EXAMINING THE IMPACT OF CALIBRATED PEER REVIEW (CPR) ON STUDENT WRITING DEVELOPED THROUGH WEB-BASED ECOLOGICAL INQUIRY PROJECTS

A Dissertation

by

DENISE C. ROBLEDO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Rangeland Ecology and Management

Examining the Impact of Calibrated Peer Review (CPR) on Student Writing Developed

Through Web-Based Ecological Inquiry Projects

EXAMINING THE IMPACT OF CALIBRATED PEER REVIEW (CPR) ON STUDENT WRITING DEVELOPED THROUGH WEB-BASED ECOLOGICAL INQUIRY PROJECTS

A Dissertation

by

DENISE C. ROBLEDO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	X. Ben Wu
	Stephanie Knight
Committee Members,	Steven G. Whisenant
	Hersh Waxman
Head of Department,	Steven G. Whisenant

May 2011

Major Subject: Rangeland Ecology and Management

ABSTRACT

Examining the Impact of Calibrated Peer Review (CPR) on Student Writing Developed Through Web-Based Ecological Inquiry Projects. (May 2011)

Denise Celeste Robledo, B.A., Texas A&M University; M.Ed., Texas A&M University

Co-Chairs of Advisory Committee: Dr. X. Ben Wu
Dr. Stephanie Knight

E-learning tools such as Calibrated Peer Review (CPR) have made writing assignments easier to implement and grade; however, we have limited knowledge of how CPR affects student scientific writing. Past CPR research has examined how CPR generated scores change across multiple CPR writing assignments for the purpose of reporting student learning gains. This study will not rely on CPR generated score data. This study (1) independently evaluated the impact of CPR on student writing of ecological inquiry report components using a grading criteria instrument and (2) explored how the revision process influenced the quality of ecological inquiry report components through text analysis.

A web-based science inquiry project was implemented in a large (up to 500 students) introductory ecology course. Students observed grizzly bears at McNeil River Falls in Alaska using Bear Cam picture stills. They developed and tested hypotheses about grizzly bear spatial distribution and interactions and reported findings in individual ecological inquiry reports. Students submitted reports to CPR and anonymously reviewed three peer reports and self-assessed their own. Finally, students were given one-week following CPR to revise reports based on peer reviews and submit online.

A 28-item grading criteria instrument (9 scales) was used to examine how students revised ecological inquiry reports post CPR. Eight paired t-tests were used to assess the pre-post CPR changes in scores for individual grading criteria scales or components. Cohen's d effect size was used to explore how achievement or performance level, ethnicity, gender and major influenced student text changes to ecological inquiry report components post CPR. Text analysis using a subset of 27 sample reports (pre-post CPR) assessed the amount and location of text changes and the impact of these revisions on the quality of ecological inquiry report components. Common errors in ecological inquiry report components post CPR were also analyzed.

Results showed that CPR and revision significantly improved the scores related to the objective, sampling and discussion scales. Analyses using Cohen's d effect sizes illustrated interesting but inconsistent patterns related to the influence of student performance level, gender, ethnicity, and major on pre-post CPR score gains. Text analysis revealed the majority of helpful revisions were related to making the objective identifiable, reporting of sample size and discussion of study limitations and future questions raised by individual ecological inquiry projects. Text analysis shows three common reasons participants failed to meet grading criteria post CPR. Un-testable hypotheses, insufficient descriptions for sample selection, data analysis, variables collected and revisions of only easy grading criteria components. This study provided direct evidence of CPR's effects on student writing and provided a greater understanding of pattern of revision process following CPR.

DEDICATION

To my parents, David Robledo and Hortencia Olivarez and my godparents Berta and Reyes Trevino for your unwavering support and faith in my ability to attain my Ph.D. To my sister, Candace Alma and brother, David Angelo, who have always motivated and inspired me to work harder. To the Robledo-Alfaro and Olivarez-Trevino family for always tending to my roots and keeping my feet firmly grounded. To my grandparents, Jose and Consuelo Robledo and Cruz and Balbina Olivarez for instilling a legacy in their children to pursue education and be leaders in their community.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Dr. X. Ben Wu, my mentor who was pivotal in my success as a Ph.D. student. It has sincerely been a privilege to work with you and your colleagues. I would also like to acknowledge my co-chairs, Drs. X. Ben Wu and Stephanie Knight whose constant guidance and advice elevated my dissertation research. Thank you for always illuminating the direction and focus of my study. Additionally, I would also like to extend my gratitude to my committee members, Dr. Steve Whisenant and Hersh Waxman. Dr. Whisenant, I would like to recognize you for hiring me all those years ago and setting me down this academic journey. I truly appreciate all of the support you have provided me over the years as department head.

I would like to recognize Dr. Manuel Pina and the Hispanic Leaders in Agriculture and the Environment (HLAE) program for funding my Ph.D. studies at Texas A&M University. Dr. Pina, I will eternally be grateful for all the opportunities you provided me while in graduate school. I would also like to show my gratitude to Dr. Roel Lopez, whose funding and assistance in my last year has been essential to my Ph.D. completion. Dr. Lopez, thank you for all your kindness and always encouraging me to get the dissertation done!

Very special thanks to those who assisted in various tasks associated with my dissertation research over the years: Christopher Cheleuitte, Xi Chen, Sungae Yoo, Cheryl Ann Peterson. Thanks also go to my ESSM department friends, faculty and staff for making my time in the Department of Ecosystem Science and Management a great experience. I would like to acknowledge my tribe, Linda Langlitz, Maria Gutierrez, Elke Aguilar, Camy Sturdivant and JoAnna Thorton whose friendship always provided a constant source of humor, inspiration and energy to pursue my Ph.D. studies.

Finally, I would like to recognize those who put up with me when I was struggling with this dissertation and loved me when I was at my worst. Xochitl Flores,

thanks for always talking to me about the future. Casey Ball, what can I say, thanks for never letting me feel too intimidated by academia and making me feel like I could make my own and noteworthy contribution. Matthew Sifuentes, thank you, if you didn't keep me laughing I would have been crying. You never allowed me to lose sight of what was most important. Gaby Sosa, thank you for being a constant reminder of my community and its loving spirit. I am eternally grateful for your constant encouragement to blaze toward my goals (my tribute to your fascination with fire – for restoration purposes only of course). Christopher Cheleuitte, thanks for not letting me lose myself while I was suffering from a vulcan mind-meld. You reflect beams of light that cast shadows on all fear, doubt and worry. Thanks again for all your assistance with inter-rater reliability training – your help was fundamental to me completing my dissertation. Candace Robledo, you're the greatest best friend and sister a girl could have. I am so blessed to have been born your twin. You are the stars in my night sky – always a reminder to remain optimistic for tomorrow brings plenty of endless opportunities. Thanks for all of your counseling and advice through this dissertation. This would have been impossible to complete without you by my side. Dr. Meera Alagaraja, without your friendship I would have literally faded into gray. You and Candace were the guardians of my hope and self-confidence. Thanks for all the laughs, long conversations and pep-talks. The best for last – Mom, Tia, Dad, Tio and David – I can't even begin to thank you for all you have given me. Your love, hard work and sacrifices have allowed me to live the life of my dreams and pursue this Ph.D. Thank you for always supporting me. You always gave me perspective and kept me motivated. At times you believed more in my abilities than I did and that was reason enough for me to finish this dissertation. Mom and Tia thank you for all your prayers and encouragement. Dad – I got in, got it done and now I'm on my way up! Tio – Por fin, tal vez puedo realizar una vida normal! Gracias por todo tu amor y apollo durante mis estudios. Dave – Seriously, I wouldn't work as hard as I do if I didn't have you in my life supporting me every step of the way. Thanks hermanito.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

1. INTRODUCTION

Calibrated Peer Review (CPR) is a web-based system developed by UCLA researchers from the Division of Molecular Sciences in 1995. The CPR system is based on the scientific inquiry model and provides students an opportunity to practice scientific manuscript submission and anonymous peer review. CPR was developed to promote student writing and reviewing skills. CPR provides instructors with tools to develop and manage student writing projects and provides students with tools to self-assess and anonymously peer review writing projects assigned by instructors.

A web-based science inquiry project was implemented in a large (500 students) introductory ecology course in 2007, 2008 and 2009. Students observed grizzly bears at McNeil River Falls in Alaska using Bear Cam picture stills. They developed and tested hypotheses about grizzly bear spatial distribution and interactions and reported research and findings in individual ecological inquiry reports. Students submitted reports to CPR and anonymously reviewed three peer reports and self-assessed their own. Finally, students were given one-week following CPR to revise ecological inquiry reports based on peer reviews and re-submit online for further grading by the Teaching Assistant.

The reporting of changes in CPR generated scores across multiple CPR assignments has yielded evidence that CPR has a positive effect on student writing and reviewing ability (Gerdeman et al. 2007; Margerum et al. 2007; Gunersel et al. 2008; Gunersel and Simpson 2009). Assessment of student writing and reviewing ability using CPR generated scores, however, is dependent on the quality of CPR grading criteria instruments, the nature of CPR writing assignments, and instructor-set CPR scoring regime. Instead of relying on CPR generated score data, this study (1) evaluated the impact of CPR on student writing through independent evaluation of student report using a grading rubric and (2) explored the pattern of the revision process through text

This dissertation follows the style of CBE-Life Sciences Education.

analysis and its relationship to improvement in quality of writing.

The objective for section two was to evaluate how CPR and the revision process affected the quality of student writing with respect to specific components of their ecological inquiry reports. In addition, this section explored how achievement level, gender, ethnicity, and major influenced the impact of CPR on student writing.

Section three explored the revision process associated with CPR in students' development of ecological inquiry reports by characterizing the types, frequency and impact of revisions generated by students post CPR. Moreover, common errors present in post CPR reports that prevented students from meeting grading criteria were also identified and analyzed.

This study attempted to identify future research needs to better our understanding of CPR impact on student scientific writing skills. Study findings can contribute to the understanding of the revision process following web-based, anonymous peer review. Improved understanding the relationship between online anonymous peer review and the revision process can contribute to the effective development of future CPR assignments and more effective instructional approaches to promote student scientific writing and learning.

2. THE INFLUENCE OF CALIBRATED PEER REVIEW (CPR) ON THE QUALITY OF STUDENT ECOLOGICAL INQUIRY REPORTS

Introduction

Undergraduate science instructors face a challenge. Recent educational reform challenges science instructors to implement scientific inquiry and writing projects in their courses to promote critical thinking and writing skills (Reynolds and Moskovitz, 2008). The time required to plan, design, implement and assess scientific inquiry and writing projects can easily overwhelm even the most experienced instructors. High student course enrollments can further increase time required by the instructor to carry out scientific inquiry and writing projects.

Increasingly, instructors are adopting e-learning tools and technology that are capable of helping instructors manage and author scientific inquiry and writing assignments. One such learning tool is Calibrated Peer Review (CPR). UCLA researchers from the Division of Molecular Sciences developed Calibrated Peer Review (CPR) in 1995. The CPR system was designed based on a model of scientific inquiry and functions to replicate scientific manuscript submission and anonymous peer review (Carlson and Berry, 2003, 2008). The purpose of CPR is to develop student writing and anonymous peer review skills in order to promote critical thinking and mastery of science content (Kennicutt et al. 2008). CPR is a web-based system that serves two main functions. First, CPR provides instructors with tools to develop and manage student writing projects. Second, CPR provides students with tools to self-assess and anonymous peer review writing projects assigned by instructors. CPR is made up of four structured workspaces (Carlson and Berry, 2003) and students are required to complete them in a sequential manner.

- Text entry workspace: Provides students with a writing prompt, instructor provided grading criteria and source material. Upon completion of their writing task, students submit their writing to the text entry workspace.

- CPR calibration workspace: Students grade three benchmark writing samples (low, average and high quality) of their current writing assignment. The three benchmark writing samples are graded by students using the instructor provided grading criteria provided in the text-entry workspace. Student grading is compared to instructor grading of the three benchmark writing samples. Ultimately, students are given a reviewers competency index (RCI) score (1 to 6) representing how well their grading matched the instructors. If students fail to receive an acceptable RCI score (previously set by the instructor) they will have to repeat CPR calibration.

- CPR peer review workspace: Students are randomly assigned three essays written by their classmates to anonymously peer review. During CPR peer review students will grade randomly assigned essays using the instructor provided grading criteria they used in the text-entry and calibration workspaces.

- CPR self-assessment workspace: During CPR self-assessment stage students are required to grade their own writing essay originally submitted in the text-entry workspace at the start of CPR. Again, students will grade their own writing using the instructor provided grading criteria used in the three previous CPR workspaces.

A CPR assignment library is available to instructors who are not interested in authoring their own CPR assignments. Instructors who choose to author and implement their own CPR assignment can expect to complete the following tasks.

- Create CPR assignment goals, instructions and timeline.

- Develop CPR writing assignment criteria (writing prompt and length requirements).

- Generate grading criteria to be used for all CPR workspaces (text entry, calibration, peer review and self-assessment).

- Produce and grade writing project samples of low, average and high quality to serve as benchmarks for calibration. Instructor grades for writing samples will be compared to student generated grades during CPR calibration.

- Organize and upload supplemental resource material for CPR writing assignment.

- Set-up CPR student scoring. For example, instructors set acceptable percent of style and content calibration questions students must answer correctly in order to pass CPR calibration (See T able 2.1 in Appendix A).

- Review student generated grades in CPR and identify and re-score student peer reviews that were graded by students who failed the CPR calibration process.

