quality control mechanism
    d. Be able to distinguish the normal level of discussion that occurs at the cutting edge of research from true dissent or controversy over the validity of scientific results.
    e. Have a broad appreciation of the interesting questions, state of knowledge in the field of biology (molecular to ecosystems levels of complexity)
    f. Understand the social and natural context of knowledge (role of science in society, influence of society on science)
    g. Be able to debate the ethical implications of science (difference between our abilities and our values).
    h. Develop an appreciation for the history of ideas and the development of the major fields of biology

How to provide useful feedback: BIOL 102 Fall 2004

Useful feedback is:

- **specific and concrete,**
- **focuses on the quality** of the author's argument (e.g. are conclusions logical and well supported by the evidence/data?) **rather than on the mechanics** of writing, (e.g. spelling or grammar),
- **identifies assumptions** behind **or consequences** of author's ideas which the author has not explicitly discussed
- tells the writer **why** you think they did or did not meet the criteria.
- **In sum, is likely to result in meaningful revisions or new content being added to the paper.**

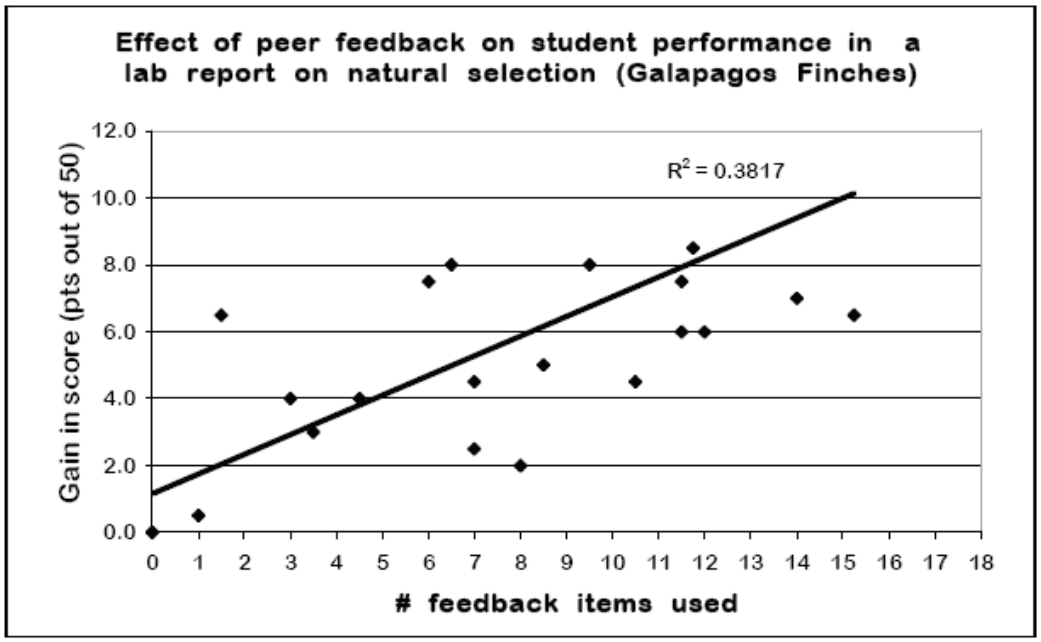
You will be prompted by the CPR website to provide feedback for each criterion. Please make an effort to provide useful feedback to your classmates. Your TA will also randomly select one of your three reviews and grade the quality of the feedback you provide. Useful items earn 1 pt, partially useful items earn 0.5 pts and non-useful feedback earn 0 points and each review is worth up to 10 points. You may write as many feedback items as you would like, but you must provide at least one piece of feedback in response to each prompt in CPR. Reviewing other students' papers should also give you insight into strengths and weaknesses in your own paper. The benefits you will receive from this exercise are directly correlated with the effort you put into it. Below are some examples to help you understand what constitutes useful feedback.

Useful feedback can improve your grade (see graph on the back).

Examples:

Feedback item	Useful?	How to improve the feedback
1. Your paper is GREAT! How did you come up with your idea?	NO	Provides no actual information to the writer on HOW to improve the paper.
2. At the end of paragraph 2, you say you think this was a sex-linked cross. Is this your hypothesis? What traits do you think the parents had? Why do you think this is the best explanation?	Yes	Full of detail about where and why the reviewer was lost and if the writer answers the reviewer's questions, the paper will have a clearer statement of the hypothesis, a consideration of alternative explanations and logical connection between hypotheses, data and conclusions.
3. Your argument makes no sense. What is your evidence?	Partially	Asking for evidence is useful, but reviewer does not indicate which part of the paper is confusing them or what exactly they don't understand.
4. Your argument depends on weight being an inherited trait. What evidence do you have to support this assumption?	Yes	The reviewer has identified an assumption made by the writer and pointed out how the validity or invalidity of this assumption could impact the writer's conclusion.
5. Which of your hypotheses is best supported by the data?	Partially	The reviewer is specific in indicating that the writer did something well (posed multiple explanations) but provides only a vague indication that the writer needs to discuss the data more without indicating how or where they felt the writer's conclusions were lacking.

Please put effort into your peer reviews – the feedback can make a real difference your peers! As a writer you should note that for every three pieces of useful feedback used, writers saw approximately a two point gain in their final grade.



Appendix 3. Universal Rubric for Laboratory Reports

Universal Rubric for Lab Reports

developed by

Briana Timmerman
University of South Carolina
Department Biological Sciences

For more information, please contact:
Briana Timmerman
timmerman@schc.sc.edu

The rubric was developed as a means of measuring how well students are achieving the stated curriculum goals of the USC Biology Curriculum (http://www.biol.sc.edu/undergrad/curriculum.html). The rubric was created and refined in consultation with a wide variety of faculty, instructors and educational specialists, both within and outside the department (see acknowledgements page). The criteria were selected as the minimum framework one would expect to see in any good biology lab report or other scientific communication. The levels of student performance are intended as a roadmap of the probable learning trajectory of a typical undergraduate student. The "Proficient" level describes the performance we would hope an exceptional undergraduate or beginning graduate student would achieve. Instructors are encouraged to select and use only the criteria and levels of student performance that they feel are relevant to their student population and assignment goals. A scoring guide (rubric plus examples of student work at each level of performance) has also been developed and is available upon request. Feedback or comments would be most appreciated if sent to Briana Timmerman at the contact information listed above. The Rubric underwent formal reliability testing (Timmerman *et al.* 2007) producing a three rater reliability score of $g = 0.85$ using biology graduate students as raters and generalizability analysis. Further details are also available from Briana Timmerman.

Please cite as:
Timmerman, B.E., Johnson, R.L. and Payne, J. 2007. Development of a universal rubric for assessing students' science inquiry skills. *National Association of Research in Science Teaching 2007 Annual Meeting* New Orleans LA, April 15-18th

Support for this project was provided by NSF Award 0410992 to Timmerman.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Introduction: Context				
Demonstrates a clear understanding of the big picture; Why is this question important/ interesting in the field of biology?	• The importance of the question is not addressed. • How the question relates within the broader context of biology is not addressed.	• The writer provides a generic or vague rationale for the importance of the question. • The writer provides vague or generic references to the broader context of biology.	• The writer provides one explanation of why others would find the topic interesting. • The writer provides some relevant context for the research question(s).	• The writer provides a clear sense of why this knowledge may be of interest to a broad audience • The writer describes the current gaps in our understanding of this field and explains how this research will help fill those gaps
Introduction: Accuracy and relevance				
Content knowledge is accurate, relevant and provides appropriate background for reader including defining critical terms.	• Background information is missing or contains major inaccuracies. • Background information is accurate, but irrelevant or too disjointed to make relevance clear • Primary literature references are absent or irrelevant. May contain website or secondary references **websites or review papers are not primary**	• Background omits information or contains inaccuracies which detract from the major point of the paper. • Background information is overly narrow or overly general (only partially relevant). • Primary literature references, if present, are inadequately explained.	• Background information may contain minor omissions or inaccuracies that do not detract from the major point of the paper. • Background information has the appropriate level of specificity to provide relevant context. • Primary literature references are relevant and adequately explained but few.	• Background information is completely accurate • Background information has the appropriate level of specificity to provide concise and useful context to aid the reader's understanding. • Primary literature references are relevant, adequately explained, and indicate a reasonable literature search.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Hypotheses: Testable and consider alternatives				
Hypotheses are clearly stated, testable and consider plausible alternative explanations	• No hypothesis is indicated. • The hypothesis is stated but too vague or confused for its value to be determined • A clearly stated, but not testable hypothesis is provided. • A clearly stated and testable, but trivial hypothesis is provided.	• A single relevant, testable hypothesis is clearly stated • The hypothesis may be compared with a "null" alternative which is usually just the absence of the expected result.	• Multiple relevant, testable hypotheses are clearly stated. • Hypotheses address more than one major potential mechanism, explanation or factors for the topic.	• A comprehensive suite of testable hypotheses are clearly stated which, when tested, will distinguish among multiple major factors or potential explanations for the phenomena at hand.
Hypotheses: Scientific merit				
Hypotheses have scientific merit.	• Hypotheses are trivial, obvious, incorrect or completely off topic.	• Hypotheses are plausible and appropriate though likely or clearly taken directly from course material.	• Hypotheses indicate a level of understanding beyond the material directly provided to the student in the lab manual or coursework.	• Hypotheses are novel, insightful, or actually have the potential to contribute useful new knowledge to the field.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Methods: Controls and replication				
Appropriate controls (including appropriate replication) are present and explained. If the student designed the experiment:	• Controls and/ors replication are nonexistent, • Controls and/or replication may have been present, but just not described or • Controls and/or replication were described but were inappropriate.	• Controls consider one major relevant factor • Replication is modest (weak statistical power).	• Controls take most relevant factors into account • Controls include positive and negative controls if appropriate • Replication is appropriate (average sample size with reasonable statistical power).	• Controls consider all relevant factors • Controls have become methods of differentiating between multiple hypotheses. • Replication is robust (sample size is larger than average for the type of study).
If the instructor designed the experiment:	• Student fails to mention controls and/or replication or mentions them, but the description or explanation is incomprehensible.	• Student explanations of controls and/or replication are vague, inaccurate or indicate only a rudimentary sense of the need for controls and or replication	• Student evidences a reasonable sense of why controls/ replication matter to this experiment • Explanations are mostly accurate.	• Explanations of why these controls matter to this experiment are thorough, clear and tied into sections on assumptions and limitations
Methods: Experimental design				
Experimental design is likely to produce salient and fruitful results (tests the hypotheses posed.) Methods are:	• inappropriate • poorly explained / indecipherable	• appropriate • clearly explained • drawn directly from coursework • not modified where appropriate	• appropriate • clearly explained • modified from coursework in appropriate places • or drawn directly from a novel source (outside the course)	• appropriate • clearly explained • a synthesis of multiple previous approaches or an entirely new approach