Prior to participating in CPR students are required to complete a CPR tour and training module. After training students are allowed to enter the CPR system. Students must complete all CPR tasks according to the instructor set timeline. This will ensure all students participating in CPR have submitted their writing assignments to the CPR text entry workspace before any student can proceed to the calibration, peer review and self-assessment workspaces. After students complete the CPR self-assessment workspace (the last stage in CPR) they are allowed to review results of peer review and their overall writing assignment grade.

As an educational tool, CPR is often met by student reluctance to accept the technology. This reluctance stems from the belief that peers are unable to grade fairly; however, CPR research studies have shown that students' distrust of their peer's grading ability declines significantly with repeated CPR use (Robinson, 2001; Prichard, 2005; Kennicutt et al., 2008). Further promoting the use of CPR as an educational tool are recent research findings that repeated use of CPR has led to improved student writing and reviewing skills (McCarty et al., 2005; Gunersel et al., 2008; Gunersel and Simpson, 2009; Gragson and Hagen, 2010).

CPR use and research has been reported in a variety of science disciplines such as chemistry, biology, zoology, physiology and neuroscience. However, very little research reports how student use of CPR impacts student writing, specifically scientific writing. The absence of a CPR revision workspace is a hindrance to evaluating CPR impacts on student scientific writing (Prichard, 2005). Once students complete a CPR assignment, they have no incentive to make revisions based on peer review feedback unless the instructor assigns multiple CPR assignments. Further complicating the task of assessing CPR's effect on student scientific writing is the discipline-specific context and structure of CPR writing assignments (Gunersel et al., 2008). CPR writing assignments vary in difficulty, design and their ability to promote higher order and critical thinking skills (Reynolds and Moskovitz, 2008).

CPR generated scores are used to help instructors' track students' writing and peer reviewing performance within the CPR system (See Table A.1). When reporting how the use of CPR affects student writing and peer review skills researchers have relied on examining changes in CPR generated scores. Instructors assign multiple CPR assignments and evaluate how CPR generated scores changed from the first to the last CPR assignment. Often changes in CPR generated scores are used in research to report changes in students' writing and reviewing abilities (Gerdeman et al., 2007; Margerum et al., 2007; Gunersel et al., 2008; Gunersel and Simpson, 2009).

CPR generated scores commonly used to report changes in students' writing and reviewing ability as a result of CPR use include the text rating (TR) and reviewer competency index (RCI) scores. The TR score (1 to 10 points) is a weighted average of three peer reviews given to a CPR writing assignment. Changes in TR scores for subsequent CPR assignments are often cited as evidence of changes in a student's writing ability. The RCI score (1 to 6 points) represents how well students applied the grading criteria during CPR calibration to peer review low, average and high quality benchmark writing samples. If student grading of benchmark writing samples deviates too much from the instructor's grading the student would receive a low RCI score. Therefore a peer review by a student with a low RCI score is given less weight than a student peer review that has a high RCI score. Recent CPR research uses the RCI score as evidence of a students' reviewing ability.

The reporting of changes in CPR generated scores across multiple CPR assignments has yielded evidence that CPR has a positive effect on student writing and reviewing ability. One study substantiates this finding by reporting a gradual increase in student RCI scores and low standard deviations between students' self assessment scores and peer review scores for students completing multiple CPR writing assignments that increased in level of difficulty (Margerum et al., 2007). Furthermore, when comparing student CPR generated scores between a low and high scoring difficulty assignment, researchers reported that students' ability to sustain relatively the same CPR scores between assignments substantiated learning had occurred (Carlson and Berry, 2008). Contrary to the findings above that CPR does improve student writing and reviewing ability, Walvoord et al. (2008), found after repeated CPR use students did not improve technical writing (grammar, style) nor their ability to convey scientific knowledge (hypotheses, methods, data collected, interpretation of results and support of hypothesis) after analyzing grades provided by three independent experts. An explanation provided by Walvoord et al. (2008) for the difference in findings compared to other CPR studies was that the rubric used may have encouraged students to write simplified essays.

CPR generated scores have also been used to group students by achievement level for the purposes of characterizing low, average and high achieving student performance in CPR. Gunersel and Simpson (2009) implemented multiple CPR assignments and found CPR did not favor students who entered the CPR system better prepared (performed well in calibration phase for first CPR assignment). That is, when students were grouped into low, moderate and high performers based on the TR and RCI of the first CPR assignment, low performers exhibited the most gains over time in (TR) and self review competency (RCI) scores (Gerdeman et al., 2007; Margerum et al., 2007; Gunersel et al., 2008; Gunersel and Simpson, 2009).

Using changes in CPR generated scores to report improvement in student writing and reviewing skills has several limitations. For example, weighting of CPR scores is left up to the instructor's discretion and therefore may lack consistency among research studies (Walvoord et al., 2008). Aside from inconsistent weight designation by instructors on CPR generated scores, other problematic issues arise when students fail to perform well in calibration or complete CPR. CPR generated scores may be inadequate for research when a students' submitted text does not receive three peer reviews, graded by at least two reviewers who exceeded the allowed deviations set by the instructor, or reviewed by a student with a low RCI score. Lastly, ambiguous grading criteria questions that are used throughout all CPR stages can influence the quality of CPR generated scores. Therefore, sole reliance on CPR generated scores for the purposes of evaluating CPR effects on student writing and reviewing skills should be minimized.

This study will not rely on CPR generated score data to evaluate student learning or to group students by achievement level. The study objective was to evaluate how CPR affects student writing of ecological inquiry reports, and will address the following two research questions.

1. How does student use of CPR impact the quality of ecological inquiry reports?
2. Do achievement level, gender, ethnicity, and major influence the impact of CPR on the quality of ecological inquiry reports?

Examination of these two research questions can contribute to the understanding of the relationship between online anonymous student peer review and the revision process. Improved understanding of this relationship can contribute to the development of more effective instructional strategies to enhance student writing in science.

Methods

Context of study

This study was conducted at a Texas University in a Fundamentals of Ecology course. All research activities were certified by the Institutional Review Board in 2007 and renewed in 2008 and 2009 (IRB Protocol Id: 2007-0534). Co-taught by two professors, Fundamentals of Ecology enrolled up to 500 students. Students were required to complete weekly online quizzes, 4 major exams, and a web-based ecological inquiry project that involved the use of CPR (4 weeks).

The web-based ecological inquiry project was designed specifically to promote student understanding of how ecologists conduct and report research. The project required individual students to complete all the following tasks in order to receive full credit for the ecological inquiry project.

1. Observe Alaska grizzly bear behavior and spatial patterns at McNeil River Falls in Alaska using picture stills captured by a remote Bear Cam (Griffing, 2007).
2. Develop a study objective and testable hypothesis within the limits of the data provided (Alaska Grizzly Bear Cam picture stills), to explain observed grizzly bear behavior and spatial patterns.
3. Design a study to test the hypothesis developed.
4. Collect and analyze relevant data using the Alaska Grizzly Bear Cam pictures.
5. Interpretation of the results.
6. Participate in online discussion (10 students per discussion group) to exchange ideas and feedback on research objectives, hypotheses, study design and data analyses.
7. Write an individual one-page ecological inquiry report with an Introduction (Objective/Hypothesis), Methods (Data Collection/Analysis), Results and Discussion/Conclusion section. Directions for inquiry projects and writing assignment are provided in Appendix B.
8. Submit individual ecological inquiry reports to CPR in order to participate in online, anonymous peer review.
9. Revise one-page ecological inquiry report based on CPR feedback.
10. Re-submit ecological inquiry reports online to WebCT Vista (Blackboard Vista) a week after CPR completion for teaching assistant grading.
11. Complete online survey to provide feedback on usefulness of CPR. In 2008 and 2009, students also were asked to rate the usefulness of their discussion group member's participation.

The grading criteria for the web-based ecological inquiry project was revised from 2007 to 2009 (See Table A.2). Throughout the three years the Alaska Grizzly Bear inquiry project was implemented the CPR workspace scoring criteria and the CPR calibration score criteria set by the instructor remained the same (See Table A.3 and A.4).

CPR grading criteria were developed to evaluate student ecological inquiry reports that resulted from the completion of the web-based Alaska Grizzly Bear inquiry project. CPR grading criteria served three purposes:

1. Outlined requirements for individual students’ one-page ecological inquiry reports (pre-post CPR);
2. Provided grading criteria for student use in CPR stages (calibration, peer review and self-assessment); and
3. Assisted teaching assistant and researcher grading of individual student pre and post CPR reports.

CPR grading criteria evaluate 9 separate tenets or scales (with 28 dichotomous questions; 3 for each scale) of scientific writing (See Table A.5). Each grading criteria question evaluates the presence or absence of a specific element of scientific writing. The grading criteria in Table A.5 were used by students in CPR for the 2008 and 2009 Fall semesters. In 2007, the grading criteria used by students in CPR consisted of only 10 rubric questions (see Table A.6 in Appendix A). The rubric changed in 2008 in response to student feedback that grading criteria was difficult to use.

Study participants

Study participants included sophomore students registered for Fundamentals of Ecology in 2007, 2008 and 2009 Fall semesters. Study participants included only sophomore students who had attended the university for at least 2 semesters. Sophomore students who transferred from another university or college were eliminated from the study sample because there was no GPR (grade point average) on record for these students at the time of course enrollment. Also, sophomore students were eliminated from the study sample if they failed to complete both the pre and post CPR assignments. Respective sample sizes for study samples in 2007, 2008 and 2009 were 89, 74 and 66. Table A.7 in Appendix A summarizes the demographic characteristics of study samples by year.

Inter-rater reliability

Prior to blindly grading student pre and post CPR reports with the grading criteria (9 scales), the researcher and a PhD candidate majoring in Ecosystem Science and Management conducted inter-rater reliability assessment. First, inter-rater reliability judges (researcher and PhD candidate) completed training on how to apply the grading criteria. During training the judges collaboratively graded 12 sample scientific reports using the rubric. Sample scientific reports were equally selected among low, average and high quality reports previously graded by the Fundamentals of Ecology teaching assistant in 2008. During rubric training judges developed detailed criteria for each of the 28 questions to later assist the judges with pre and post CPR report grading. See Appendix C for individual question grading criteria developed during inter-rater judge training. Once rubric training was completed inter-rater reliability judges independently graded 12 reports. The 12 reports for inter-rater reliability were randomly selected among low, high and average quality reports previously graded by the Fundamentals of Ecology teaching assistant in 2008.

Cohen's Kappa (κ) was used to assess inter-rater reliability and corrects for nominal-scale chance agreement between raters (Cohen, 1960). Kappa (κ) ranges from -1 (perfect disagreement) to 1 (perfect agreement beyond chance agreement). A zero Kappa (κ) score means inter-rater reliability judges failed to agree anymore than would be expected by chance. Negative Kappa (κ) values indicate judges agreed less than what would be expected by chance. Landis and Koch (1977) suggest kappa values from 0.41 to 0.60 are considered moderate while values above 0.61 to 0.80 are considered substantial. For this study a Kappa Score ≥ 0.500 was considered an acceptable level of agreement between raters considering all disagreements carried equal weight. Kappa Scores for inter-rater reliability between judges for each of the 9 tenets evaluated by the 28 rubric questions are provided in Table A.8 (Appendix A).

Data sources

After inter-rater reliability was completed all pre and post CPR reports for qualifying sophomore student participants in the 2007, 2008 and 2009 Fall semesters were given a unique random identifier. All pre and post CPR reports were ordered by the unique random identifier and blindly graded by the researcher using the grading criteria (See Table A.5). The researcher graded a total of 458 pre and post CPR reports (2007, 2008 and 2009 samples). The pre and post grades generated for individual grading criteria scales and individual questions for each pre and post student scientific report served as the primary data source for this study. The following items from the hypothesis, content placement and writing grading criteria scales were scored but not intended to be included in later analyses. These questions were deleted to make ecological inquiry component scales (3 points each) comparable.

- Hypothesis Question 7: Is part of the hypothesis testable with available data?
- Content Placement Question 23: Is the report organized in three sections (Introduction & Hypothesis, Methods, Results & Discussion)?
- Writing Question 27 and 28 (Writing Scale for the Rubric).

Individual student data related to GPR at time of class enrollment, ethnicity, gender and reported major were requested from the university. Student GPR at time of class enrollment was used to group students into low, average and high performing groups. Also, due to a limited representation of students in some ethnicity and majors, ethnicity data was grouped by Hispanic and White sub categories, and majors were categorized by college and grouped into majors and non-majors categories.

Statistical analysis

All statistical analyses were conducted with SPSS. The Shapiro-Wilks test with an alpha of 0.05 was used to test the normality of the data generated by the grading criteria at the scale level. Our interest was focused on comparing the difference between the pre-CPR and post-CPR reports in eight specific aspects (scales) in order to examine the effect of CPR on the quality of student writing. Given the fact that we intentionally chose participants representing low, average and high performance levels, paired comparisons would be more appropriate than non-paired mean comparisons such as F-tests using ANOVA. Based on these considerations, the paired t-test was used to test the significance of pre-CPR or post-CPR changes in scores for each of the eight scales. The significance level of alpha=0.05 was used for these t-tests. Participants were grouped into low (GPR ≤ 2.0), average (2 < GPR ≥ 3) and high (GPR > 3) performing groups based on the 25th, 50th and 75th percentiles.

The Cohen's d effect size was used to compare pre-post CPR mean scores of low, average and high performers for grading criteria scales in order to provide data characterizing how students who differ by achievement level use CPR. Additionally, Cohen's d was used to examine the influence of participant gender, ethnicity and major/non-major status on pre-post CPR mean scores for grading criteria scales. Cohen's d analyses of achievement level, gender, ethnicity and major/non-major characteristics were analyzed by 2007, 2008 and 2009 sample years. Cohen's (1988) benchmarks for effect sizes were used to report low, moderate and large effect sizes. According to Cohen's benchmarks an effect size of 0.2 was considered low, 0.5 represented a moderate effect and 0.8 represented a large effect. Lastly, the frequency of students that did not meet requirements, met requirements to a degree and met requirement for grading criteria scales and individual items was reported.

Results

Paired T-test results

The Shapiro-Wilks test of normality showed that score data for grading criteria scales in 2007, 2008 and 2009 were all normally distributed. Paired t-test results for 2007 show significant differences between pre- and post-CPR scores for the sampling (p=0.033) and discussion (p=0.002) scales, but no other grading criteria scale in 2007 had significant changes. For 2008 and 2009, there were significant differences between the pre- and post-CPR scores for the objective (p=0.015 and 0.010), sampling (p=0.002 and 0.038), and discussion (p=0.000 and 0.006) scales, and no significant differences for any other scales (Refer to Tables A.9 – A.11 in Appendix A).

These results reveal a consistent pattern of significant changes for the objective, sampling and discussion scales in 2007, 2008 and 2009, except for objective scale in 2007. Therefore, further analysis based on descriptive statistics and effect sizes were conducted for these three scales to explore the effect of performance level, gender, ethnicity, and major on the changes in student writing quality.

Effect of performance level

Figure 2.1 shows all 2008 and 2009 low, average and high performers met the objective scale to a degree for pre CPR reports. In addition, low performers scores are consistently lower than average and high participant scores. With the exception of 2009 high performers, the majority of 2007, 2008 and 2009 performers on pre CPR reports met the sampling and discussion scales to a degree. Figure 2.2 summarizes Cohen's d effect sizes for CPR pre-post scores for low, average and high performers.

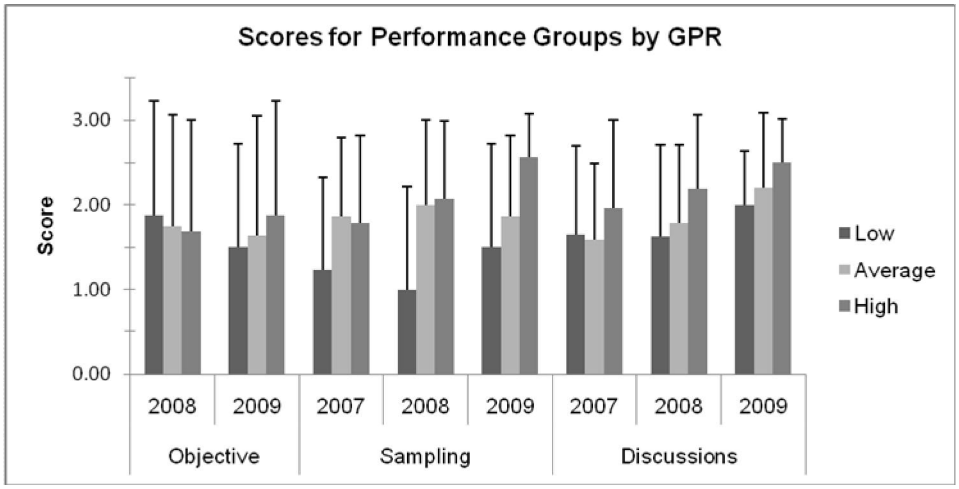

Figure 2.1. Overall pre-CPR performance of low, average and high performers for the objective, sampling and discussion scales.

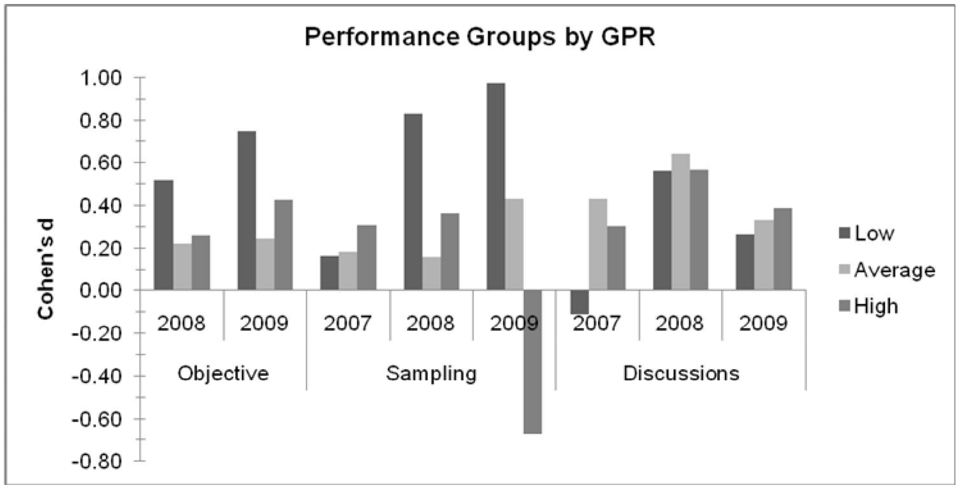

Figure 2.2. Cohen's d pre-post comparisons of low, average and high performers for objective, sampling and discussion scale.

Cohen’s d analysis indicate 2008 low performers show a moderate improvement to objective scale scores post CPR (d=0.518), compared to 2008 average and high performers that show only a small improvement (d=0.220, 0.261 respectively). 2009 low performers made high gains post CPR for the objective scale (d=0.750); however, 2009 average and high groups made small improvements (d=0.244, 0.429 respectively).

Cohen’s d analysis show 2007 low, average and high performers made small improvements to sampling scale scores post CPR (d=0.164, 0.183 and 0.307 respectively). Results indicate 2008 low performers made high gains to sampling scale scores post CPR (d=0.829) and 2008 average and high performers show only small gains (d=0.158, 0.366). 2009 low performers made high improvements to sampling scale scores (d=0.971), 2009 average performers made moderate gains (d=0.430) and 2009 performers show a moderate decline in sampling scores post CPR (d=-0.671).

Cohen’s d analysis indicate 2007 low performers show no improvements for the discussion scale pre-post CPR (d=-0.111); however, 2007 average and high performers show small gains in discussion scale post CPR scores (d=0.431 and 0.302 respectively). 2008 low, average and high performers show moderate improvements in discussion scale scores post CPR (d=0.560, 0.640 and 0.565). Results reveal 2009 low, average and high performers show a small effect for the discussion scale post CPR reports (d=0.263, 0.331 and 0.389 respectively). Refer to Table A.12 for Cohen’s d analysis for low, average and high performers.

Gender: Grading criteria scale performance

Figure 2.3 shows 2009 performers had the highest scores for the sampling and discussion scales pre CPR. The majority of female and male participants per year met the objective, sampling and discussion scale requirements to a degree. Figure 2.4 summarizes Cohen’s d effect sizes for CPR pre-post scores for female and male participants.

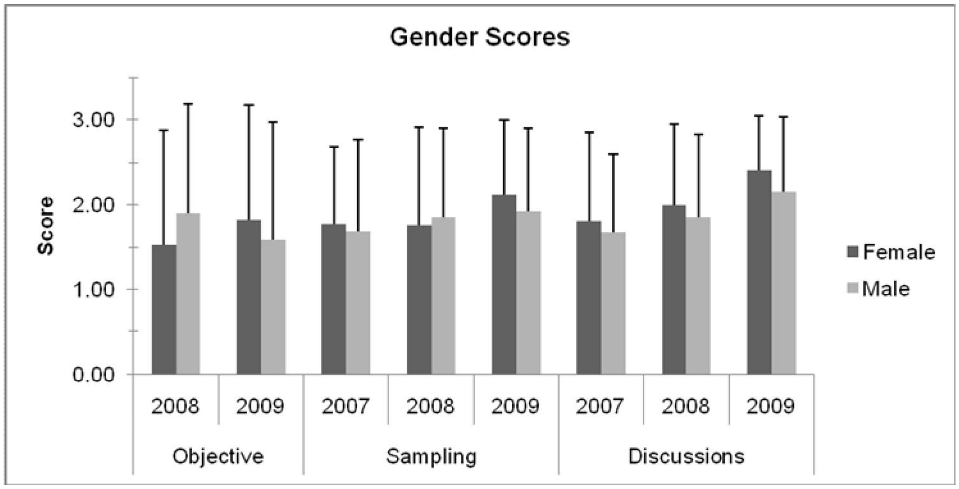

Figure 2.3. Overall pre-CPR performance of low, average and high performers for the objective, sampling and discussion scales.

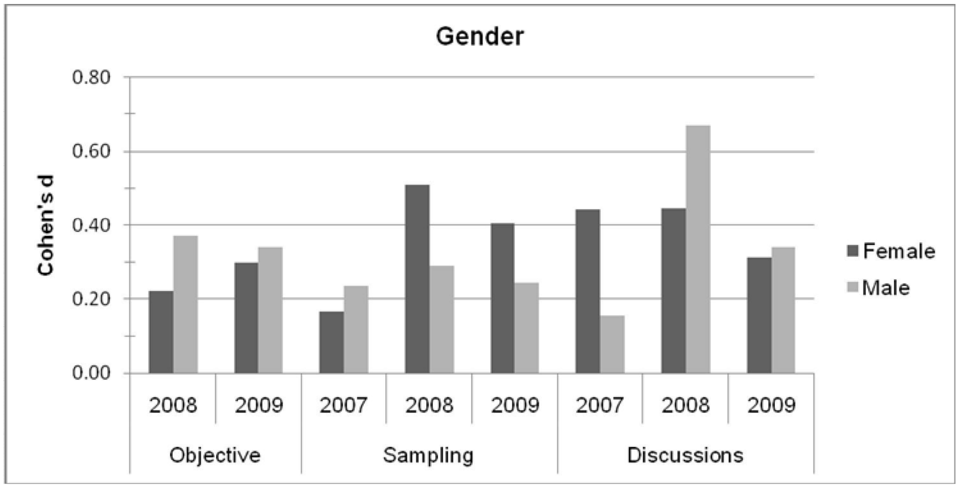

Figure 2.4. Cohen's d pre-post comparisons of female and male participants for objective, sampling and discussion scale.

Cohen’s d analysis indicate 2008 and 2009 female and male participants showed a small gain in pre-post CPR reports for the objective scale (d=0.222 and 0.370 respectively).

Cohen’s d analysis show 2007 female and male participants showed a small improvement from pre-post CPR reports for the sampling scale (d=0.168 and 0.235 respectively). 2008 female participants moderately improved post CPR scores for the sampling scale (d=0.508) compared to 2008 male participants who only showed a small improvement (d=0.292). Results also reveal 2009 female and male participants showed a small improvement in sampling scale scores post CPR (d=0.403 and 0.244 respectively).

Cohen’s d analysis show 2007 female and male participants made small improvements in discussion scale scores post CPR (d=0.442 and 0.155 respectively). Results also indicate 2008 female participants made a small improvement to discussion scale scores post CPR (d=0.445) and 2008 males moderately increased scores (d=0.668). 2009 female and male participants showed a small improvement for the discussion scale post CPR (d=0.312 and 0.339 respectively). Refer to Table A.13 for Cohen’s d analysis of female and male participants.

Ethnicity: Grading criteria scale performance

Figure 2.5 depicts the majority of Hispanic and White participants per year met the objective, sampling and discussion scale requirements to a degree for pre CPR reports. 2009 Hispanics slightly outperformed White participants for the objective and sampling scales pre CPR. Figure 2.6 summarizes Cohen’s d effect sizes for CPR pre-post scores of Hispanic and White participants.

.

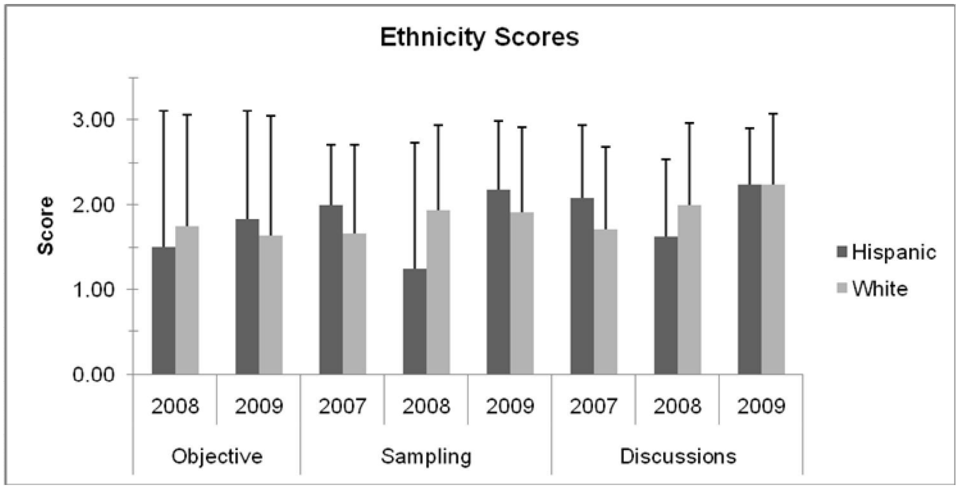

Figure 2.5. Overall pre-CPR performance of Hispanic and White participants for the objective, sampling and discussion scales.