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Results: Data selection				
Data are comprehensive, accurate and relevant.	• Data are too incomplete or haphazard to provide a reasonable basis for testing the hypothesis	• At least one relevant dataset per hypothesis is provided but some necessary data are missing or inaccurate • Reader can satisfactorily evaluate some but not all of writer's conclusions.	• Data are relevant, accurate and complete with any gaps being minor. • Reader can fully evaluate whether the hypotheses were supported or rejected with the data provided.	• Data are relevant, accurate and comprehensive. • Reader can fully evaluate validity of writer's conclusions and assumptions. • Data may be synthesized or manipulated in a novel way to provide additional insight.
Results: Data presentation				
Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete. <u>Presentation of data:</u>	• Labels or units are missing which prevent the reader from being able to derive any useful information from the graph. • Presentation of data is in an inappropriate format or graph type • Captions are confusing or indecipherable.	• contains some errors in or omissions of labels, scales, units etc., but the reader is able to derive some relevant meaning from each figure. • is technically correct but inappropriate format prevents the reader from deriving meaning or using it. Captions are missing or inadequate	• contains only minor mistakes that do not interfere with the reader's understanding and the figure's meaning is clear without the reader referring to the text. • Graph types or table formats are appropriate for data type. • includes captions that are at least somewhat useful.	• contains <u>no</u> mistakes • uses a format or graph type which highlights relationships between the data points or other relevant aspects of the data. • may be elegant, novel, or otherwise allow unusual insight into data • has informative, concise and complete captions.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Results: Statistical analysis				
Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained.	• No statistical analysis is performed. • Statistics are provided but are inappropriate, inaccurate or incorrectly performed or interpreted so as to provide no value to the reader.	• Appropriate, accurate descriptive statistics only are provided. • Inferential statistics are provided but either incorrectly performed or interpreted or an inappropriate test was used. • Appropriate, correct inferential statistics are provided, but lack sufficient explanation.	• Appropriate inferential (comparative) statistical analysis is properly performed and reasonably well explained. • Explanation of significant value may be limited or rote (e.g. use of $p<0.05$ only)	• Statistical analysis is appropriate, correct and clearly explained • includes a description of what constitutes a significant value and why that value was chosen as the threshold (may choose values beyond $p<0.05$).
Discussion: Conclusions based on data selected				
Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.	• Conclusions have little or no basis in data provided. • Connections between hypothesis, data and conclusion are non-existent, limited, vague or otherwise insufficient to allow reasonable evaluation of their merit. • Conflicting data are not addressed.	• Conclusions have some direct basis in the data, but may contain some gaps in logic or data or are overly broad. • Connections between hypothesis, data and conclusions are present but weak. • Conflicting or missing data are poorly addressed.	• Conclusions are clearly and logically drawn from and bounded by the data provided with no gaps in logic. • A reasonable and clear chain of logic from hypothesis to data to conclusions is made. • Conclusions attempt to discuss or explain conflicting or missing data.	• Conclusions are completely justified by data. • Connections between hypothesis, data, and conclusions are comprehensive and persuasive. • Conclusions address and logically refute or explain conflicting data • Synthesis of data in conclusion may generate new insights.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Discussion: Alternative explanations				
Alternative explanations are considered and clearly eliminated by data in a persuasive discussion. Alternative explanations:	• are not provided • are trivial or irrelevant • are mentioned but not discussed or eliminated.	• are provided in the discussion only • may include some trivial or irrelevant alternatives. • Discussion addresses some but not all of the alternatives in a reasonable way.	• Some alternative explanations are tested as hypotheses; those not tested are reasonably evaluated in the discussion. • Discussion of alternatives is reasonably complete, uses data where possible and results in at least some alternatives being persuasively dismissed.	• have become a suite of interrelated hypotheses that are explicitly tested with data. • Discussion and analysis of alternatives is based on data, complete and persuasive with a single clearly supported explanation remaining by the end of the discussion.
Discussion: Limitations of design				
Limitations of the data and/or experimental design and corresponding implications discussed. Limitations:	• are not discussed.	• are discussed in a trivial way (e.g. "human error" is the major limitation invoked).	• are relevant, but not addressed in a comprehensive way • Conclusions fail to address or overstep the bounds indicated by the limitations.	• are presented as factors modifying the author's conclusions. • Conclusions take these limitations into account.
Discussion: Significance of research				
Paper gives a clear indication of the significance of the research and its future directions. Future directions and significance of this research:	• are not addressed.	• are vague, implausible (not possible with current technologies or methodologies), trivial or off topic.	• are useful, but indicate incomplete knowledge of the field (suggest research that has already been done or is improbable with current methodologies) • suggest a fruitful line of research, but lack detail to indicate motivations for or implications of the future research.	• are salient, plausible and insightful • suggest work that would fill knowledge gaps and move the field forward.

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Use of Primary Literature				
Relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided). <u>Primary literature is defined as:</u> - peer reviewed - reports original data - authors are the people who collected the data. - published by a non-commercial publisher.	• Primary literature references are not included.	• Primary literature references are limited (only one or two primary references in the whole paper) • References to the textbook, lab manual, or websites may occur. • Citations are at least partially correctly formatted. Note that proper format includes a one-to-one correspondence between in-text and end of text references (no references at end that are not in text and vice versa) as well as any citation style currently in use by a relevant biology journal.	• Primary literature references are more extensive (at least one citation for each major concept) • Literature cited is predominantly (≥ 90%) primary literatures. • Primary literature references are used primarily to provide background information and context for conclusions • Primary literature references	• Primary literature references indicate an extensive literature search was performed. • Primary literature references frame the question in the introduction by indicating the gaps in current knowledge of the field. • Primary literature references are used in the discussion to make the connections between the writer's work and other research in the field clear • Primary literature references are properly and accurately cited

Appendix 3. Universal Rubric for Laboratory Reports

Criteria	Not addressed	Novice	Intermediate	Proficient
Writing quality				
Grammar, word usage and organization facilitate the reader's understanding of the paper. (some elements inspired by the SC State Dept. Education Extended-Response Scoring Guide for English Language Arts.)	• Grammar and spelling errors detract from the meaning of the paper. • Word usage is frequently confused or incorrect. • Subheadings are not used or poorly used. • Information is presented in a haphazard way.	• Grammar and spelling mistakes do not hinder the meaning of the paper. • General word usage is appropriate, although use of technical language is may have occasional mistakes. • Subheadings are used and aid the reader somewhat. • There is some evidence of an organizational strategy though it may have gaps or repetitions.	• Grammar and spelling have few mistakes. • Word usage is accurate and aids the reader's understanding. • Distinct sections of the paper are delineated by informative subheadings. • A clear organizational strategy is present with a logical progression of ideas.	• Correct grammar and spelling. • Word usage facilitates reader's understanding. • Informative subheadings significantly aid reader's understanding. • A clear organizational strategy is present with a logical progression of ideas. There is evidence of an active planning for presenting information; this paper is easier to read than most.

Acknowledgements

Many thanks to the following people for their time and efforts in developing and/or providing feedback on the rubric in its various stages of development:

USC Department of Biological Sciences Faculty:

Renae Brodie
Susan Carstensen
Erin Connelly
Brian Helmuth
Laurel Hester
Robert Lawther
David Lincoln
Timothy Mousseau
Richard Showman, Curriculum Committee Chair
Dick Vogt
Sally Woodin, Department Chair

USC Biology Graduate Students/Instructors:

Kyle Aveni-Deforge
Sarah Berke
Pamela Brannock
Andre Brock
Lisa Cox
Rebecca Cozart
Debra Davis
Kenny Fernandez
Sierra Jones
Jennifer Jost
Dino Marshalonis
Kenneth Oswald
Suzanne Pickard
Cliff Ramsdell
Sarah Refi
Denise Strickland
Lauren Szathmary
Michelle Vieyra

University of South Carolina:
Claudia Benitez-Nelson, USC Marine Science Program
Robert Johnson, USC College of Education, Department of Educational Psychology, Research, and Foundations
Loren Knapp, Assistant Dean, USC College of Arts and Sciences
Jonathan Singer, USC College of Education, Instruction and Teacher Education

Outside USC
Barbara Buckley, Concord Consortium
David Treagust, Curtin University, Australia

Scoring Guide for the
Universal Rubric for Lab Reports
developed by
Briana Timmerman
University of South Carolina
Department Biological Sciences

INSTRUCTIONS TO SCORERS:
Using the point values provided, indicate the appropriate level achieved for each criterion by the papers that have been provided to you. A technique some people find useful for scoring is called "the wedge." For each criterion, first decide if the response is in the top or bottom half of the rating scale. Then focus on the two criteria within that "half" of the scale. For example, if a response is in the upper half of the rating scale, is it more like the intermediate response or the proficient response? Once you decide on a rating (e.g. "intermediate"), does the description of the intermediate level match the response in question, or is it a little better or a little worse? **You may use "+" and "–" to indicate if a response does not fit clearly within a category.** For example, 2- would indicate a response whose quality most closely aligns with intermediate but which is slightly less proficient than a typical intermediate response; 1+ would indicate a response that is slightly better than the average novice response.

Student Proficiency Scoring Level			
0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+

A "Not addressed" response may either be completely missing (0) or something is written, but it is completely irrelevant (0+).

"Examples of student responses are presented as regular text in quotations." [If a portion of a student's response was removed to reduce the length of the text, a description of the edited material will appear in square brackets at the point where it was removed.] ***Suggestions about how to apply the criteria to those student responses follow each example in bold italics.***

****Remember, you are scoring, not grading.*** There is no expectation for a majority of the papers to achieve any particular score. Just place the quality of each paper as accurately as you can on the scales provided.

Support for this project was provided by NSF Award 0410992 to Timmerman.