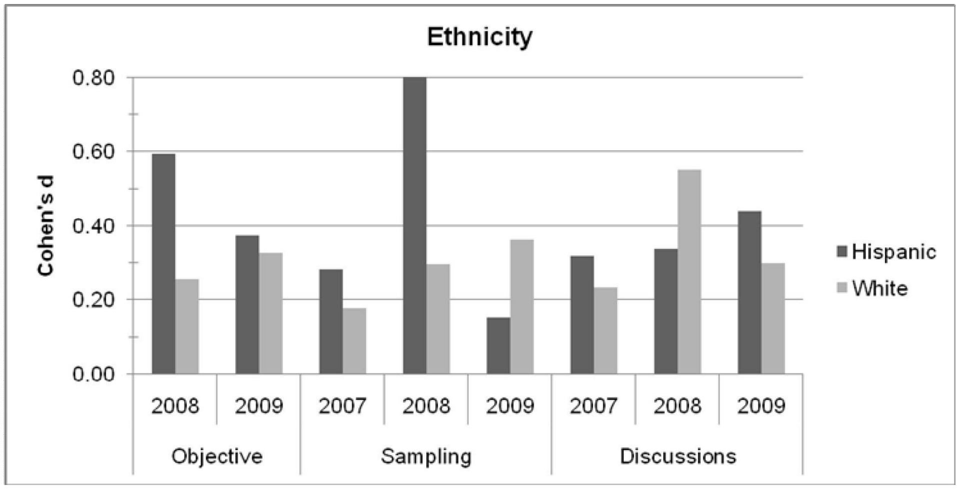

Figure 2.6. Cohen's d pre-post comparisons of Hispanic and White participants for objective, sampling and discussion scale.

Cohen's d analysis show 2008 Hispanic participants made a moderate improvement to the objective scale post CPR (d=0.594) compared to 2008 White participants who show only a small improvement (d=0.254). Results indicate 2009 Hispanic and White participants showed only small improvements to the objective scale post CPR (d=0.373 and 0.327 respectively).

Cohen's d analysis show 2007 Hispanic and White participants show only small improvements to the sampling scale scores post CPR (d=0.283 and 0.179 respectively). 2008 Hispanic participants made high gains to sampling scale scores post CPR (d=0.800), while 2008 White participants made small improvements (d=0.297). In addition, 2009 Hispanic and White participants show only small improvements to sampling scale scores post CPR (d=0.153 and 0.362 respectively).

Cohen's d analysis for 2007 Hispanic and White participants show small improvements to discussion scale scores (d=0.318 and 0.234 respectively). Results also reveal 2008 Hispanic students show a small improvement in post CPR discussion scores (d=0.338), while 2008 White participants made moderate gains (d=0.551). 2009 Hispanic and White participants showed small improvements to the discussion scale scores post CPR (d=0.439 and 0.298 respectively). Refer to Table A.14 for Cohen's d analysis for Hispanic and White participants.

Majors/Non-majors: Grading criteria scale performance

Figure 2.7 show that the majority of majors/non-majors meeting the objective, sampling and discussion criteria to a degree. 2008 majors slightly outperformed non-majors for the objective, sampling and discussion scales. Figure 2.8 summarizes Cohen's d effect sizes for CPR pre-post scores for Majors/Non-Majors.

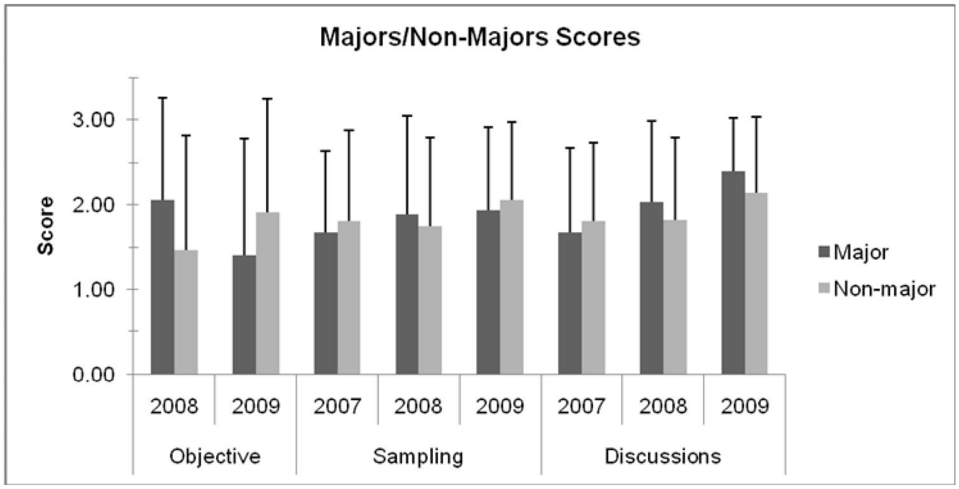

Figure 2.7. Overall pre-CPR performance of Majors/Non-majors for the objective, sampling and discussion scales.

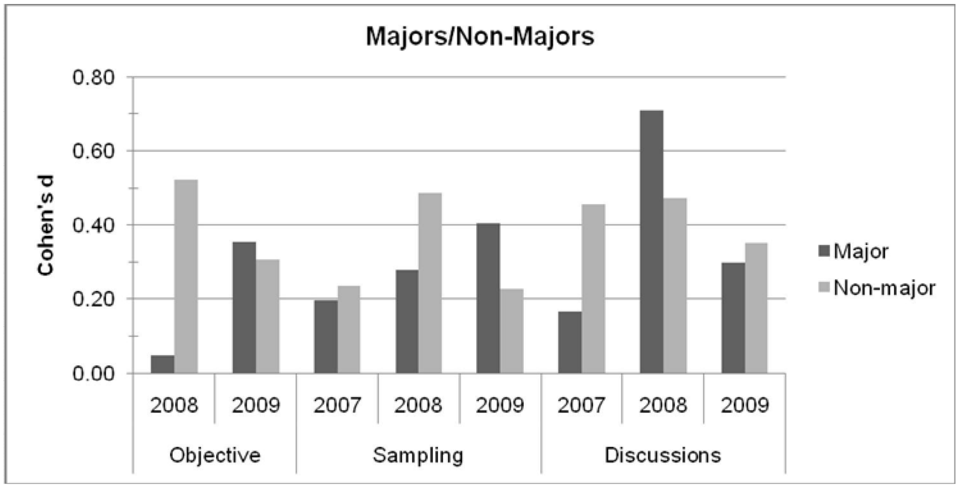

Figure 2.8. Cohen’s d pre-post comparisons of majors/non-majors for objective, sampling and discussion scale.

Cohen’s d analysis show 2008 majors show no improvement to objective scale post CPR (d=0.049), while 2008 non-majors made moderate gains to Objective scale post CPR scores (d=0.522). Results indicate 2009 majors and non-majors made a small improvement to Objective scale scores post CPR (d=0.353 and 0.306 respectively).

Cohen’s d analysis indicate 2007 majors and non-majors show small improvements to post CPR sampling scale scores (d=0.196 and 0.234 respectively). Analysis also show 2008 majors made a small improvement to sampling scale scores post CPR (d=0.281); while 2008 non-majors show moderate improvement post CPR (d=0.488). Cohen’s d results reveal 2009 majors and non-majors show small improvements to sampling scale scores post CPR (d=0.403 and 0.227 respectively).

Cohen’s d analysis show 2007 majors made small improvements to discussion scale scores post CPR (d=0.166) and 2007 non-majors show a moderate gain in scores post CPR (d=0.473). Results indicate 2008 majors and non-majors show moderate gains in post discussion scale scores (d=0.710 and 0.473 respectively). 2009 majors and non-majors show a small improvement for discussion scale pre-post CPR (d=0.300 and d=0.351 respectively). Refer to Table A.15 for Cohen’s d analysis for majors/non-majors.

Discussion

CPR impact on quality of ecological inquiry reports

CPR and the revision process impacted three ecological inquiry report components. These included the objective, sampling and discussion grading criteria scales. Provided, if participants received CPR feedback to improve discussion, objective or sampling criteria, some of these revisions might have been easier to do compared to revisions of some of the other components which showed less improvements post CPR. For example, making revisions to improve hypothesis and data collection/analysis criteria would have required participants to re-design and conduct another ecological inquiry study. Conducting another study was not entirely feasible considering participants had one week to revise and re-submit their ecological inquiry reports for additional grading by the teaching assistant. Therefore, time constraints rather than lack of CPR feedback may have prevented participants from revising criteria for some of the ecological inquiry report components that showed less improvements post CPR. Furthermore, some of the grading criteria scales and questions may inherently have required participants to use higher order thinking skills which they may need more practice in developing. For example, students may need practice in generating testable hypotheses within the limits of Bear Cam data or interpreting results.

Influence of achievement level, gender, ethnicity and major/non-major

The impact of CPR on student writing varied with performance levels. Low performers had moderate or high effect sizes in the majority of the cases for the objective, sampling and discussion scales in the three years. Most notably, the scores of 2008 low performers increased with moderate to high effect sizes in all three grading criteria scales. Low performers in 2009 had moderate to high increases for the objective and sampling scales. The majority of average and high performers showed little improvements to objective, sampling and discussion scales post CPR, except for the discussion scale in 2008, as indicated by small effect sizes. High performers in 2009 showed a significant decrease in pre to post CPR scores for the sampling scale and this decrease was directly related to students' providing a insufficient description of how they collected their ecological inquiry data sample. Past studies examining how multiple CPR use affects student writing report low performers gain the most over time while high performers who initially perform well on the first CPR assignment decline (Gunersel et al., 2008; Gunersel and Simpson, 2009). In this study pre-CPR scores for the objective, sampling and discussion scales were consistently lower for low performers compared to average and high performers. Therefore, low performers likely received lower CPR scores and possibly feedback from peers in CPR which might have prompted them to revise these components. Since these components had considerable room for improvement and were not difficult to change for the most part, the revision efforts of low performers would likely result in more significant improvements to post CPR scores.

There are three factors that could have contributed to the low performers' large gain in 2008 which was a substantial improvement from 2007. First, the 10 question CPR grading criteria instrument used for CPR calibration, peer review and self-assessment in 2007 was replaced with a 28 question CPR grading criteria instrument that more explicitly described the expectations (see Table A.5 and A.6 in Appendix A). Second, the instructor used an entire class session to explain and train students how to apply the new 28 question CPR grading criteria instrument. Third, in response to student requests, the course instructor and teaching assistant held a voluntary evening session prior to CPR in Fall 2008 to discuss and answer student questions on the writing of ecological inquiry reports. Students were encouraged to bring laptops, drafts of their ecological inquiry reports, and questions related to their ecological inquiry projects. It is plausible that students who attended the evening session could have enhanced their understanding of the instrument and the expectations and therefore could have been better prepared to revise post CPR report if they received low CPR peer scores on their initial submission.

Hartberg et al. (2008) reported little difference between the performance of female and male students when comparing CPR generated scores. Similarly, the results of this study show no consistent pattern of gender difference. Interestingly, however, there were gains of moderate effect sizes in 2008 - females for the sampling scale and males for the discussion scale. In these cases, groups had lower pre-CPR scores and therefore had higher effect sizes or gains post CPR.

There is no consistent pattern with respect to ethnicity across the three sample years. In 2008, however, Hispanics had gains of moderate to high effect sizes and white students had a low one for the objective and sampling scales. Furthermore, in 2008 for the discussion scale, White students had a moderate effect size and Hispanics had a low effect size. Hispanics had lower pre-CPR scores compared to Whites for the respective scales and this may have contributed to their gains of a higher effect size. It would be interesting to explore if cultural backgrounds of different races play a role in differential performance for different components of the ecological report.

It would seem logical to expect majors to outperform non-majors given their more extensive background related to the contents and process of ecological inquiry. Non-majors, however, made more gains with moderate effect sizes than majors. Particularly in 2008, non-majors had gains with moderate effect sizes for the objective, sampling and discussion scales. Majors did have a gain with moderate effect size for the discussion scale in 2008. Our definition for major and non-majors based on the college affiliation may be a limitation. Some of the majors in the College of Agriculture and Life Sciences (for example, Recreation, Parks and Tourism Science) may actually have less biological or ecological background than those from other colleges (for example, Biology and Geography). A more refined classification based on individual majors may be more appropriate or rather a classification based on the number of science courses participants have completed at the time of course enrollment.

Conclusion

Calibrated Peer Review was developed to promote anonymous peer review of writing, a critical component to the scientific inquiry process. As an educational tool CPR can offer instructors in any discipline a way to develop and implement peer review writing assignments. Past research studying impacts of CPR on student scientific writing and reviewing skills have largely relied on examining changes in CPR generated scores. Therefore, those findings may be specific to CPR assignment grading criteria design, the nature of CPR writing assignments, and CPR instructor-set scoring regimes. This study independently examined student scientific writing separate from CPR holistic scoring measures. This independent examination using a rubric scoring instrument enabled more consistent and reliable assessment of the effect of CPR and revision on the quality of the individual components of the ecological inquiry reports.

The results show significant impact of CPR and revision on the quality of three components, the objective, sampling and discussion scales, of ecological inquiry reports. These findings are promising in affirming the positive effect of CPR on student learning. They also beg questions on why CPR had an effect on these components but not others. The nature of the individual grading criteria components may require different types and level of skills. For example, creating a testable hypothesis or interpretation of findings may take considerably more skills and effort on the part of the student. The design and weighting of the rubric questions may influence student understanding and behavior in using the rubric for writing and revision of ecological inquiry reports. Further research is needed to explore these aspects for purposes of substantially enhancing student learning.

No consistent pattern emerged in this study on the influence of performance level, gender, ethnicity and major on the effect of CPR and revision to the quality of student writing. There are many interesting patterns related to these factors, such as the gains of moderate to high effective sizes for low performers, the limited but differential pattern of noticeable gains by gender, the noticeable gains of Hispanics in objective and sampling scales and gains of Whites for discussion scale, and the more common noticeable gains for non-majors. Many of these hints of pattern were observed in 2008 when substantial planned and non-planned instructional changes occurred. These hints suggest that further studies are warranted.

Future research assessing the impact of CPR on scientific writing should further assess how grading criteria instruments developed for CPR assignments affect student writing outcomes. In addition to explicitly identify the components and associated expectations, differential weighting schemes may also be necessary. For example, accurately describing the sampling procedure is less challenging than developing a hypothesis that is testable given the limitations of Bear Cam data. Within the discussion scale, interpretation of the results to explore possible mechanisms or explanations requires considerably greater higher-order thinking skills than identifying a limitation of the study and suggesting possible future studies. Differential weighting can encourage students to make more efforts on those more important tasks hence help them further develop the essential competencies. With equal weight designation of points among components and grading criteria, some students may opt to revise easier criteria in order to reach an adequate grade. Differential weighting may also help generate more meaningful evaluations of student learning, the impact of CPR, and the influence of possible factors such as performance level, gender, ethnicity, and majors. It is essential that future CPR research include evaluations of the types of changes CPR grading criteria instruments promote and provide recommendations on how to develop more effective CPR assignments.

3. EXPLORING THE REVISION PROCESS FOLLOWING CALIBRATED PEER REVIEW (CPR) THROUGH TEXT ANALYSIS AND ITS IMPACT ON QUALITY OF ECOLOGICAL INQUIRY REPORTS

Introduction

Increasingly, instructors are adopting e-learning tools and technology that are capable of helping instructors manage and author scientific inquiry and writing assignments. One such learning tool is Calibrated Peer Review (CPR). UCLA researchers from the Division of Molecular Sciences developed Calibrated Peer Review (CPR) in 1995. The CPR system was designed based on a model of scientific inquiry and functions to replicate scientific manuscript submission and anonymous peer review (Carlson and Berry, 2003, 2008). The purpose of CPR is to develop student writing and anonymous peer review skills in order to promote critical thinking and mastery of science content (Kennicutt et al., 2008). CPR is a web-based system that serves two main functions. First CPR provides instructors with tools to develop and manage student writing projects. Second, CPR provides students with tools to self-assess and anonymously peer review writing projects assigned by instructors. CPR is made up of four structured workspaces (Carlson and Berry, 2003) and students are required to complete them in a sequential manner.

- Text entry workspace: Provides students with a writing prompt, instructor provided grading criteria and source material. Upon completion of their writing task, students submit their writing to the text entry workspace.
- CPR calibration workspace: Students grade three benchmark writing samples (low, average and high quality) of their current writing assignment. The three benchmark writing samples are graded by students using the instructor provided grading criteria provided in the text-entry workspace. Student grading is compared to instructor grading of the three benchmark writing samples. Ultimately, students are given a reviewers competency index (RCI) score (1 to 6) representing how well their grading matched the instructors. If students fail to receive an acceptable RCI score (previously set by the instructor) they will have to repeat CPR calibration.

- CPR peer review workspace: Students are randomly assigned three essays from their class to anonymously peer review. During CPR peer review students will grade randomly assigned essays using the instructor provided grading criteria they used in the text-entry and calibration workspaces.

- CPR self-assessment workspace: During CPR self-assessment stage students are required to grade their own writing essay originally submitted in the text-entry workspace at the start of CPR. Again, students will grade their own writing using the instructor provided grading criteria used in the three previous CPR workspaces.

A CPR assignment library is available to instructors who are not interested in authoring their own CPR assignments. Instructors who choose to author and implement their own CPR assignment can expect to complete the following tasks.

- Create CPR assignment goals, instructions and timeline.

- Develop CPR writing assignment criteria (writing prompt and length requirements).

- Generate grading criteria to be used for all CPR workspaces (text entry, calibration, peer review and self-assessment).

- Produce and grade writing project samples of low, average and high quality to serve as benchmarks for calibration. Instructor grades for writing samples will be compared to student generated grades during CPR calibration.

- Organize and upload supplemental resource material for CPR writing assignment.

- Set-up CPR student scoring. For example, instructors set acceptable percent of style and content calibration questions students must answer correctly in order to pass CPR calibration (See Table A.1 in Appendix A).

- Review student generated grades in CPR and identify and re-score student peer reviews that were graded by students who failed the CPR calibration process.

Prior to participating in CPR students are required to complete a CPR tour and training module. After training students are allowed to access the CPR system. Students must complete all CPR tasks according to the instructor set timeline. This will ensure all students participating in CPR have submitted their writing assignments to the CPR text entry workspace before any student can proceed to the calibration, peer review and self-assessment workspaces. After students complete the CPR self-assessment workspace (the last stage in CPR) they are allowed to review results of peer review and their overall writing assignment grade.

CPR does not include a revision workspace, and therefore; once students finish the CPR self-assessment stage (last stage in CPR); there is no continuation of the peer review process. One study reported 40% of upper-level students and 63% of introductory-level students opted to revise at least one essay when provided with an opportunity to incorporate feedback post CPR for extra-credit (Pritchard, 2005). CPR assignments without post revision opportunities prevent students from practicing the "revise and re-submit" stage of peer review. The "revise and re-submit" stage of peer review is critical to developing students' higher-order level thinking skills. During the "revise and re-submit" stage of peer review students make decisions to incorporate or disregard peer feedback; invariably learning how to support their research decisions or reflect on how they can make them better. With the exclusion of a CPR revision workspace, there is no guarantee students will review and incorporate CPR feedback through revision of their writing unless the instructor assigns multiple CPR assignments for the same writing project.

The absence of a CPR revision workspace also hinders researchers from evaluating CPR impact on student scientific writing. Further complicating the task of assessing CPR's effect on student scientific writing is the discipline-specific context and structure of CPR writing assignments (Gunersel et al., 2008). CPR writing assignments vary in difficulty, design and their ability to promote higher order and critical thinking skills (Reynolds and Moskovitz, 2008). Recent research findings across a range of disciplines such as chemistry, biology, zoology, physiology and neuroscience have found that repeated use of CPR has led to improved student writing and reviewing skills (McCarty et al., 2005; Gunersel et al., 2008; Gunersel and Simpson, 2009; Gragson and Hagen, 2010). However, these studies only evaluated impact of CPR generated scores across multiple CPR assignments. Due to varying quality of peer grading and inconsistency among CPR scoring weight regimes, it is difficult to compare and generalize the findings of these studies which also offer limited insights to how post CPR revisions directly impact writing.

CPR generated scores are used to help instructors track students' writing and peer reviewing performance within the CPR system (See Table A.1 in Appendix A). When reporting how the use of CPR affects student writing and peer review skills researchers have relied on examining changes in CPR generated scores. Instructors assign multiple CPR assignments and evaluate how CPR generated scores changed from the first to the last CPR assignment. Often changes in CPR generated scores are used in research to report changes in students' writing and reviewing abilities (Gerdeman et al., 2007; Margerum et al., 2007; Gunersel et al., 2008; Gunersel and Simpson, 2009).

CPR generated scores commonly used to report changes in students' writing and reviewing ability as a result of CPR use are the text rating (TR) and reviewer competency index (RCI) scores. The TR score (1 to 10 points) is a weighted average of three peer reviews given to a CPR writing assignment. Changes in TR scores for subsequent CPR assignments are often cited as evidence of changes in a student's writing ability. The RCI score (1 to 6 points) represents how well students applied the grading criteria during CPR calibration to peer review low, average and high quality benchmark writing samples. If student grading of benchmark writing samples deviates too much from the instructor's grading the student would receive a low RCI score. Therefore a peer review by a student with a low RCI score is given less weight than a student peer review that has a high RCI score. Recent CPR research uses the RCI score as evidence of a students' reviewing ability.

The reporting of changes in CPR generated scores across multiple CPR assignments has yielded evidence that CPR has a positive effect on student writing and reviewing ability. One study substantiates this finding by reporting a gradual increase in student RCI scores and low standard deviations between students' self assessment scores and peer review scores for students completing multiple CPR writing assignments that increased in level of difficulty (Margerum et al., 2007). Furthermore, when comparing CPR generated scores in low and high difficulty assignments, researchers reported that students' ability to sustain relatively the same CPR scores between assignments substantiates learning had occurred (Carlson and Berry, 2008). Contrary to the above research findings, Walvoord et al. (2008) enlisted three graders (writing expert, zoology doctoral candidate and a zoology professor) to independently score a random sample of 60 reports from 20 different students and found repeated CPR use did not improve technical writing (grammar, style) nor students' ability to convey scientific knowledge (hypotheses, methods, data collected, interpretation of results and support of hypothesis). Dissimilar findings between CPR studies examining CPR generated scores and those enlisting independent graders indicate the need to further explore CPR impacts to student writing and reviewing skills.

Sole reliance on CPR generated scores for the purposes of evaluating CPR effects on student writing and reviewing skills should be minimized. Using changes in CPR generated scores to report improvement in student writing and reviewing skills has several limitations. For example, weighting of CPR scores is left up to the instructor's discretion and therefore may lack consistency among research studies (Walvoord et al., 2008). Aside from inconsistent weight designation by instructors on CPR generated scores, other problematic issues arise when students fail to perform well in calibration or complete CPR. CPR generated scores may be inadequate for research when a students' submitted text does not receive three peer reviews, graded by at least two reviewers who exceeded the allowed deviations set by the instructor, or reviewed by a student with a low RCI score. Lastly, ambiguous grading criteria questions that are used throughout all CPR stages can influence the quality of CPR generated scores.

This study will not examine changes in CPR generated score data to evaluate revisions in students' writing of ecological inquiry reports. The objective of this study was to develop a better understanding of the revision process following CPR by examining the frequency, distribution and impact of post CPR text changes to students' ecological inquiry reports. Exploration of post CPR revisions can assist instructor in identifying gaps in instruction and providing students with additional guidance specifically targeting ecological inquiry report components of low quality. Understanding the post CPR revision process can lead to improved designs of CPR assignments more capable of developing essential scientific writing skills in students.

This study will address the following research questions.

1. What are the type and distribution of revisions made by students post CPR and how do the revisions impact the quality of individual components of ecological inquiry reports.
2. What are the common errors present within individual students' ecological inquiry reports post CPR revision?

Examination of these research questions will contribute to the understanding of the influence of online anonymous student peer review on the revision process. This understanding can help guide future development of effective instructional strategies to enhance student understanding of scientific writing and peer review.

Methods

Context of study

This study was conducted at a Texas university in a introductory ecology course. All research activities were certified by the Institutional Review Board (IRB) of the university in 2007 and renewed in 2008 and 2009 (IRB Protocol Id: 2007-0534). Co-taught by two professors, the course Fundamentals of Ecology enrolled up to 500 students. Students were required to complete weekly online quizzes, 4 major exams and a web-based ecological inquiry project that involved the use of CPR (4 weeks).

The web-based ecological inquiry project was designed specifically to promote student understanding of how ecologists conduct and report research. The project required individual students to complete all the following tasks in order to receive full credit for the ecological inquiry project.

1. Observe grizzly bear behavior and spatial patterns at McNeil River Falls in Alaska using picture stills captured by a remote Bear Cam (Griffing, 2007).
2. Develop a study objective and testable hypothesis within the limits of the data provided (Alaska Grizzly Bear Cam picture stills) to explain observed grizzly bear behavior and spatial patterns.
3. Design a study to test the hypothesis developed.
4. Collect and analyze relevant data using the Alaska Grizzly Bear Cam stills.
5. Interpret the results and draw conclusions.
6. Participate in online discussion (10 students per discussion group) to exchange ideas and feedback on research objectives, hypotheses, study design and data analyses.
7. Write an individual one-page ecological inquiry report with three sections: Introduction (Objective and Hypothesis), Methods (Data Collection and Analysis), Results and Discussion/Conclusion. Directions for inquiry projects and writing assignment are provided in Appendix B.
8. Submit individual ecological inquiry reports to CPR and participate in online, anonymous peer review.
9. Revise one-page ecological inquiry reports based on CPR feedback.
10. Re-submit ecological inquiry reports online to WebCT Vista (Blackboard Vista) a week after CPR completion for teaching assistant grading.
11. Complete online survey to provide feedback on usefulness of CPR. In 2008 and 2009, students also were asked to rate the usefulness of their discussion group member's participation.

The grading criteria for the ecological inquiry project was revised from 2007 to 2009 (See Table A.2 in Appendix A). Throughout the three years when the inquiry project was implemented, the CPR workspace scoring criteria and the CPR calibration score criteria set by the instructor remained the same (See Table A.3 and A.4 in Appendix A).

CPR grading criteria were developed to evaluate student ecological inquiry reports that resulted from the completion of the web-based inquiry project. CPR grading criteria served three purposes:

1. Outlined requirements for individual students' one-page ecological inquiry reports (pre-post CPR);
2. Provided grading criteria for student use in CPR stages (calibration, peer review and self-assessment); and
3. Assisted teaching assistant and researcher grading of individual student pre and post CPR reports.

CPR grading criteria evaluate 9 separate scales (28 dichotomous questions) of scientific writing (See Table A.5). Each grading criteria question evaluates the presence or absence of a specific tenet of scientific writing. The grading criteria in Table A.5 were used by students in CPR for the 2008 and 2009 Fall semesters. In 2007, the grading criteria used by students in CPR consisted of only 10 rubric questions (see Table A.6). The rubric changed in 2008 and 2009 as a result of student feedback that grading criteria was unclear.

Study participants

The participants of this study included sophomore students registered for Fundamentals of Ecology in 2007, 2008 and 2009 Fall semesters. Study participants included only sophomore students who had attended the university for at least 2 semesters. Sophomore students who transferred from another university or college were eliminated from the study sample because there was no GPR on record for these students at the time of course enrollment. Also, sophomore students were eliminated from the study sample if they failed to complete both the pre and post CPR assignments. Respective sample sizes for sophomore student study samples in 2007, 2008 and 2009 were 89, 74 and 66. Table A.7 in Appendix A summarizes the demographic characteristics of study samples by year.