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Introduction: Context				
Demonstrates a clear understanding of the big picture; Why is this question important/ interesting in the field of biology?	• The importance of the question is not addressed. • How the question relates within the broader context of biology is not addressed.	• The writer provides a generic or vague rationale for the importance of the question. • The writer provides vague or generic references to the broader context of biology.	• The writer provides one explanation of why others would find the topic interesting. • The writer provides some relevant context for the research question(s).	• a clear sense of why this knowledge may be of interest to a broad audience • writer provides explanation of gaps in understanding and how this research will help fill those gaps
Example of a novice (1-) response: "Gene sequencing is done by first PCR'ing the DNA and then cutting it with primers to determine the sequence of nucleotides. Many scientists like to know the sequence of genes so that they can perform other experiments on those organisms." ***("…so they can perform other experiments" is too vague – no sense of context is provided.)***				
Example of a novice (1) response: "Plant competition is an frequent ecological interaction." [Definition of intra vs. inter specific competition provided.] "There are several factors that influence competition between plant species: density of plants, and amount of available nutrients and sunlight….Tilman proposes that the outcome of competition can either be at equilibrium, or can oscillate over time." ***(While a reference to someone else's work is provided, there is no sense of why competition is worth studying or what the consequences are of various competitive outcomes (no sense of importance or broader context.)***				
Example of an intermediate (2) response: "One of biology's major ideas relates to the question of why do offspring resemble their parents." [Historical context is then followed by definition of phenotype and Mendel's three laws and a rationale for why the fruit fly is used as a model organism for genetics.] "The purpose of this experiment is to determine whether sex-linked traits are identifiable by the ratios of phenotypes in offspring. This experiment is important because by examining and analyzing inheritance patterns (i.e. using Punnett squares) scientists can determine the probability of inheriting certain traits." ***(implicit sense of importance, historical context and justification for organism provided)***				
Example of an proficient (3) response: "The effect of spore dispersal distance on fungal biodiversity is a subject that is superficially well-understood by scientists, however much of the intricate workings of genetic variation in individuals and in the resulting populations are yet to be discerned. Fungi have the ability to reproduce both sexually and asexually, but the relative impact of these propagation methods on their genome is unknown and paramount to comprehending their evolutionary and reproductive history. Such results would enlighten our understanding of the co-evolution of plants and fungi as well as provide a foundation for assessing the health of populations and fungal biodiversity." ***(broad context is provided as well as identification of gaps in knowledge and why they matter.)***				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Introduction: Accuracy and relevance				
Content knowledge is accurate, relevant and provides appropriate background for reader including defining critical terms. Background information:	• is missing or contains major inaccuracies. • is accurate, but irrelevant or too disjointed to make relevance clear • primary literature references are absent or irrelevant. May contain website or secondary references **websites or review papers are not primary**	• omits information or contains inaccurate information which detracts from the major point of the paper. • is overly narrow or overly general (only partially relevant). • Primary literature references if present inadequately explained.	• may contain minor omissions or inaccuracies that do not detract from the major point of the paper. • has the appropriate level of specificity to provide relevant context. • Primary literature references are relevant and adequately explained but few.	• is completely accurate • has the appropriate level of specificity to provide concise and useful context to aid the reader's understanding. • Primary literature references are relevant, adequately explained, and indicate a reasonable literature search.
Example of a not addressed (0+) response: "The Galapagos Islands are an archipelago consisting of 16 islands scattered over 36,000 square mile near the equator. The total land area is 3,028 square miles and the islands all rose form the ocean floor as the tops of volcanoes, possibly during the Plicoene era and have never been connected to the mainland ("Galapagos Ecosystem" website). ***(While accurate, this information is completely irrelevant to a report on evolution – the writer makes no connection between it and any other part of the report.)***				
Example of a novice (1) response: "Tilman describes the outcome of interspecific plant competition for a single resource as '…either a true equilibrium or it may be an oscillatory solution in which resources level and population densities fluctuate' (Tilman 1982). Here Tilman proposes that the outcome is either at equilibrium or the more competitive species fluctuates with its different advantages over time. This is a possibility we can look for when we test interspecific competition." ***(While a relevant primary reference is provided, the writer does not clarify it or explain it or its relevance in any meaningful way.)***				
Example of an intermediate response: "The main way that plants respond to competition is to change their root-shoot ratio (Turkington, 2005). Specifically, if plants are competing for resources primarily underground, they will allocate their growth to their roots. In doing this, they will gain competitive advantage and find more nutrients by foraging greater distances. Or, if competing for sunlight, they will allocate greater biomass to shoots so they will extend their branches above others. Root-shoot ratio is therefore a common measure of the effect of competition on plants." ***(a single primary reference which is relevant and adequately explained)***				
Example of an proficient response: "The systemin signaling pathway is a well-known defense pathway in plants and involves a cascade of mitogen activated protein kinases (MAPKs) which phosphorylate other molecules which then activate or inactivate the targeted molecule (Ryan 2000). MAPK targets are often transcription factors that turn on or off aspects of the wound response. They therefore have an important role in signal transduction upon wounding, pathogen attack or abiotic stress (Guillaume et al. 2001; Xing et al.2001) and are therefore useful indicators of wound responses." ***(multiple primary references whose relevance is clear)***				

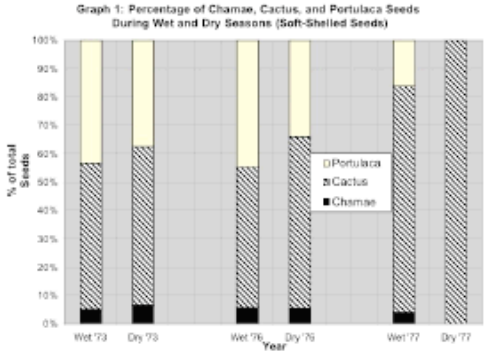
Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Hypotheses: Testable and consider alternatives				
Hypotheses are clearly stated, testable and consider plausible alternative explanations	• None indicated. • The hypothesis is stated but too vague or confused for its value to be determined • A clearly stated, but not testable hypothesis is provided. • A clearly stated and testable, but trivial hypothesis is provided.	• A single relevant, testable hypothesis is clearly stated • The hypothesis may be compared with a "null" alternative which is usually just the absence of the expected result.	• Multiple relevant, testable hypotheses are clearly stated. • Hypotheses address more than one major potential mechanism, explanation or factors for the topic.	• A comprehensive suite of testable hypotheses are clearly stated which, when tested, will distinguish among multiple major factors or potential explanations for the phenomena at hand.
Example of a novice (0+) response: "The hypothesis to explain the crisis event of 1977 that caused many finches to die and some to survive is a series of events." [Four sentences of descriptive narrative follows] ***(This could have been a (1) if the chain of events were identified as a series of sequential hypotheses (each with a null of no change). It would not count as a (2) because there are multiple explanations at each step in the chain of events which are not addressed. Linked single hypotheses are not multiple hypotheses relevant to the same issue.)***				
Example of a novice (1) response: "The purpose of the experiment is to determine if sex-linked traits produce different ratios of phenotypes than non-sex-linked traits. We hypothesize that the sex-linked trait will produce a 50/50 mix of red and white-eyed offspring. The null hypothesis is that there will be no difference between the ratio of the sex linked cross and the autosomal mono-hybrid cross." ***(This is a single relevant testable hypothesis compared to a null.)***				
Example of an intermediate (2-) response: "The hypotheses that have been formulated are that radishes and collards grown alone will have a higher germination rate, taller shoot heights and greater stem weights than the radish and collard mix. Also, we hypothesized that the radishes and collards at low density will be more successful with a higher germination rate, taller shoot heights and greater stem weights than radishes and collards at high density." ***(this earns a "2" because there are two specific, testable hypotheses which address multiple factors (each has an implicit null hypotheses of no difference), but a "-" because the factors are not identified or explained and no connections are made to the topic of inter-specific competition described in the introduction.)***				
Example of an proficient response (3): "We observe that dandelions in regularly mowed fields are shorter than those in un-mowed fields and determined that the height difference has a genetic basis. We hypothesize that this evolutionary shift is due to the selective pressure generated by mowing. Other explanations include genetic drift or mutation. If genetic drift is responsible for the height differences, than populations released from mowing will remain stunted on average due to reduced genetic diversity. If a mutation is responsible, than short populations will all contain unique genetic markers not found in un-mowed populations. If natural selection currently producing the shorter dandelions in the mowed field, than populations released from mowing should show a range of heights initially with a gradual increase in average height over time and un-mowed populations should respond to early season mowing with a decrease in average height at time of blooming." ***(multiple, plausible, interconnected hypotheses)***				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Hypotheses: Scientific merit				
Hypotheses have scientific merit. Hypotheses:	• are trivial, obvious, incorrect or completely off topic.	• are plausible and appropriate though likely or clearly taken directly from course material.	• indicate a level of understanding beyond the material directly provided to the student in the lab manual or coursework.	• are novel, insightful, or actually have the potential to contribute useful new knowledge to the field.
Example of a not addressed (0+) response: "The purpose of the experiment is to determine if sex-linked traits produce different ratios of phenotypes than non-sex-linked traits. We hypothesize that the sex-linked trait will produce a 50/50 mix of red and white-eyed offspring. The null hypothesis is that there will be no difference between the ratio of the sex linked cross and the autosomal mono-hybrid cross." ***(Student appears trapped by a "cookbook" assignment, but the question should at least be not whether sex-linked crosses have different ratios than autosomal (that is already known), but if the trait in question is sex-linked or autosomal.)***				
Example of a novice (1) or intermediate (2) response: "The hypotheses that have been formulated are that radishes and collards grown alone will have a higher germination rate, taller shoot heights and greater stem weights than the radish and collard mix. Also, we hypothesized that the radishes and collards at low density will be more successful with a higher germination rate, taller shoot heights and greater stem weights than radishes and collards at high density." ***(The merit of these hypotheses depends entirely on whether or not the factors of density and single vs. mixed species were provided to the students or derived by the students on their own. If the measures of germination rate, shoot height and stem weight were suggested by the instructor or lab manual, than this is a novice response (1). If the students identified and chose these on their own, then it is intermediate (2).)***				
Example of an proficient (3) response: "Previous predictions of species response to climate change do not consider the possibility that changes in environmental parameters such as air temperature could translate into non-linear changes in organismal body temperature (Somero 2005). Here we use a biophysical model to explore the relationship between two key climatic parameters (air and water temperature) and changes to the body temperature of an ecologically dominant species of intertidal invertebrate (the mussel Mytilus californianus) across 1600 km of its geographic range. We hypothesize that the relative importance of air and water temperature to body temperature varies both geographically and with vertical position within locations, thus contradicting past assumptions that a simple change in air or water temperature will have a consistent effect across the species distribution." ***(The proposed hypothesis, if correct, would challenge a widely held assumption currently in use by others working in this field (that organisms' body temperatures respond in simple, linear ways to air/water temperature) and would require other scientists to rethink their experimental design and results. Thus this hypothesis would make a major contribution to the field and likely change how others in the field think about these types of research questions.)***				