Inter-rater reliability

Prior to blindly grading student pre and post CPR reports with the grading criteria (9 scales), the researcher and a PhD candidate majoring in Ecosystem Science and Management conducted inter-rater reliability. First, inter-rater reliability judges (researcher and PhD candidate) completed training on how to apply the grading criteria.

During training the judges collaboratively graded 12 sample scientific reports using the rubric. Sample scientific reports were equally selected among low, average and high quality reports previously graded by the Fundamentals of Ecology teaching assistant in 2008. During rubric training judges developed criteria for each of the 28 questions to later assist the judges with pre and post CPR report grading. See Appendix C for individual question grading criteria developed during inter-rater judge training. Once rubric training was completed inter-rater reliability judges independently graded 12 reports. The 12 reports for inter-rater reliability were again randomly selected among low, high and average quality reports previously graded by the Fundamentals of Ecology teaching assistant in 2008.

Cohen's Kappa (κ) was used to assess inter-rater reliability and corrects for nominal-scale chance agreement between raters (Cohen, 1960). Kappa (κ) ranges from -1 (perfect disagreement) to 1 (perfect agreement beyond chance agreement). A zero Kappa (κ) score means inter-rater reliability judges failed to agree anymore than would be expected by chance. Negative Kappa (κ) values indicate judges agreed less than what would be expected by chance. Landis and Koch (1977) suggest kappa values from 0.41 to 0.60 are considered moderate while values above 0.61 to 0.80 are considered substantial. For this study a Kappa Score ≥ 0.500 was considered an acceptable level of agreement between raters considering all disagreements carried equal weight. Kappa scores for inter-rater reliability between judges for each of the 9 scales evaluated by the 28 rubric questions are provided in Table A.8 (Appendix A).

Data sources

After inter-rater reliability was completed all pre and post CPR reports for qualifying sophomore student participants in the 2007, 2008 and 2009 Fall semesters were given a unique random identifier. All pre and post CPR reports were ordered by the unique random identifier and blindly graded by the researcher using the grading criteria (See Table A.5). The researcher graded a total of 458 pre and post CPR reports (2007, 2008 and 2009 samples). The pre and post grades generated for individual grading criteria scales and individual questions for each pre and post student scientific report served as the primary data source for this study. Pre and post CPR grades for individual ecological inquiry reports were used to examine post CPR revision impact to student scores of grading criteria questions and scales. Also, pre and post CPR grades were summarized to report overall student performance on individual components of ecological inquiry reports. The writing component grading criteria questions were not included in the present analyses (Table A.5). The writing grading criteria did not have the same point or weight designation as other grading criteria criterion scales.

Individual student data related to GPR at time of class enrollment was requested from the university. Student GPR at time of class enrollment was used to group students into low, average and high performing groups. Three pre and post CPR reports (a total of 6) previously graded by the researcher were randomly selected from low, average and high groups for each sample year. Stratified random sampling of ecological inquiry reports for each sample year ensured the sample was well representative of students enrolled in Fundamentals to Ecology. A total of 54 (27 pre-post CPR) sample reports were selected and submitted to Turnitin.com. Turnitin.com analysis of pre-post CPR reports for 27 students yielded the text changes or revisions of ecological inquiry report components.

The turnitin.com text analysis of reports, pre-post CPR grades, ecological inquiry grading criteria and coding schemes in Figures 3.1 and 3.2 were used to report the distribution and quality impact of text changes to individual components featured within ecological inquiry reports. When no text changes were present the researcher reviewed post CPR scores to record how well students met the requirements for each grading criteria scale and individual question. In order to identify the types of revisions, the researcher recorded all text changes to grading criteria scales for each individual post CPR report analyzed using turnitin.com. When post CPR reports had <90% text match to pre reports, the pre-reports were reviewed by the researcher for the purpose of recording additions or deletions from pre to post CPR reports. In a separate examination of post CPR grades, the researcher identified where individual students failed to meet the requirements of the ecological inquiry grading criteria. After this, the researcher reviewed individual student post CPR reports for which grading criteria requirements were not met and recorded common errors present within individual components of ecological inquiry reports.

Data analysis

Analyses to summarize text analysis data were conducted with Microsoft Excel. Once text analysis was completed using turnitin.com, the frequency of coding measures (refer to Figure 3.1 and 3.2) used to assess impact on scores to pre-post CPR reports for individual grading criteria criterion questions and scales were calculated. Lastly, the frequency of students that did not meet requirements, met requirements to a degree and met requirement for grading criterion scales and individual items was also calculated. This analysis helped identify where text changes were distributed among grading criterion scales and items and characterized what components of ecological inquiry reports participants changed as a result of CPR use and revision. Furthermore, the analyses for this study examined how text changes impacted scores for grading criteria scales and individual questions. The researcher also summarized the types of revisions to individual ecological inquiry components that led to improvement in scores and common errors present in post CPR reports that kept individual students from meeting grading criteria requirements.

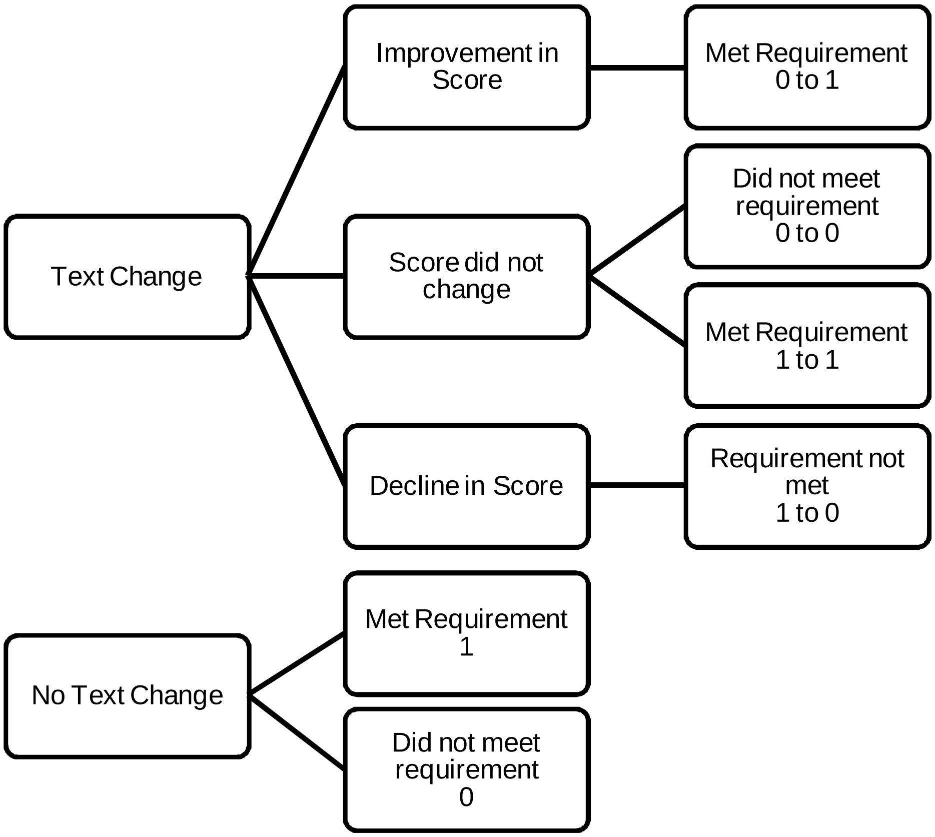

Figure 3.1. Individual grading criteria questions coding scheme characterizing impact of text changes and no text changes to post CPR scores of participant ecological inquiry reports.

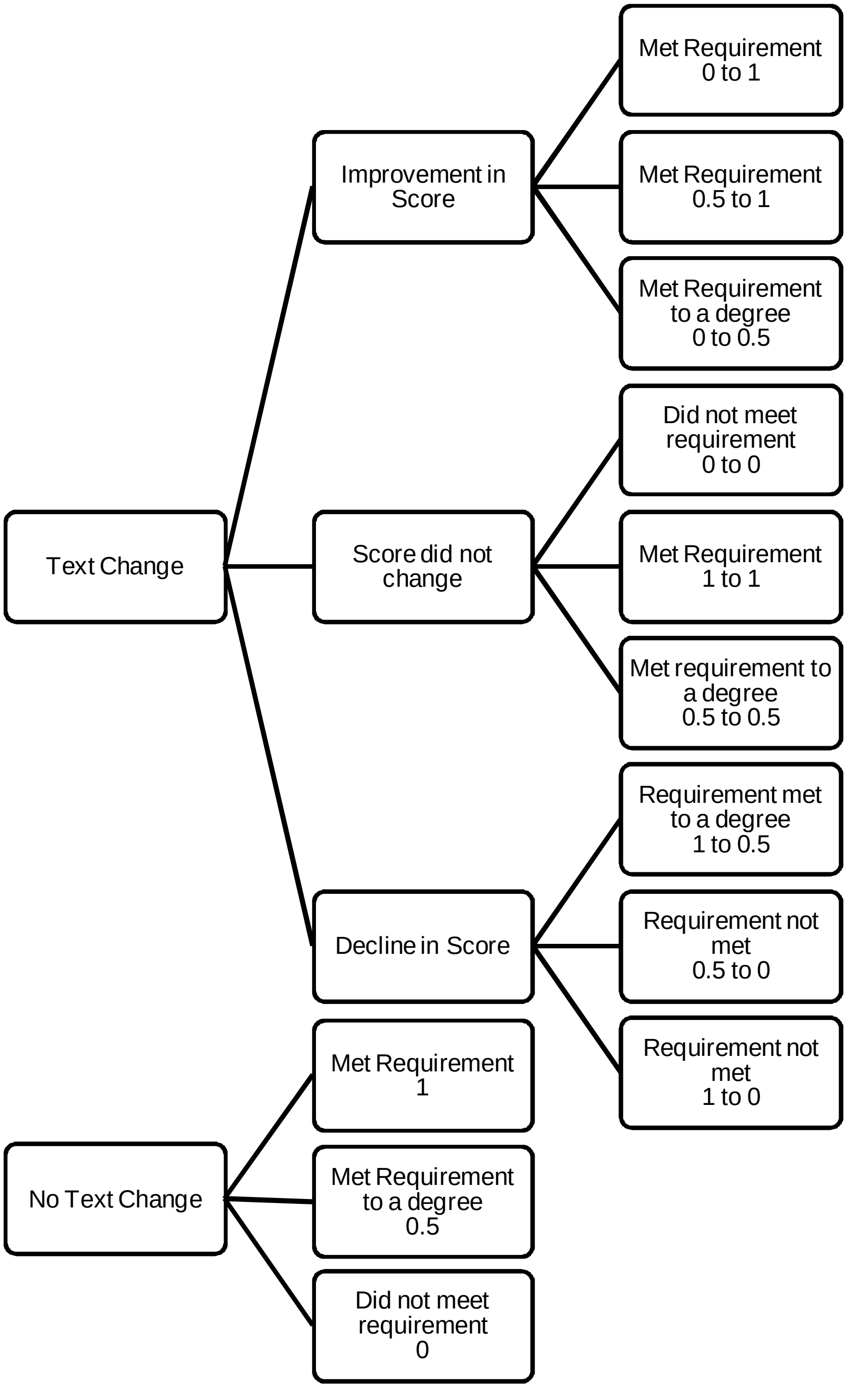
Text Change
Improvement in Score
Met Requirement 0 to 1
Met Requirement 0.5 to 1
Met Requirement to a degree 0 to 0.5
Score did not change
Did not meet requirement 0 to 0
Met Requirement 1 to 1
Met requirement to a degree 0.5 to 0.5
Decline in Score
Requirement met to a degree 1 to 0.5
Requirement not met 0.5 to 0
Requirement not met 1 to 0
No Text Change
Met Requirement 1
Met Requirement to a degree 0.5
Did not meet requirement 0

Figure 3.2. Grading Critiera Scales coding scheme characterising impact of text changes and no text changes to post CPR scores of participant ecological inquiry reports.

Results

Objective grading criteria results

The objective grading criteria scale included the following three questions for pre-post CPR text change analysis.

- Q1: Is the objective (what one is attempting to investigate) clearly stated?
- Q2: Is the objective clear? No objective = 0
- Q3: Is the objective reasonably specific given the study?

Post CPR grades reveal only 37.04% of participants met the requirements of the objective grading criteria scale (Figure 3.3). The majority of participants post CPR, met the objective scale requirement to a degree (40.74%). Text change analysis revealed that on average participant's revised roughly 26 to 30% of text related to objective scale grading criteria (Figure 3.4).

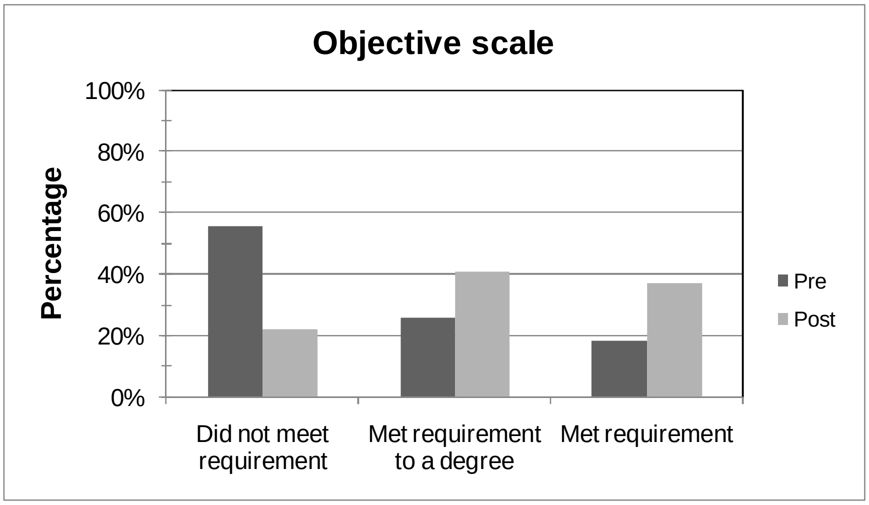

Figure 3.3. Overall participant performance on objective. Percentages of participants who failed to meet the requirement (score=0), met the requirement to a degree (score=0.5) and met the requirement (score =1) for the Objective scale are depicted.

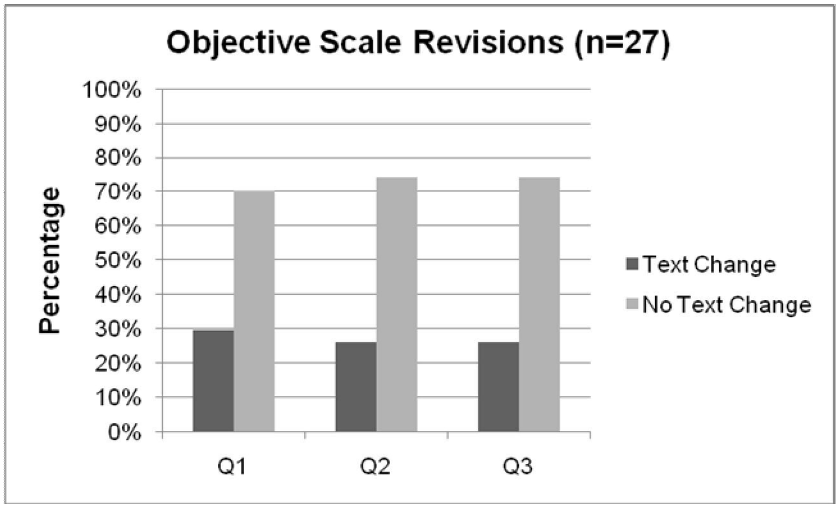

Figure 3.4. Percent text change to objective scale grading criteria for sample CPR reports.

Text change analysis for the objective scale (Table D.1) show that 25.93% of students make revisions and improved their score at the scale level, 7.41% made revisions but no change in score, and none made revisions that led to a decline in score. At the individual item level (Table D.2), 18.52% of the students made revisions that improved their score for both Q1 and Q2 of the Objective scale. Only 11.11% of the students made revisions associated with Q3 that improved their scores compared to 14.81% of them who made revisions related to Q3 but did not improve their scores. No decline in scores at the individual item level under the Objective scale.

Analysis of ecological inquiry reports shows 6 of the 27 sample reports had revisions to the objective grading criteria that led to an improvement in post CPR scores. A summary of the types of these revisions are provided below.

- Q1 (n=2) Revisions associated with stating an objective for their ecological inquiry report studies.
- Q2 (n=4) Text added post CPR made the objective clearly identifiable to the reader. Text changes included, “My objective” or “The purpose of my investigation.”
- Q3 (n=1) Revisions to text post CPR made the objective more specific to the hypothesis or question.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting objective scale requirements. These errors were not specifically related to revisions.

- Q1 (n=4) Participant failed to include a study objective in their ecological inquiry report.
- Q2 (n=4) Did not include text that made study objectives easily identifiable.
- Q3 (n=8) Included a broad study objective not specific to participants’ research hypothesis or question.
- Q3 (n=2) Incorported the instructor-provided learning objective of the ecological inquiry project rather than an objective specific to participants’ individual investigations.

Hypothesis grading criteria results

The hypothesis grading criteria scale included the following four questions for pre-post CPR text change analysis.

- Q4: Is hypothesis presented?
- Q5: Is the hypothesis logical?
- Q6: Is entire hypothesis testable with available data?
- Q7: Is part of the hypothesis testable with available data?

Post CPR grades show a little over half of participants met the requirements of the hypothesis grading criteria scale while 48.15% met the requirements to a degree (Figure 3.5). Text change analysis revealed that on average participant's revised 40.74% of text related to Q4 of the hypothesis scale grading criteria (Figure 3.6). 22.22% of text revised was related to Q5 while 14.81% of text revisions were associated with Q6 and Q7.

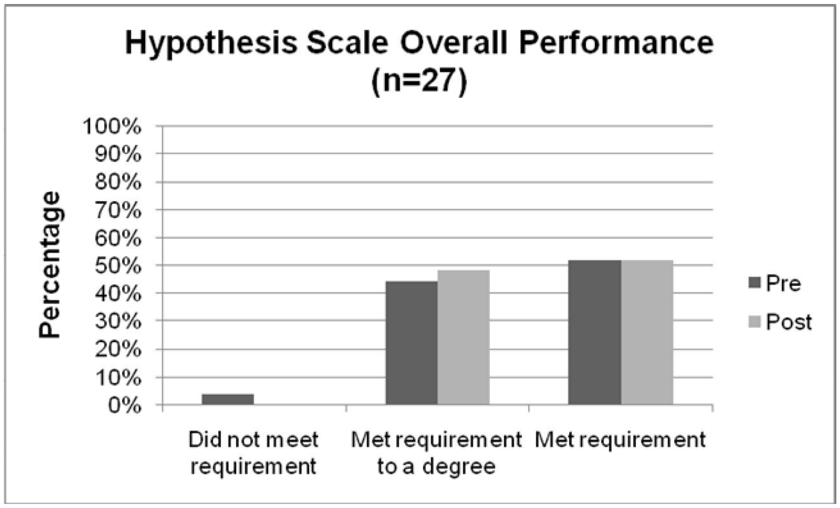

Figure 3.5. Overall participant performance on hypothesis grading criteria scale. Percentages of participants who failed to meet the requirement (score=0), met the requirement to a degree (score=0.5) and met the requirement (score =1) for the Objective scale are depicted.

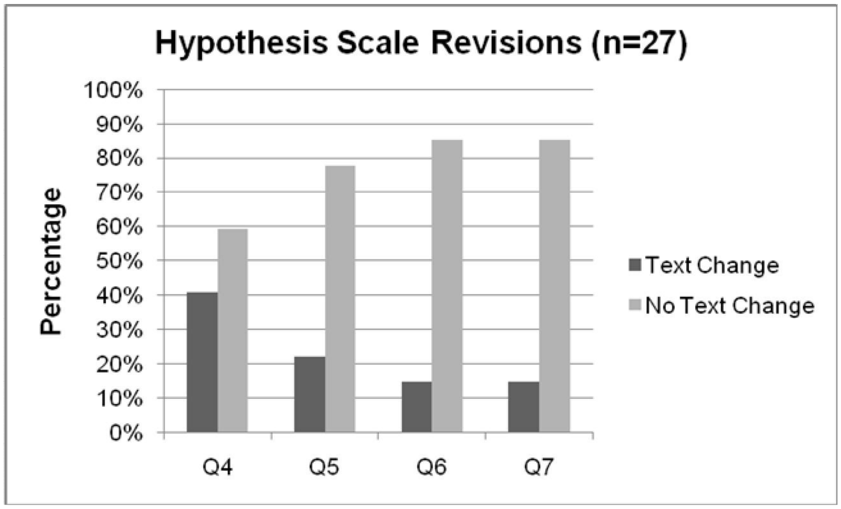

Figure 3.6. Percent text change to hypothesis scale grading criteria for sample CPR reports.

Text change analysis for the hypothesis scale (Table D.3) show 7.41% of revisions improved post CPR scores to a degree. 29.63% of revisions did not impact scores for the hypothesis scale compared to 3.70% of text changes that decreased post CPR scores. Text change analysis for individual hypothesis criteria (Table D.4) indicate 37.04% of the 40.74% text changes associated with Q4 did not impact or change post CPR scores. Half of text revisions associated with Q5 improved post CPR scores (11.11% of 22.22%). 7.41% of Q5 text changes made no impact to scores and 3.70% of revisions led to a decline in post CPR scores. A majority of revisions associated with Q6 and Q7 did not change scores (11.11% of 14.81%). Only 3.70% of revisions associated with Q6 and Q7 improved CPR post scores.

Analysis of ecological inquiry reports shows only 4 of the 27 sample reports had revisions associated with the hypothesis scale. Revisions for the hypothesis scale mainly included simple edits and not revisions which changed the nature of the participants' hypotheses. For example, all 4 reports added text post CPR that made the hypothesis clearly identifiable to the reader. Text included, "I hypothesize" or "My hypothesis." It is likely that improved post CPR scores were not associated with specific revisions to make hypotheses easily identifiable but rather grader disagreement of whether the hypothesis was logical and testable.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting hypothesis scale requirements. These errors were not specifically associated with text changes.

- Q4 (n=1) Hypothesis was mislabeled as the study objective.
- Q5 (n=7) Hypotheses included obscure language or terminology which prevented participants from receiving credit.
- Q6/Q7 (n=9) Contained un-testable research hypotheses that did not consider the limitations of the Alaska Grizzly bear data provided by the instructor. For example many un-testable hypotheses dealt specifically with attempting to examine a bears' fishing success in a particular location of McNeil River.

Sampling grading criteria results

The sampling grading criteria scale included the following three questions for pre-post CPR text change analysis.

- Q8: Is the number of samples (pictures) reported?
- Q9: Is the number of samples sufficient (>=30; or the maximum available if <30, but no less than 5)?
- Q10: Is there sufficient description for sample selection (how)?

Post CPR grades indicate only 29.63% of participants met the sampling scale requirement compared to 59.26% of participants who met sampling scale requirement to a degree (Figure 3.7). Text change analysis of ecological inquiry reports revealed that on average participant's revised 25.93% of text related to Q8, Q9 and Q10 of the sampling scale grading criteria (Figure 3.8).

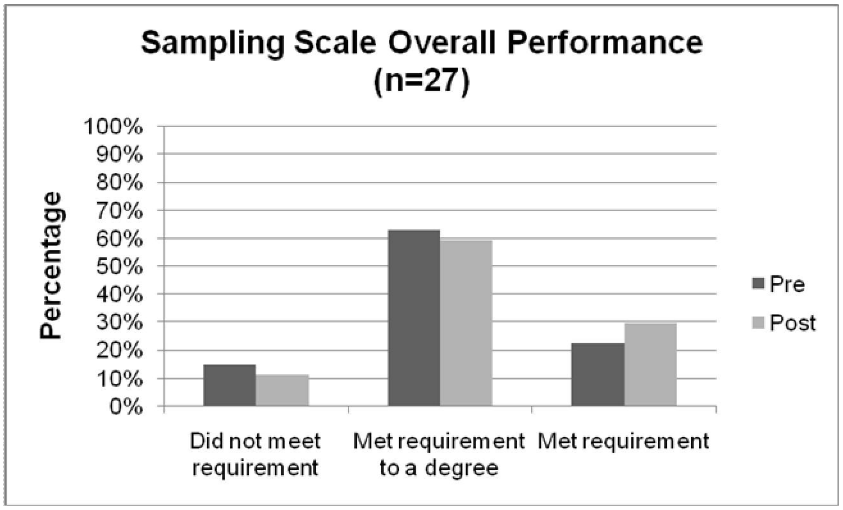

Figure 3.7. Overall participant performance on sampling grading criteria scale. Percentages of participants who failed to meet the requirement (score=0), met the requirement to a degree (score=0.5) and met the requirement (score =1) for the Objective scale are depicted.

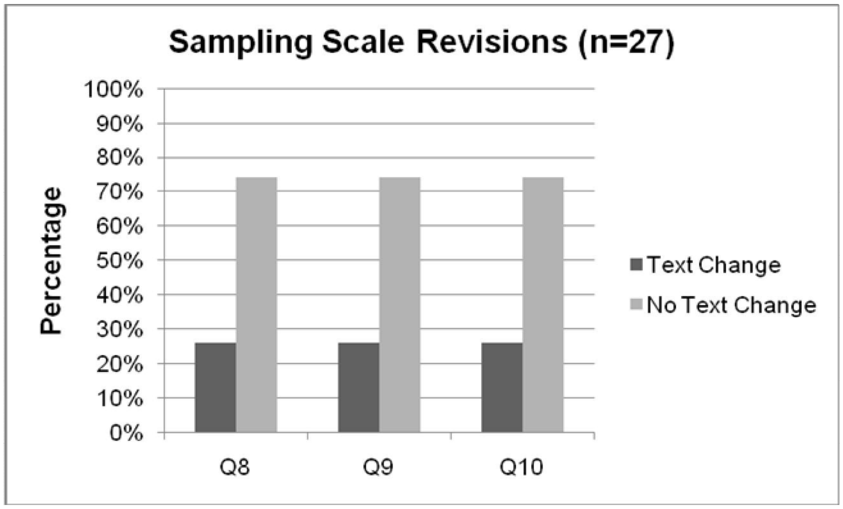

Figure 3.8. Percent text change to sampling scale grading criteria for sample CPR reports.

Text change analysis for the sampling scale (Table D.5) show that 11.11% of revisions led to a gain in post CPR scores. 25.93% of revisions had not impact on post CPR scores of the sampling scale. The majority of participants who failed to impact or change sampling scale scores post CPR with their text revisions met the sampling scale requirement to a degree (18.42%). Text change analysis for individual sampling criteria (Table D.6) indicate 18.52% of revisions improved scores for Q8 and Q9; while 7.41% of text changes led to no impact or change in scores post CPR. 11.11% of text changes related to Q10 improved post CPR scores compared to 14.81% of revisions that resulted in no change in scores post CPR. No declines in post CPR scores related to sampling grading criteria were found.