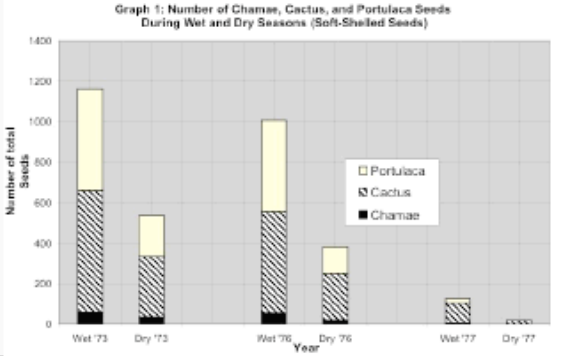
Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Methods: Controls and replication				
Appropriate controls (including appropriate replication) are present and explained. If student designed the experiment:	• Controls and/ors replication are nonexistent, • Controls and/or replication may have been present, but just not described or • Controls and/or replication were described but were inappropriate.	• Controls consider one major relevant factor • Replication is modest (weak statistical power).	• Controls take most relevant factors into account • Controls include positive and negative controls if appropriate • Replication is appropriate (average sample size with reasonable statistical power).	• Controls consider all relevant factors • Controls have become methods of differentiating between multiple hypotheses. • Replication is robust (sample size is larger than average for the type of study).
If instructor designed the experiment:	• Student fails to mention controls and/or replication or mentions them, but the description or explanation is incomprehensible.	• Student explanations of controls and/or replication are vague, inaccurate or indicate only a rudimentary sense of the need for controls and or replication	• Student evidences a reasonable sense of why controls/ replication matter to this experiment • Explanations are mostly accurate.	• Explanations of why these controls matter to this experiment are thorough, clear and tied into sections on assumptions and limitations
Example of a not addressed (0+) response: "Two culture tubes were obtained with three females and two males placed in each tube. After two weeks, the phenotypes of the newly hatched offspring were recorded and a statistical test used to determine if the data matched the hypothesis." ***(Replication is negligible. Results section sample size indicates that class data were pooled, but student appears to not recognize the importance of this distinction; no discussion of controls such as ensuring that the females were virgin.***				
Example of an intermediate (2-) response: "For the purposes of testing the toxicity of statin to over the entire Amphiascus lifecycle, individuals were raised in separate environments from nauplius stage through adulthood. Seawater (30 ppt) was filtered (0.45 micron) and changed every three days. Culture checks, water changes and feeding schedules were uniform across samples. Three 96-well plates (n=288 individuals) were used per treatment and for the control. Nauplii ages were synchronized before being loaded into the wells. (Coolidge, Howard Hughes August 2005) ***(Multiple controls are present and replication is reasonable, though acknowledgement of a potential for "pseudo-replication" is not addressed (all wells on same plate might be considered a single sample). Only an explanation of why the controls are necessary is lacking.)***				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Methods: Experimental design				
Experimental design likely to produce salient and fruitful results (tests the hypotheses posed.) Methods are:	• inappropriate • poorly explained / indecipherable	• appropriate • clearly explained • drawn directly from coursework • not modified where appropriate	• appropriate • clearly explained • modified from coursework in appropriate places • or drawn directly from a novel source (outside the course)	• appropriate • clearly explained • a synthesis of multiple previous approaches or an entirely new approach
Example of a novice (1-) response: "We placed the following seed groupings each in their own pots: 8 radish seeds, 16 radish seeds, 32 radish seeds, 8 collard seeds, 16 collard seeds, 32 collard seeds, 4 radish + 4 collard seeds, 8 radish + 8 collard, 16 radish + 16 collard….[care and feeding instructions]…we counted the number of plants germinated weekly. After the fifth week we also measured average number leaves per individual, average leaf weight per individual, average shoot length, average shoot weight and stem density. Our data was [sic] combined with other students' data and an ANOVA test with two-factor replication on Microsoft Excel was used." ***(Although the experimental design is appropriate for the hypotheses given earlier in the paper and relatively clear, the 'recipe-like' quality of the writing indicates that the student had very little cognitive input to this experimental design. The student provides no sense of how these methods relate to the various hypotheses.)***				
Example of an intermediate or proficient (2 or 3) response: "Fibroblasts were isolated from neonatal rat heart tissue minced and digested with collagenase (Carver 2003). Those designated for mechanical stretch were plated onto silastic membranes and allowed to reach confluence before glucose treatment (Carver 1991)." *(Zhang HH Aug 2005)* ***(If this student originated the idea to combine the methodologies from the 2003 and 1991 papers, this is an proficient response. If the mentor (Carver) suggested the new protocol than it is an intermediate response.)*** ***This criterion obviously requires familiarity with the course and/or students' background in order to judge the quality of the student's response.***				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Results: Data selection				
Data are comprehensive, accurate and relevant	• Data are too incomplete or haphazard to provide a reasonable basis for testing the hypothesis	• At least one relevant dataset per hypothesis is provided but some necessary data are missing or inaccurate • Reader can satisfactorily evaluate some but not all of writer's conclusions.	• Data are relevant, accurate and complete with gaps being minor. • Reader can fully evaluate whether the hypotheses were supported or rejected with the data provided.	• Data are relevant, accurate and comprehensive. • Reader can fully evaluate validity of writer's conclusions and assumptions. • Data may be synthesized or manipulated in a novel way to provide additional insight.



The graph on the LEFT is a student graph. The graph on the RIGHT is an improved version. The student's hypothesis is that finches died because there were not enough seeds to eat.



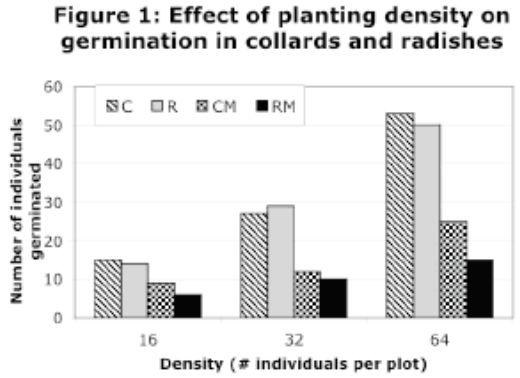
Example of a novice response: *The data selection on the LEFT is a novice response (1-) because the data only show how seed types are changing, with no indication of how seed abundance changed. (In other words, it gives the impression that there are plenty of cactus seeds.)*

Example of an intermediate (2): *The figure on the RIGHT communicates a much more complete message showing both the change in seed type and the decline of overall seed numbers making it clear that not only were the finches affected by a change in seed type, but that there simply wasn't much to eat.*

Example of an proficient response (3): *If the figure on the RIGHT also included the relative proportion and total abundance of hard-shelled seeds (as well as soft-shelled) the reader would then have a complete picture of how the finches' food sources changed over time.*

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Results: Data presentation				
Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete. Presentation of data:	• is missing labels or units which prevent the reader from being able to derive any useful information from the graph. • is in an inappropriate format or graph type • Captions are confusing or indecipherable.	• contains some errors in or omissions of labels, scales, units etc., but the reader is able to derive some relevant meaning from each figure. • is technically correct but inappropriate format prevents the reader from deriving meaning or using it. Captions are missing or inadequate	• contains only minor mistakes that do not interfere with the reader's understanding and the figure's meaning is clear without the reader referring to the text. • graph types or table formats are appropriate for data • includes captions that are at least somewhat useful.	• contains no mistakes • uses a format or graph type which highlights relationships between the data points or other relevant aspects of the data. • may be elegant, novel, or otherwise allow unusual insight into data • has informative, concise and complete captions.

Example of a novice (1-) and/or intermediate (2) response: "Figure 1 showed the effects of competition on the germination rates among radishes and collards. There is notable interspecific and intraspecific competition….[an] increase in the density of plants leads to a decrease in the rate of germination."

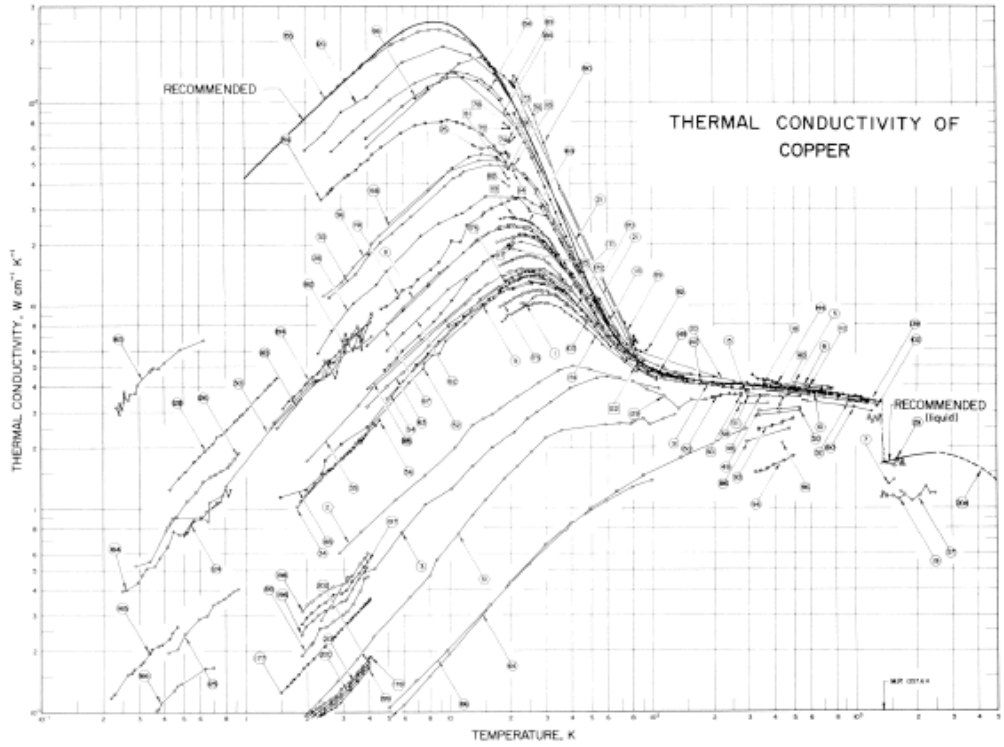


(This figure is complete and accurate with all items labeled, but earns (1-) because it does not actually report the data referenced in the text (germination rate) forcing the reader to try and calculate % germination for each data point making this otherwise acceptable graph fairly useless.

If the absolute number of plants germinated were useful information, and what was referenced in the text, then this would be an INTERMEDIATE RESPONSE (2). Suggestions for making it a 2+ include writing out what "C," "R," "CM" and "RM" stand for, adding error bars and indicating which categories are significantly different, darkening the gridlines to make reading values off the y-axis easier and providing more contrast between the categories (lightening "R").

See the next page for an example of an proficient response.

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
				Proficient response (3): *(This graph takes an immense amount of information (the values reported in over 200 papers – each line represents a set of data reported in a single paper with the paper's citation indicated by the number in the circle attached to the line) and summarizes it in an immediate and truly effective way. The reader can now make comparisons and conclusions that would have been impossible had the data been presented in any other way.)* *Graph from Yo, Powell and Liley. 1974. Thermal conductivity of the elements: A comprehensive review. J Physical Chem Reference Data 3(supplement 1): 1-244 as illustrated in Tufte, E. 2001. The visual display of quantitative information. Graphics Press: Cheshire CT (p. 49)*



Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Results: Statistical analysis				
Statistical analysis is appropriate for hypotheses tested and appears correctly performed with relevant values reported and explained.	• No statistical analysis is performed. • Statistics are provided but are inappropriate, inaccurate or incorrectly performed or interpreted so as to provide no value to the reader.	• Appropriate, accurate descriptive statistics only are provided. • Inferential statistics are provided but either incorrectly performed or interpreted or an inappropriate test was used. • Appropriate, correct inferential statistics are provided, but lack sufficient explanation.	• Appropriate inferential (comparative) statistical analysis is properly performed and reasonably well explained. • Explanation of significant value may be limited or rote (e.g. use of $p<0.05$ only)	• Statistical analysis is appropriate, correct and clearly explained • includes a description of what constitutes a significant value and why that value was chosen as the threshold (may choose values beyond $p<0.05$).
Example of a (0+) response: "The average beak size of the finches increased from 10.1 to 10.7 from 1973 to 1978 as seen in graph 1." *(BET) (This response barely earns a (0+) because no actual statistical test is performed and the writer fails to provide units for the change in beak size leaving the reader completely unable to assess the potential significance of the change (an increase of 0.6 cm is very different than 0.6 mm))*				
Example of a novice (1) response: "The p-value for the autosomal hybrid cross is less than 0.05 and the value for the sex-linked monohybrid is less than 0.05 for the males, but larger than 0.05 for the females. Thus, the hypothesis for the autosomal monohybrid cross can be accepted, but the hypothesis for the sex-linked cross must be rejected." *(While an appropriate test was performed, the writer's understanding and explanation of the test is limited and vague.)*				
Example of an intermediate (2+) response: "The plant heights were significantly different when compared by a 2 factor ANOVA, thus supporting our hypothesis. It should be noted that the p-value generated by these samples was 0.049 however. As we used the conventional threshold for significance ($p=0.05$) our result is borderline in its significance should not be as strongly considered as our second result which generated a p value of 0.0037." (BET) *(The writer simply used the conventional 0.05 cut-off for significance, but her follow-up sentence indicates that she realizes significance is a spectrum, not an either/or situation.)*				
Example of an proficient (3) response: "The Bonferroni method of comparing least squared means was used to analyze the developmental period from nauplius to adult and indicated a significant difference between the control and all treatment groups of about 1 day (control was 15.11$\pm$ 1.34 days vs. 16.14$\pm$ 1.31 days (50 ng/L treatment) to 16.39 $\pm$ 1.72 days (250 ng/L) ($p<0.0001$ for each comparison). This experiment demonstrates the effectiveness of simvastatin to alter the development of Amphiascus, even at minute concentrations. The ecological impact of this effect is uncertain however. Survivorship did not differ among groups, but the change in timing could have implications for reproductive success." *(Coolidge HH Aug 2005) (This response earns (3) because the statistics are appropriate, well explained and the significance of the result viewed in a broader, biologically relevant context.)*				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Discussion: Conclusions based on data selected				
Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.	• Conclusions have little or no basis in data provided. • Connections between hypothesis, data and conclusion are non-existent, limited, vague or otherwise insufficient to allow reasonable evaluation of their merit. • Conflicting data are not addressed.	• Conclusions have some direct basis in the data, but may contain some gaps in logic or data or are overly broad. • Connections between hypothesis, data and conclusions are present but weak. • Conflicting or missing data are poorly addressed.	• Conclusions are clearly and logically drawn from and bounded by the data provided with no gaps in logic. • A reasonable and clear chain of logic from hypothesis to data to conclusions is made. • Conclusions attempt to discuss or explain conflicting or missing data.	• Conclusions are completely justified by data. • Connections between hypothesis, data, and conclusions are comprehensive and persuasive. • Conclusions address and logically refute or explain conflicting data • Synthesis of data in conclusion may generate new insights.
Example of a (0+) response: "The Chi-square value was zero which would indicate that the observed and expected values were equal. The results did not show exactly a 3:1 ratio for the autosomal cross and a 1:1 for the sex-linked cross, but the two crosses were definitely different types of ratios." *(Either statistics show the observed values to be different from the expected values or not, the ratios cannot be "different" if the Chi-squared was non-significant. The only merit of this response is that it wasn't left blank.)*				
Example of a novice (1-) response: "Graph 2 indicates the number of tribulus (hard-shelled) seeds produced during the wet and dry seasons. This supports the hypothesis by showing that more tribulus seeds were eaten by the large beaked finches during the dry seasons." *(While the number of seeds is a relevant dataset for determining why the finches died, a graph showing seed production does not provide any direct knowledge of the number of seeds eaten by finches and there are no data in the project at all which tell what size birds consumed the most seeds (no one collected seed shells out of bird droppings and reported those numbers).)*				
Example of an intermediate (2) response: "As can be seen in Fig. 6, the allele frequencies for the T/T genotypes have greatly increased from 1995 to 2005, especially in the northern regions. There were significant increases at Muir Beach, Half Moon Bay and Santa Cruz sites. Correspondingly, the frequency of G/G genotypes is reduced at the northern sites and unchanged in the southern sites. These data refute the hypothesis that hybridization is responsible for the decline in G/G, but support the hypothesis that M. galloprovincialis is retreating southward down the coast." *(Jones HH Aug 2005) (Specific data are directly linked to a logical conclusion and an alternative explanation is refuted with data. More detail about the trends shown in the data, which comparisons are significant and a greater discussion of ambiguous or conflicting data would be required for this to be an proficient response.)*				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Discussion: Alternative explanations				
Alternative explanations are considered and clearly eliminated by data in a persuasive discussion. Alternative explanations:	• are not provided • are trivial or irrelevant • are mentioned but not discussed or eliminated.	• are provided in the discussion only • may include some trivial or irrelevant alternatives. • Discussion addresses some but not all of the alternatives in a reasonable way.	• Some alternative explanations are tested as hypotheses; those not tested are reasonably evaluated in the discussion section. • Discussion of alternatives is reasonably complete, uses data where possible and results in at least some alternatives being persuasively dismissed.	• have become a suite of interrelated hypotheses that are explicitly tested with data. • Discussion and analysis of alternatives is based on data, complete and persuasive with a single clearly supported explanation remaining by the end of the discussion.
Example of a (0+) response: "This hypothesis [a narrative chain of 4-5 separate ideas] is supported by much data and no refuting evidence could be found. Other viable hypotheses could also serve to explain the finch crisis of 1977 such as the conjecture that in the dry season of 1977 the overall finch population decreased due to a decrease in overall seed production. There is no evidence to refute this idea, however it is not as strong as the first hypothesis because it is very general and does not specify what type of seeds decreased in number, nor what type of finches were affected." *(The "alternative" (as the student states) is not a viable or plausible hypothesis and therefore not worth points. Alternatives must be realistic – something for which another student might argue.)*				
Example of a novice (1-) response: "The collard mix had a significantly lower stem weight in the medium density pot than the collards alone. This data supports my first hypothesis that radishes and collards grown alone will have greater stem weights than the mixed pots. However, radishes had the opposite result, the radishes in the mix had greater stem weights than radishes alone. Possible explanations are that the majority of radish growth occurs underground; therefore the majority of their resources are allocated to the roots." *(While the first sentence does support the hypothesis, the student does not discuss what happened in the other collard pots (ignoring data weakens the conclusion) and the third sentence makes little sense, why would allocating the majority of resources below ground produce greater stem weights?)*				
Example of an intermediate (2) response: "It is suggested by the data that the M. galloprovincialis population is retreating down the coast to more southern sites. It is also possible that the decline in the G/G genotype in the north is due to G/G mollusks cross-breeding with the T/T genotype mussels to yield more G/T genotypes. The data did not support this alternative hypothesis however. The G/T frequencies at northern sites were similar to those found in 1995." *(Jones HH Aug 2005) (Two reasonable explanations are compared and one rejected because the data refute it (the data supporting the other idea are provided further along in the paper).*				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Discussion: Limitations of design				
Limitations of the data and/or experimental design and corresponding implications discussed. Limitations:	• are not discussed.	• are discussed in a trivial way (e.g. "human error" is the major limitation invoked).	• are relevant, but not addressed in a comprehensive way • Conclusions fail to address or overstep the bounds indicated by the limitations.	• are presented as factors modifying the author's conclusions. • Conclusions take these limitations into account.
Example of a novice (1) response: "Environmental factors may have led to the deaths of many flies. It cannot be confirmed that the environment in which the Drosophila were kept was constant or that the correct flies were put in the vials." *(This is the generic "human error" response. These limitations are trivial because they are simply the specter of failed controls or sloppy work and unless the student has specific evidence that this has occurred, it is not useful to say that these might be sources of error.)*				
Example of an intermediate (2) response: "Even though the lab only took into account the flies that were alive, we could not conclude that the dead flies died in the same phenotypic ratio as the live flies. We should have counted the dead flies as well." *(This student mentions this as an after-thought and while insightful and correct, the student's failure to actually modify her conclusion based on this thought make the discussion more trivial than useful. If she had modified her conclusion based on this, it would have been rated more highly.*				
Example of an proficient (3) response: "A major limitation of this of this experiment is that it simplified the actual environmental conditions in the copepods' habitat. The paper that provided the benchmark concentrations for copepod toxins (Miao and Metcalfe 2003) mentioned that in addition to simvastatin, three other statin medications were observed in sewage effluent: pravastatin, lovastatin and atorvastatin. So while simvastatin caused significant delay in copepod development across the three concentrations tested, future research should examine the effects of the mixtures of these four statin medications in a single assay to determine their combined effect on copepod development." *(This response earns a (3) because not only does the author recognize a major limitation of the experimental design and use it to modify his conclusions, but the limitation comes from the relationship of this research to other research and places this work clearly in context with other relevant research.)*				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Discussion: Significance of research				
Paper gives a clear indication of the significance of the research and its future directions. <u>Future directions and significance of this research:</u>	• are not addressed.	• are vague, implausible (not possible with current technologies or methodologies), trivial or off topic.	• are useful, but indicate incomplete knowledge of the field (e.g. suggest something that's already known or unknowable given current methodologies) • suggest a fruitful line of research, but fail to provide enough detail to indicate what the implications of those future results would be (what is the motivation for doing the future research?)	• are salient, plausible and insightful • suggest work that would fill knowledge gaps and move the field forward.
Example of a novice (1-) response: "If I were to recreate this experiment, I would allow the plants to grow for an extended period of time in order to allow their advantageous characteristics to take even more of an effect on the results." ***(This might be an effective follow-up experiment, but the description is too vague for its merit to be judged – What does extended mean? Just for a few more days? Until flowering? Until they die? What kind of "effect" on the results is the writer hoping for? Was the current result borderline in its statistical significance?***				
Example of an intermediate (2) response: "This paper reports on the effect of simvastatin on copepods. Future research should examine the effects of mixtures of the four statin medications in a single assay to determine their combined effect on copepod development." *(Coolidge HH Aug 2005)* ***(While the future directions suggested are relevant and useful and indicate a knowledge of the field, this is the last sentence in the paper and so the writer does not indicate what a knowledge of the combined effects would offer. Is it simply more complete? Likely to show completely different patterns? Are the results likely to be additive or synergistic?)***				
Example of an proficient (3) response: "Future research should attempt to find a sub-micro molar inhibitor. This will allow a more indepth study of the processes that PRMT's control or with which they are involved. As PRMTs are suspected to control nuclear receptors, such as the estrogen receptor, cytokine and p53 controlled transcription, determination of PRMT's role in these processes and would make them ideal targets for new breast cancer and heart disease therapies." *(Smith HH Aug 2005).* ***(The motivation for continuing research in this area is clear and the writer outlines the steps and rationale in sufficient detail.***				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Use of Primary Literature				
Relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided). Primary literature is defined as: - peer reviewed - reports original data - authors are the people who collected the data. - published by a non-commercial publisher. **Primary literature references:**	• are not included.	• are limited (only one or two primary references in the whole paper) • References to the textbook, lab manual, or websites may occur. • Citations are at least partially correctly formatted. Note that proper format includes a one-to-one correspondence between in-text and end of text references (no references at end that are not in text and vice versa).	• are more extensive (at least one citation for each major concept) • Literature cited is predominantly (≥ 90%) primary literatures. • are used primarily to provide background information and context for conclusions • are properly cited in the paper.	• indicate an extensive literature search was performed. • frame the question in the introduction by indicating the gaps in current knowledge of the field. • are used in the discussion to make the connections between the writer's work and other research in the field clear • are properly and accurately cited
Example of a novice (1+) response: "Research done on fifty-two lines of Drosophila showed that environmental stress can have an effect on phenotypic variance, genetic variance and survivorship; cold temperature and high density can cause a reduction in survivorship (Fowler 681)." [another sentence on a different topic follows] *(This reference is appropriate and relevant. It scores as a novice response though for two reasons: 1) it is improperly cited (uses MLA instead of scientific citation format) and 2) the student does not relate Fowler and Whitlock's highly relevant results to her own despite the fact that she later talks about how if one phenotype had a lower survivorship than the other, that could have thrown off her results.)*				
Example of an intermediate (2+) response: "Amphiascus tenuiremis (a species of copepod) is a model organism that has proved useful for testing environmental toxicity of chemicals that enter aquatic environments (e.g. Chandler et al. 2004; de Fur 2004; Chandler and Greene 1996)." *(Coolidge HH Aug 2005) (Both of the Chandler references are examples of research using the same organism to test toxicity of chemicals in aquatic environments and the de Fur article is a literature review of the use of invertebrates in endocrine disruption studies – thus the literature is all primary, relevant and sets the context of the writer's work. This paper uses primary literature in similarly effective ways in the introduction, methods and discussion. The only thing preventing this response from being rated as PROFICIENT was that the writer used the wrong format (reversing author's names both in-text and at the end of the text) creating confusion for the reader and only 7 references were cited in the whole paper indicating that the literature search while notable was not extensive.*				

Criteria	0 Not addressed 0+	1 - Novice 1+	2 - Intermediate 2+	3 - Proficient 3+
Writing quality				
Grammar, word usage and organization facilitate the reader's understanding of the paper. (some elements inspired by the SC State Dept. Education Extended-Response Scoring Guide for English Language Arts.)	• Grammar and spelling errors detract from the meaning of the paper. • Word usage is frequently confused or incorrect. • Subheadings are not used or poorly used. • Information is presented in a haphazard way.	• Grammar and spelling mistakes do not hinder the meaning of the paper. • General word usage is appropriate, although use of technical language is may have occasional mistakes. • Subheadings are used and aid the reader somewhat. • There is some evidence of an organizational strategy though it may have gaps or repetitions.	• Grammar and spelling have few mistakes. • Word usage is accurate and aids the reader's understanding. • Distinct sections of the paper are delineated by informative subheadings. • A clear organizational strategy is present with a logical progression of ideas.	• Correct grammar and spelling. • Word usage facilitates reader's understanding. • Informative subheadings significantly aid reader's understanding. • A clear organizational strategy is present with a logical progression of ideas. There is evidence of an active planning for presenting information; this paper is easier to read than most.
As this criterion encompasses the whole paper, no specific examples are provided.				

Appendix 5. Criteria list and instructions given to Natural Raters

For each criterion, please award points based on the following scale:

Points	Achievement level
0	Not addressed (student did not do this) = F or D
1	Novice (student performed task at a minimal level) = C
2	Intermediate (student performed task to a reasonable level) = B, B+ or A
3	Proficient (student performed task exceptionally well (a rare performance) = A+

You may augment the points if you wish (use "-" or "+" to indicate relative performance within a level). **Please do not feel the need to produce any type of a grade distribution** (do not expect a certain number of A's vs. C's, etc. - this is a random selection of papers.)

Introduction	**Context** Demonstrates a clear understanding of the big picture; Why is this question important/ interesting in the field of biology?
	Accuracy and relevance Content knowledge is accurate, relevant and provides appropriate background for reader including defining critical terms.
Hypotheses	**Testable and consider alternatives** Hypotheses are clearly stated, testable and consider plausible alternative explanations
	Scientific merit Hypotheses have scientific merit.
Methods	**Controls and replication** Appropriate controls (including appropriate replication) are present and explained.
	Experimental design Experimental design is likely to produce salient and fruitful results (actually tests the hypotheses posed.)
Results	**Data selection** Data chosen are comprehensive, accurate and relevant.
	Data presentation Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete.
	Statistical analysis Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained.
Discussion	**Conclusions based on data selected** Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.
	Alternative explanations Alternative explanations (hypotheses) are considered and clearly eliminated by data in a persuasive discussion.
	Limitations of design Limitations of the data and/or experimental design and corresponding implications for data interpretation and conclusions are discussed.
	Implications of research Paper gives a clear indication of the implications and direction of the research in the future.
Use of Primary Literature	**Primary Literature** Writer provides a relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided) using primary literature. Primary literature is defined as: • peer reviewed • reports original data • authors are the people who collected the data. • Journal produced by a non-commercial scientific association
Writing Quality	**Writing Quality** Grammar, word usage and organization facilitate the reader's understanding of the paper.

CLASSROOM TEST OF

SCIENTIFIC REASONING

Multiple Choice Version

Directions to Students:

This is a test of your ability to apply aspects of scientific and mathematical reasoning to analyze a situation to make a prediction or solve a problem. Make a dark mark on the answer sheet for the best answer for each item. If you do not fully understand what is being asked in an item, please ask the test administrator for clarification.

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

Revised Edition: August 2000 by Anton E. Lawson, Arizona State University. Based on: Lawson, A.E. 1978. Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1): 11-24.

1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*

   a. The pancake-shaped piece weighs more than the ball
   b. The two pieces still weigh the same
   c. The ball weighs more than the pancake-shaped piece

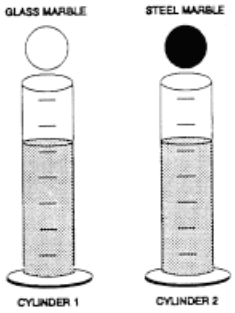
2. *because*

   a. the flattened piece covers a larger area.
   b. the ball pushes down more on one spot.
   c. when something is flattened it loses weight.
   d. clay has not been added or taken away.
   e. when something is flattened it gains weight.

3. To the right are drawings of two cylinders filled to the same level with water. The cylinders are identical in size and shape.

   Also shown at the right are two marbles, one glass and one steel. The marbles are the same size but the steel one is much heavier than the glass one.

   When the glass marble is put into Cylinder 1 it sinks to the bottom and the water level rises to the 6th mark. *If we put the steel marble into Cylinder 2, the water will rise*

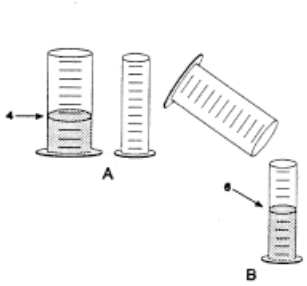


   a. to the same level as it did in Cylinder 1
   b. to a higher level than it did in Cylinder 1
   c. to a lower level than it did in Cylinder 1

4. *because*

   a. the steel marble will sink faster.
   b. the marbles are made of different materials.
   c. the steel marble is heavier than the glass marble.
   d. the glass marble creates less pressure.
   e. the marbles are the same size.

5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

a. to about 8
b. to about 9
c. to about 10
d. to about 12
e. none of these answers is correct

6. *because*

a. the answer can not be determined with the information given.
b. it went up 2 more before, so it will go up 2 more again.
c. it goes up 3 in the narrow for every 2 in the wide.
d. the second cylinder is narrower.
e. one must actually pour the water and observe to find out.

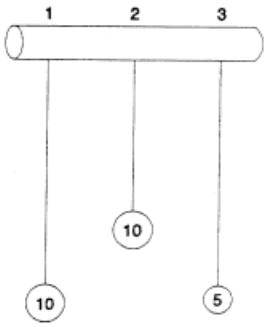
7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

a. to about 7 1/2
b. to about 9
c. to about 8
d. to about 7 1/3
e. none of these answers is correct

8. *because*

a. the ratios must stay the same.
b. one must actually pour the water and observe to find out.
c. the answer can not be determined with the information given.
d. it was 2 less before so it will be 2 less again.
e. you subtract 2 from the wide for every 3 from the narrow.

9. At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.



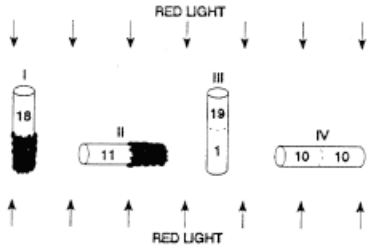
Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. *Which strings would you use to find out?*

a. only one string
b. all three strings
c. 2 and 3
d. 1 and 3
e. 1 and 2

10. *because*

a. you must use the longest strings.
b. you must compare strings with both light and heavy weights.
c. only the lengths differ.
d. to make all possible comparisons.
e. the weights differ.

11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



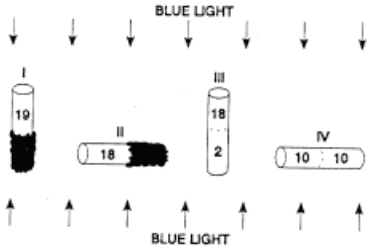
This experiment shows that flies respond to (respond means move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

12. *because*

a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
b. most flies did not go to the bottom of Tubes I and III.
c. the flies need light to see and must fly against gravity.
d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
e. some flies are in both ends of each tube.

13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.



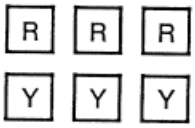
These data show that these flies respond to (respond means move to or away from):

a. blue light but not gravity
b. gravity but not blue light
c. both blue light and gravity
d. neither blue light nor gravity

14. *because*

a. some flies are in both ends of each tube.
b. the flies need light to see and must fly against gravity.
c. the flies are spread about evenly in Tube IV and in the upper end of Tube III.
d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
e. most flies are in the upper end of Tube I and the lighted end of Tube II.

15. Six square pieces of wood are put into a cloth bag and mixed about. The six pieces are identical in size and shape, however, three pieces are red and three are yellow. Suppose someone reaches into the bag (without looking) and pulls out one piece. *What are the chances that the piece is red?*

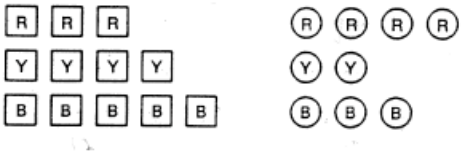


a. 1 chance out of 6
b. 1 chance out of 3
c. 1 chance out of 2
d. 1 chance out of 1
e. cannot be determined

16. *because*

a. 3 out of 6 pieces are red.
b. there is no way to tell which piece will be picked.
c. only 1 piece of the 6 in the bag is picked.
d. all 6 pieces are identical in size and shape.
e. only 1 red piece can be picked out of the 3 red pieces.

17. Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece.



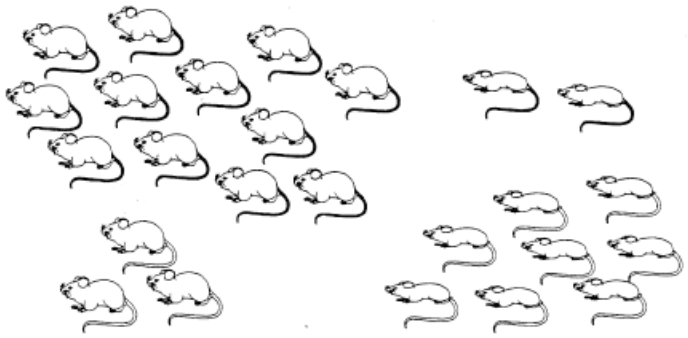
What are the chances that the piece is a red round or blue round piece?

a. cannot be determined
b. 1 chance out of 3
c. 1 chance out of 21
d. 15 chances out of 21
e. 1 chance out of 2

18. *because*

a. 1 of the 2 shapes is round.
b. 15 of the 21 pieces are red or blue.
c. there is no way to tell which piece will be picked.
d. only 1 of the 21 pieces is picked out of the bag.
e. 1 of every 3 pieces is a red or blue round piece.

19. Farmer Brown was observing the mice that live in his field. He discovered that all of them were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. Below are the mice that he captured.



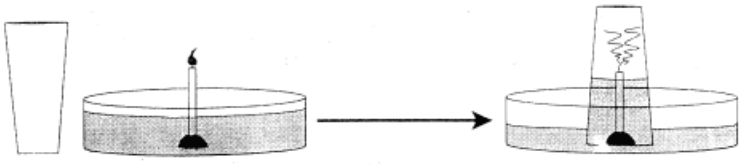
Do you think there is a link between the size of the mice and the color of their tails?

a. appears to be a link
b. appears not to be a link
c. cannot make a reasonable guess

20. *because*

a. there are some of each kind of mouse.
b. there may be a genetic link between mouse size and tail color.
c. there were not enough mice captured.
d. most of the fat mice have black tails while most of the thin mice have white tails.
e. as the mice grew fatter, their tails became darker.

21. The figure below at the left shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle, and placed in the water, the candle quickly goes out and water rushes up into the glass (as shown at the right).



This observation raises an interesting question: Why does the water rush up into the glass?

Here is a possible explanation. The flame converts oxygen into carbon dioxide. Because oxygen does not dissolve rapidly into water but carbon dioxide does, the newly formed carbon dioxide dissolves rapidly into the water, lowering the air pressure inside the glass.

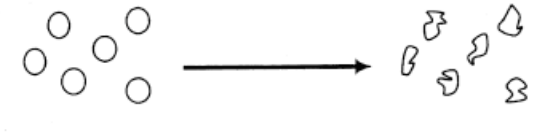
Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). *Using some or all of the materials, how could you test this possible explanation?*

a. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
b. The water rises because oxygen is consumed, so redo the experiment in exactly the same way to show water rise due to oxygen loss.
c. Conduct a controlled experiment varying only the number of candles to see if that makes a difference.
d. Suction is responsible for the water rise, so put a balloon over the top of an open-ended cylinder and place the cylinder over the burning candle.
e. Redo the experiment, but make sure it is controlled by holding all independent variables constant; then measure the amount of water rise.

22. What result of your test (mentioned in #21 above) would show that your explanation is probably wrong?

a. The water rises the same as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.

23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.



Magnified Red Blood Cells After Adding Salt Water

This observation raises an interesting question: Why do the red blood cells appear smaller?

Here are two possible explanations: I. Salt ions (Na^+ and Cl^-) push on the cell membranes and make the cells appear smaller. II. Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag, placing it in a salt solution for ten minutes and then reweighing the bag.

What result of the experiment would best show that explanation I is probably wrong?

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

24. *What result of the experiment would best show that explanation II is probably wrong?*

a. the bag loses weight
b. the bag weighs the same
c. the bag appears smaller

Introduction:

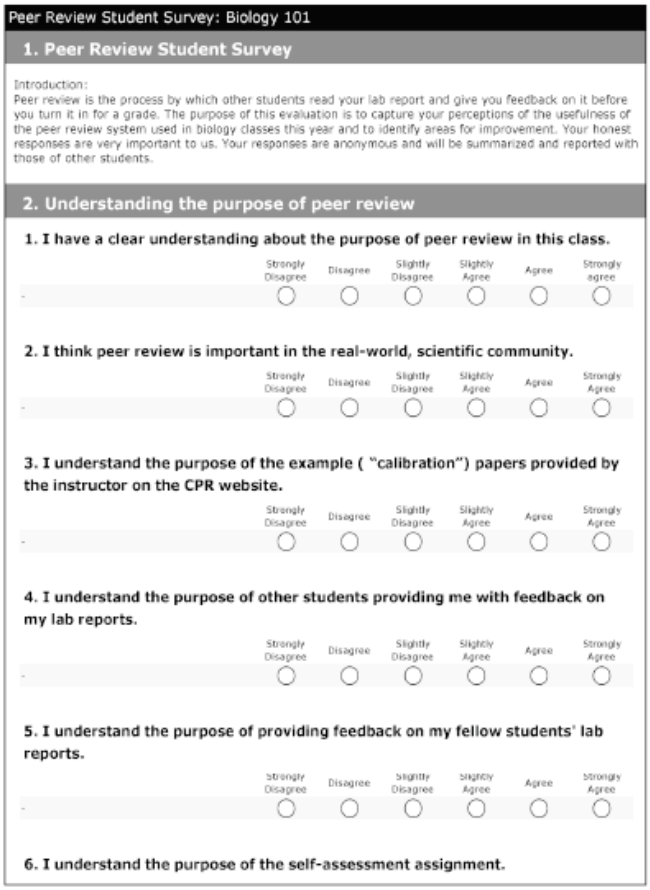
Peer review is the process by which other students read your lab report and give you feedback on it before you turn it in for a grade. The purpose of this evaluation is to capture your perceptions of the usefulness of the peer review system used in biology classes this year and to identify areas for improvement. Your honest responses are very important to us. Your responses are anonymous and will be summarized and reported with those of other students.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Understanding the *purpose* of peer review						
1. I have a clear understanding about the purpose of peer review in this class.	☐	☐	☐	☐	☐	☐
2. I understand the importance of peer review in the real-world, scientific community.	☐	☐	☐	☐	☐	☐
3. I understand the usefulness of the "calibration" papers assignment.	☐	☐	☐	☐	☐	☐
4. I understand the usefulness of other students providing feedback on my lab reports.	☐	☐	☐	☐	☐	☐
5. I understand the usefulness of providing feedback on my fellow student's lab reports.	☐	☐	☐	☐	☐	☐
6. I understand the usefulness of the self-assessment assignment.	☐	☐	☐	☐	☐	☐
Understanding the *process* of peer review						
7. My TA clearly explained the usefulness of peer review for helping me with my lab report.	☐	☐	☐	☐	☐	☐
8. My TA clearly explained why peer review skills will be useful to me later.	☐	☐	☐	☐	☐	☐
9. My TA clearly explained how peer review is used by scientists in the real world.	☐	☐	☐	☐	☐	☐

10. I have a clear understanding about how to use the peer review system.	☐	☐	☐	☐	☐	☐
11. I was provided with adequate training by my TA to successfully utilize the peer review system.	☐	☐	☐	☐	☐	☐
12. The handout I received about peer review was easy to understand.	☐	☐	☐	☐	☐	☐
13. The criteria I am supposed to use for evaluating other people's papers were understandable and easy to use.	☐	☐	☐	☐	☐	☐
14. The peer-review website was user friendly.	☐	☐	☐	☐	☐	☐
15. The time required for peer review is manageable.	☐	☐	☐	☐	☐	☐
16. I felt motivated to participate in peer review in this class.	☐	☐	☐	☐	☐	☐
Impact of the peer review system						
17. Peer review helped me improve my lab report.	☐	☐	☐	☐	☐	☐
18. Peer review facilitated my understanding of an in-class experiment.	☐	☐	☐	☐	☐	☐
19. The peer review process will be helpful in my other classes.	☐	☐	☐	☐	☐	☐
20. Peer review helped me develop my *writing* skills.	☐	☐	☐	☐	☐	☐
21. Peer review helped me develop my *editing* skills.	☐	☐	☐	☐	☐	☐
22. Peer review helped me develop my *critical thinking* skills.	☐	☐	☐	☐	☐	☐
23. Peer review helped me develop my *research* skills.	☐	☐	☐	☐	☐	☐
In your own words (please write in the open-ended text boxes)						
24. Please describe why you think we asked you to use peer review in this class.						
25. Please describe how you think real scientists use peer review in their work.						
26. What changes would you recommend to improve peer review in this class?						
27. Additional comments:						

Thank you for your participation.

These are images of the online version of the Spring 2007 Survey to demonstrate how it appeared to students. Students in both BIOL 101 and 102 received identical surveys except for the label indicating the class in which the student was enrolled.



Peer Review Student Survey: Biology 101

1. Peer Review Student Survey

Introduction:
Peer review is the process by which other students read your lab report and give you feedback on it before you turn it in for a grade. The purpose of this evaluation is to capture your perceptions of the usefulness of the peer review system used in biology classes this year and to identify areas for improvement. Your honest responses are very important to us. Your responses are anonymous and will be summarized and reported with those of other students.

2. Understanding the purpose of peer review

1. I have a clear understanding about the purpose of peer review in this class.

Strongly Disagree | Disagree | Slightly Disagree | Slightly Agree | Agree | Strongly agree

2. I think peer review is important in the real-world, scientific community.

Strongly Disagree | Disagree | Slightly Disagree | Slightly Agree | Agree | Strongly Agree

3. I understand the purpose of the example ("calibration") papers provided by the instructor on the CPR website.

Strongly Disagree | Disagree | Slightly Disagree | Slightly Agree | Agree | Strongly Agree

4. I understand the purpose of other students providing me with feedback on my lab reports.

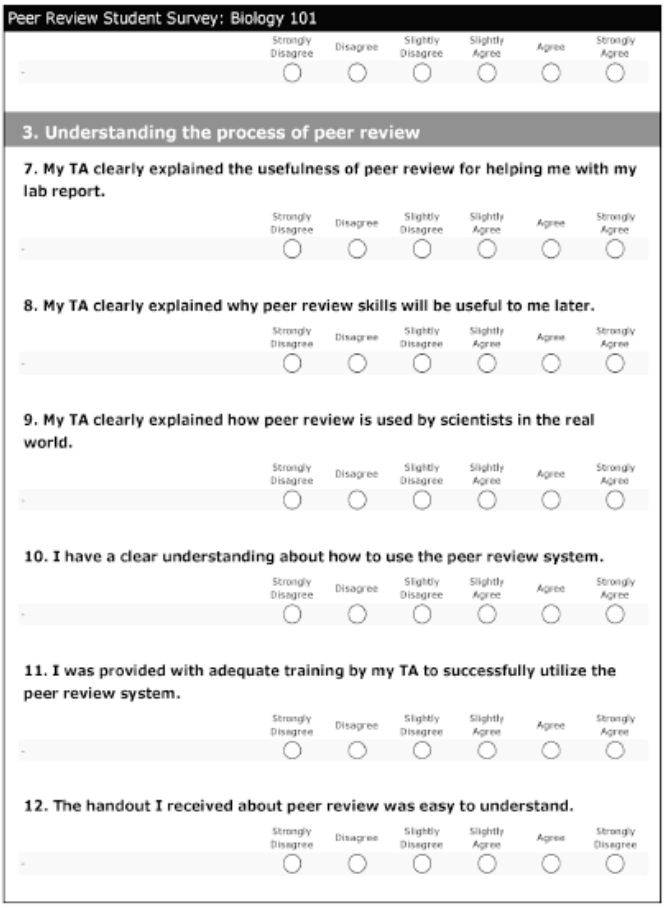
Strongly Disagree | Disagree | Slightly Disagree | Slightly Agree | Agree | Strongly Agree

5. I understand the purpose of providing feedback on my fellow students' lab reports.

Strongly Disagree | Disagree | Slightly Disagree | Slightly Agree | Agree | Strongly Agree

6. I understand the purpose of the self-assessment assignment.

Page 1



Peer Review Student Survey: Biology 101

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

3. Understanding the process of peer review

7. My TA clearly explained the usefulness of peer review for helping me with my lab report.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

8. My TA clearly explained why peer review skills will be useful to me later.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

9. My TA clearly explained how peer review is used by scientists in the real world.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

10. I have a clear understanding about how to use the peer review system.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

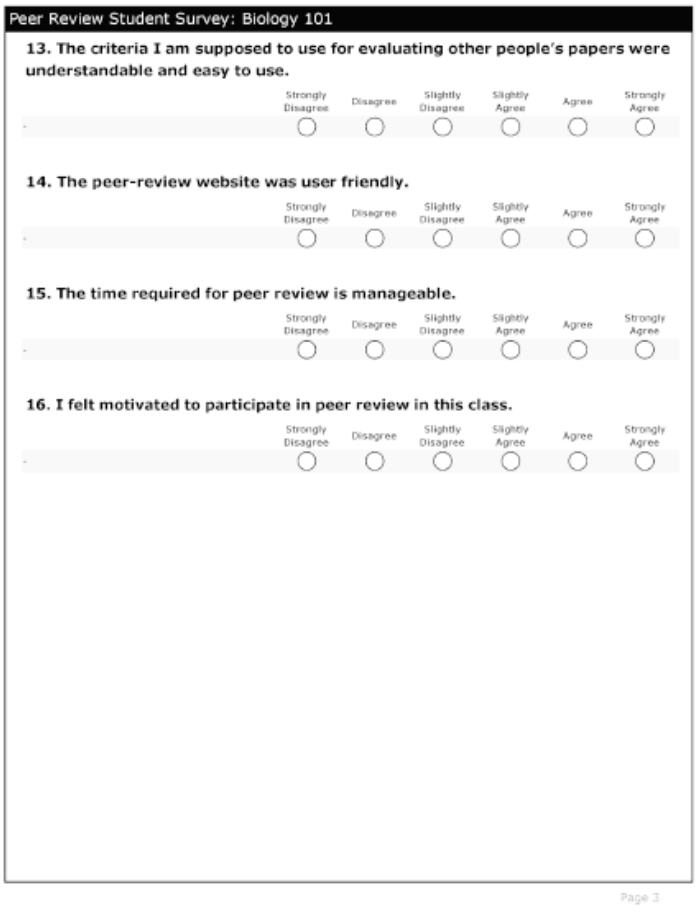
11. I was provided with adequate training by my TA to successfully utilize the peer review system.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

12. The handout I received about peer review was easy to understand.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Disagree
-	○	○	○	○	○	○

Page 2



Peer Review Student Survey: Biology 101

13. The criteria I am supposed to use for evaluating other people's papers were understandable and easy to use.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

14. The peer-review website was user friendly.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

15. The time required for peer review is manageable.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

16. I felt motivated to participate in peer review in this class.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

Page 3

Peer Review Student Survey: Biology 101

4. Impact of the peer review system

17. The example ("calibration") papers provided by the instructor on the CPR website were useful to me.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

18. The self-assessment of my paper was useful to me.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

19. Other students' feedback was helpful to me in revising my lab reports.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

20. I was satisfied with the quality of the feedback I received from other students.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

21. Providing feedback to other students helped me in making revisions to my own lab reports.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

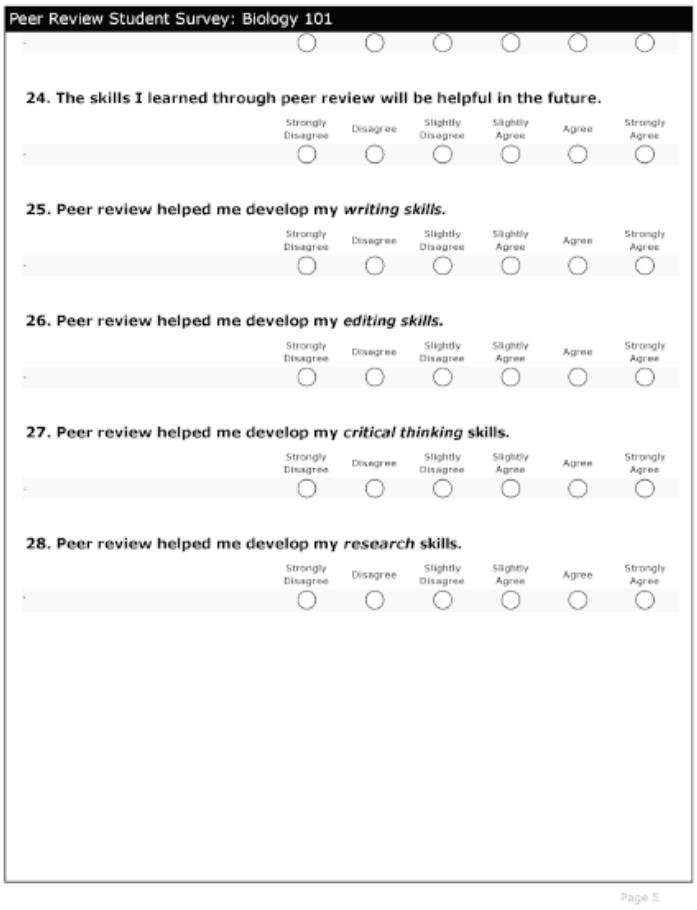
22. I provide high quality feedback I provided to other students.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
-	○	○	○	○	○	○

23. Peer review facilitated my understanding of an in-class experiment.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree

Page 4



Peer Review Student Survey: Biology 101

	○	○	○	○	○	○

24. The skills I learned through peer review will be helpful in the future.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
	○	○	○	○	○	○

25. Peer review helped me develop my *writing skills.*

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
	○	○	○	○	○	○

26. Peer review helped me develop my *editing skills.*

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
	○	○	○	○	○	○

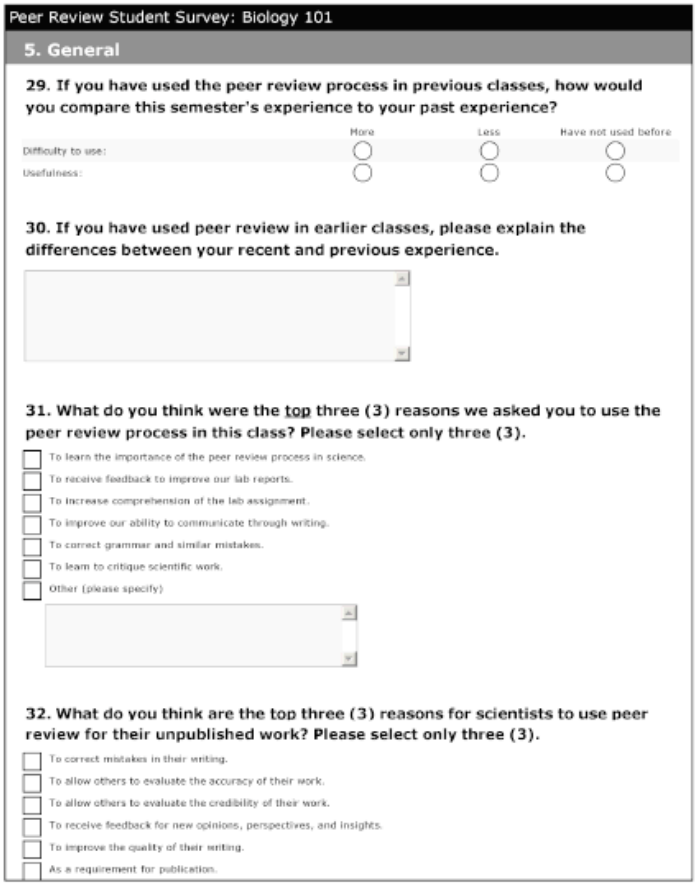
27. Peer review helped me develop my *critical thinking* skills.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
	○	○	○	○	○	○

28. Peer review helped me develop my *research* skills.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
	○	○	○	○	○	○

Page 5



Peer Review Student Survey: Biology 101

5. General

29. If you have used the peer review process in previous classes, how would you compare this semester's experience to your past experience?

	More	Less	Have not used before
Difficulty to use:	○	○	○
Usefulness:	○	○	○

30. If you have used peer review in earlier classes, please explain the differences between your recent and previous experience.

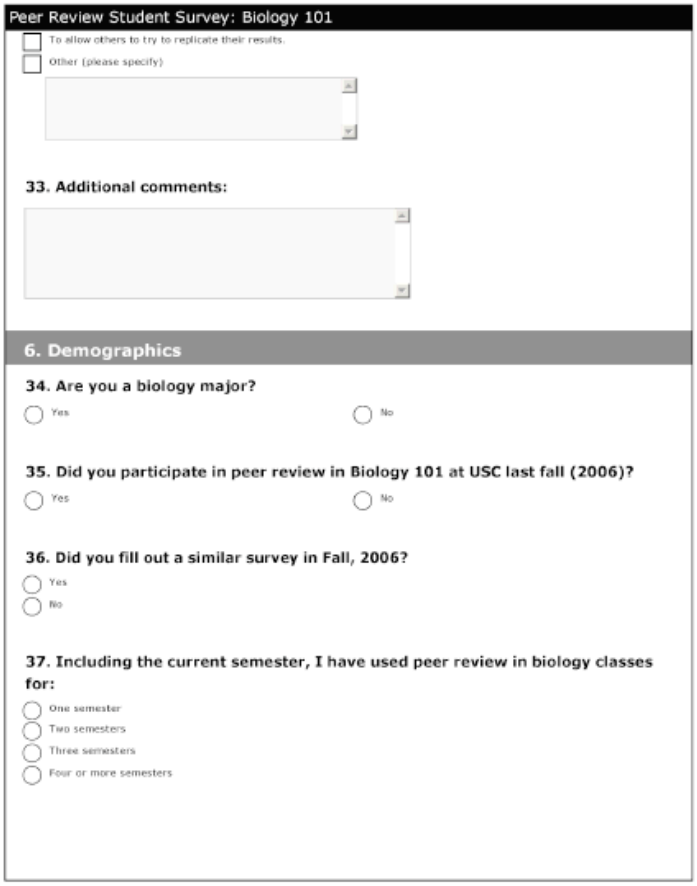
31. What do you think were the top three (3) reasons we asked you to use the peer review process in this class? Please select only three (3).

- ☐ To learn the importance of the peer review process in science.
- ☐ To receive feedback to improve our lab reports.
- ☐ To increase comprehension of the lab assignment.
- ☐ To improve our ability to communicate through writing.
- ☐ To correct grammar and similar mistakes.
- ☐ To learn to critique scientific work.
- ☐ Other (please specify)

32. What do you think are the top three (3) reasons for scientists to use peer review for their unpublished work? Please select only three (3).

- ☐ To correct mistakes in their writing.
- ☐ To allow others to evaluate the accuracy of their work.
- ☐ To allow others to evaluate the credibility of their work.
- ☐ To receive feedback for new opinions, perspectives, and insights.
- ☐ To improve the quality of their writing.
- ☐ As a requirement for publication.

Page 6



Peer Review Student Survey: Biology 101

☐ To allow others to try to replicate their results.

☐ Other (please specify)

33. Additional comments:

6. Demographics

34. Are you a biology major?

◯ Yes ◯ No

35. Did you participate in peer review in Biology 101 at USC last fall (2006)?

◯ Yes ◯ No

36. Did you fill out a similar survey in Fall, 2006?

◯ Yes
◯ No

37. Including the current semester, I have used peer review in biology classes for:

◯ One semester
◯ Two semesters
◯ Three semesters
◯ Four or more semesters

Page 7

Appendix 9. Single rater reliabilities for individual Universal Rubric criteria using trained raters

		Course			Overall
		101	102	301	Ave
Criteria	*n* =	*49*	*45*	*48*	*142*
Introduction					
Context		.40	.62	.25	.42
Accuracy and relevance		.40	.23	.38	.34
Hypotheses					
Testable		.44	.44	.59	.49
Scientific Merit		.52	.39	.41	.44
Methods					
Controls		-[1]	.00	.00	.00[2]
Experimental Design		.08	.73	.31	.37
Results					
Data Selection		.25	.27	.39	.30
Data Presentation		.53	.47	.37	.46
Statistics[2]		.32	.01	.35	.23
Discussion					
Conclusions based on data		.36	.33	.39	.36
Alternative explanations refuted		.47	.29	.46	.41
Limitations		.31	.63	.33	.42
Significance		.30	.58	.56	.48
Primary Literature		.31	.66	.84	.60
Writing Quality		.20	.15	.45	.27
Total Score		.66	.66	.65	.66

Note. Sample sizes reflect the number of unique papers scored per course. [1]The trained raters for BIOL 101 did not perceive the genetics assignment as providing a traditional control and chose as a group to not rate this criterion. [2]See section on low criteria reliabilities for explanation.