Analysis of ecological inquiry reports shows 14 of the 27 sample reports had revisions to the sampling grading criteria that led to improved post CPR scores.

- Q8 (n=6) Participants revisions included adding text post CPR that specifically cited their ecological inquiry data sample size (how many pictures were used).
- Q10 (n=8) Revisions were associated with providing a more clear description of how study samples of Bear Cam still pictures were selected.
- Q8, Q9 and Q10 (n=1) Text changes included inserting all three sampling scale criteria post CPR.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting sampling scale requirements. These errors were not specifically related to revisions.

- Q8 (n=4) Participant did not specify how many pictures were included in study samples.
- Q9 (n=3) Ecological inquiry report contained insufficient sample size. Insufficient sample size was not as a result of participants using the maxiumum amount of photos that only fit a specific set of criteria.
- Q10 (n=17) Inquiry report did not contain sufficient desscription of sample selection or specifically criteria used to select Bear Cam pictures participants' chose to use in their respective samples. For example, one student wrote, "The goal was to specifically seek out photos from the selected date that showed a bear

or several bears actively engaged in hunting or gazing into the water on any portion of the stream viewable to the camera for still pictures. All pictures/data that contributed to the objective of detailing the bear's tendencies in hunting or gazing into the shallow rapides were counted, carefully analyzed, and recoded from notes to excel spreadsheet." It was unclear to the researcher what the participant considered a actively hunting bear or gazing bear.

- Q8, Q9 and Q10 (n=3) Participants did not include any of the three sampling scale criteria in their ecological inquiry report.

Data collection/analysis grading criteria results

The data collection/analysis grading criteria scale included the following three questions for pre-post CPR text change analysis.

- Q11: Are the data (variables) collected appropriate for testing the hypothesis?
- Q12: Is there sufficient description for data collection (variables and how collected)?
- Q13: Is there sufficient description for data analysis (i.e. frequency, count, average, etc)?

Post CPR, 55.56% of participants met the data collection/analysis grading criteria scale requirement to a degree compared to 29.63% of participants who met the requirement (Figure 3.9). Text change analysis of ecological inquiry reports revealed that on average participant's revised 22.22% to 25.93% of text related to Q11, Q12 and Q13 of the data collection/analysis scale (Figure 3.10).

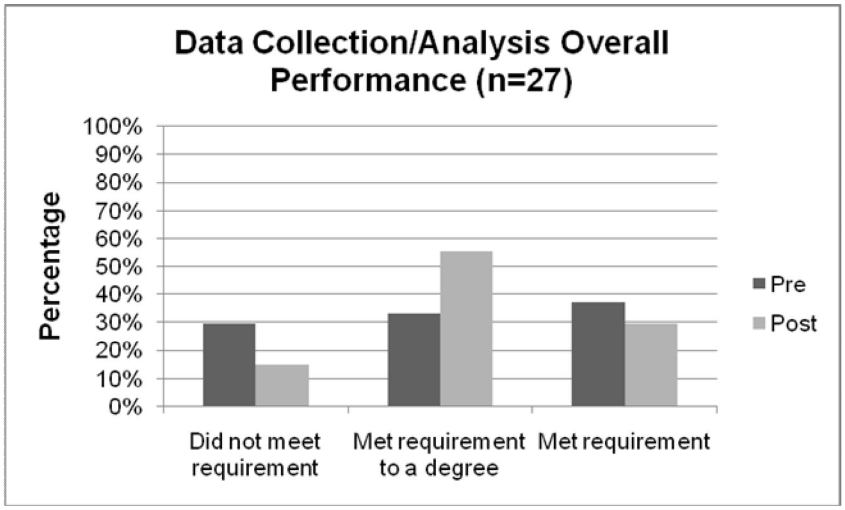

Figure 3.9. Data collection/analysis scale grading criteria percent text change for sample CPR reports.

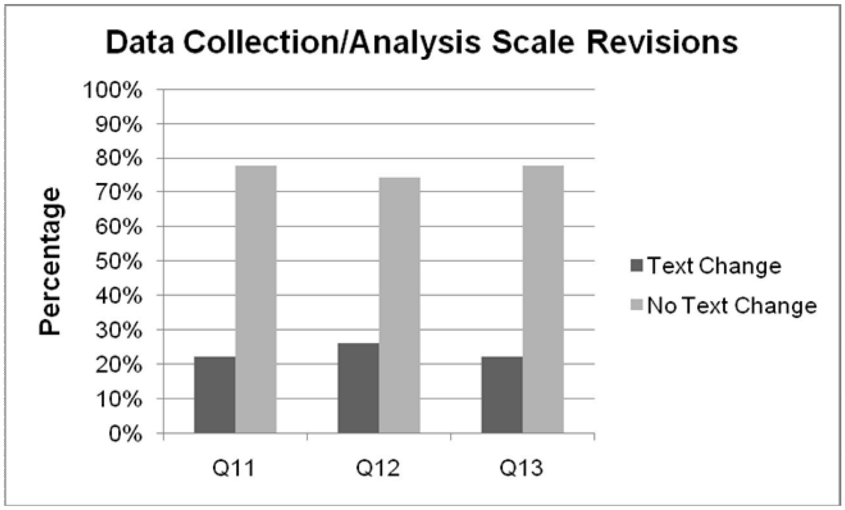

Figure 3.10. Percent text change to data collection/analysis scale grading criteria for sample CPR reports.

Text change analysis for the data collection/analysis scale (Table D.7) show that 11.11% of revisions led to a gain in post CPR scores while 18.52% of text changes resulted in no change. 3.70% of revisions led to a decline in post CPR scores for the data collection/analysis scale. Text change analysis for individual data collection/analysis criteria (Table D.8) reveal 3.70% of text changes improved scores related to Q11 and Q12. The majority of revisions related to Q11 and Q12 resulted in no impact to post CPR scores (22.22%). 11.11% of revisions associated with Q13 were equally distributed among categories that improved post CPR score and those that had no impact or change in score. No revisions resulted in a decline in data collection/analysis criteria scores.

Analysis of ecological inquiry reports shows 11 of the 27 sample reports had revisions to the data collection/analysis grading scale criteria that improved post CPR scores.

- Q12 (n=11) Ecological inquiry report included a better description of criteria used to collect specific data variables from pictures. For example, participants' specified criteria used to distinguish fishing bear from a resting bear.
- Q13 (n=9) Text revisions provided a better description of how participants' conducted data analysis of variables collected.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting data collection/analysis scale requirements. These errors were not specifically related to revisions.

- Q11 (n=8) Reported inappropriate variables for testing hypotheses. Only data variables associated with the testable portion of a participant's hypothesis were evaluated. Participants' grades were not impacted if they reported collecting variables related to part of their hypothesis that was untestable.
- Q12 (n=12) Participant did not include sufficient description of variables collected to test hypothesis in ecological inquiry report. 8 of the 12 ecological inquiry reports did not contain any text related to providing a description of how variables were collected. Insufficient description was not related to students'

reporting what data variables they collected from webcam pictures but rather specifically with criteria used by participants to collect data variables. For example, what did participants' consider criteria for identifying a male bear from a female bear or a fishing bear from an inactive bear.

- Q13 (n=12) Did not include sufficient description of data analysis. 6 of the 12 ecological inquiry reports did not contain any text related to providing a description of data analysis. 2 of the 6 reports that included a insufficient description of data analysis referred readers to a separate Excel file or figure and were not given credit. As a limitation of CPR, participants were not able to incorporate Microsoft Excel tables or figures in their ecological inquiry reports. For grading purposes, participants would not have received credit for providing sufficient description of data analysis (Question 13) if they simply referred the reader to a separate Microsoft Excel file or figure. Credit for Question 13 required students to describe what analysis was conducted in Microsoft Excel.

Result grading criteria results

The results grading criteria scale included the following three questions for pre-post CPR text change analysis.

- Q14: Are results presented?
- Q15: Are results specific (sums, averages, ratios)?
- Q16: Do results correspond to the variables specified in the Methods section? No results = 0

Post CPR grades indicate 70.37% of participants met the result scale requirement and 14.81% met the requirement to a degree (Figure 3.11). Text change analysis of ecological inquiry reports show participants revised 18.52% of text related to Q14, 22.22% of text related to Q15 and 14.81% of text related to Q16 for the results scale (Figure 3.12).

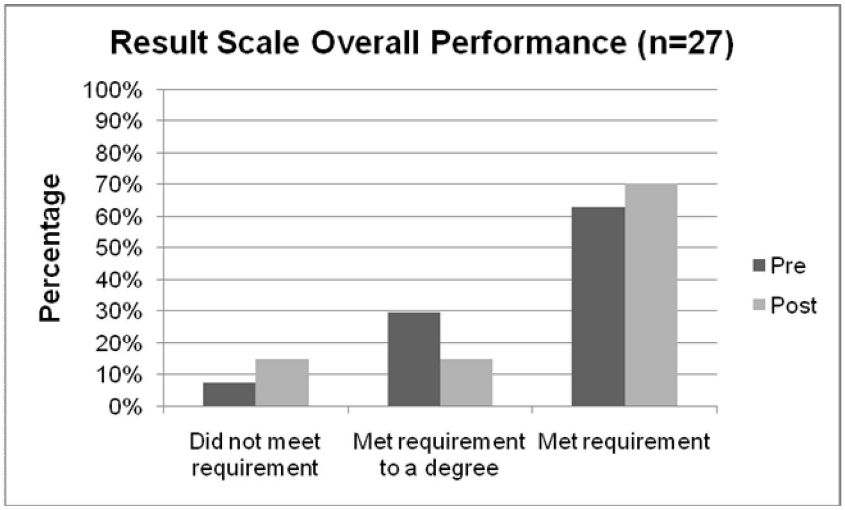

Figure 3.11. Result scale grading criteria percent text change for sample CPR reports.

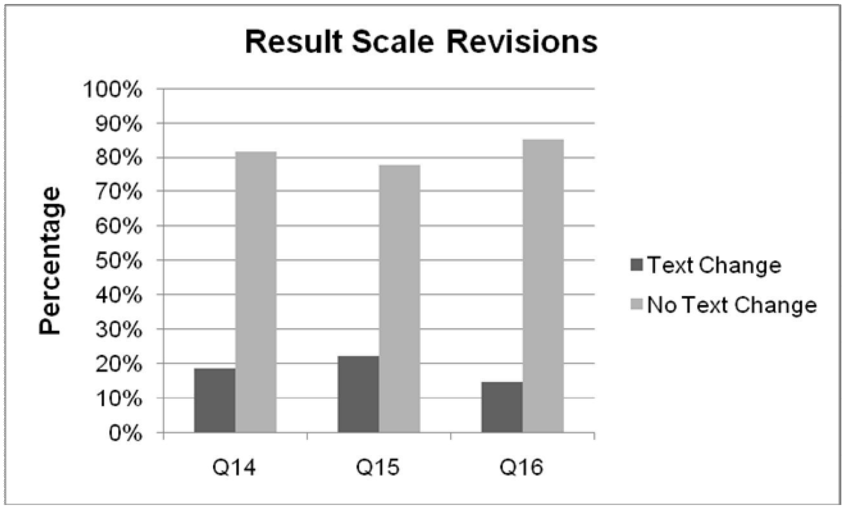

Figure 3.12. Result scale grading criteria percent text change for sample CPR reports.

Text change analysis for the result scale (Table D.9) show 14.81% of revisions led to a gain in post CPR scores while 7.41% of text changes resulted in no change. No decline in post CPR scores for the result scale was found. Text change analysis for result scale criteria (Table D.10) reveal 3.70% of text changes improved scores related to Q14 and Q16. The majority of revisions related to Q14 and Q16 resulted in no impact to post CPR scores (14.81% and 11.11% respectively). 11.11% of revisions associated with Q15 were equally distributed among categories that improved post CPR scores and those that had no impact or change in scores. No revisions resulted in a decline in scores for data collection/analysis criteria.

Analysis of ecological inquiry reports shows 3 of the 27 sample reports had revisions to the result grading criteria scale that improved post CPR scores.

- Q15 (n=3) Contained revisions that quantified results by citing actual averages or counts of reported data variables.
- Q16 (n=1) Included revisions specifically reporting results that aligned with data variables mentioned in the methods section of the report.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting result scale requirements. These errors were not specifically related to revisions.

- Q14 (n=2) No credit for result grading scale for failing to include any text related to result criteria.
- Q15 (n=5) Participant did not mention any specific results (i.e. average or count) related to analysis of data variables. 3 of the 5 ecological inquiry reports described trends that resulted from data analysis; however, they did not cite specific results that supported trends.
- Q16 (n=3) Cited results that did not align with data variables specified in the methods section of reports. Citing of results that did not align with variables in the methods section dealt specifically with participants' collecting variables that only aligned with the portion of their hypothesis that was un-testable.

Conclusion grading criteria results

The conclusion grading criteria scale included the following three questions for pre-post CPR text analysis.

- Q17: Are conclusions presented?
- Q18: Are the conclusions based solely on the results?
- Q19: Do the conclusions correspond to the hypothesis?

Post CPR grades reveal 55.56% of participants met the conclusion scale requirement and 37.04% met the requirements to a degree (Figure 3.13). Text change analysis of ecological inquiry reports show participants revised 18.52% to 22.22% of text related to Q17, Q18 and Q19 of the conclusion scale (Figure 3.14).

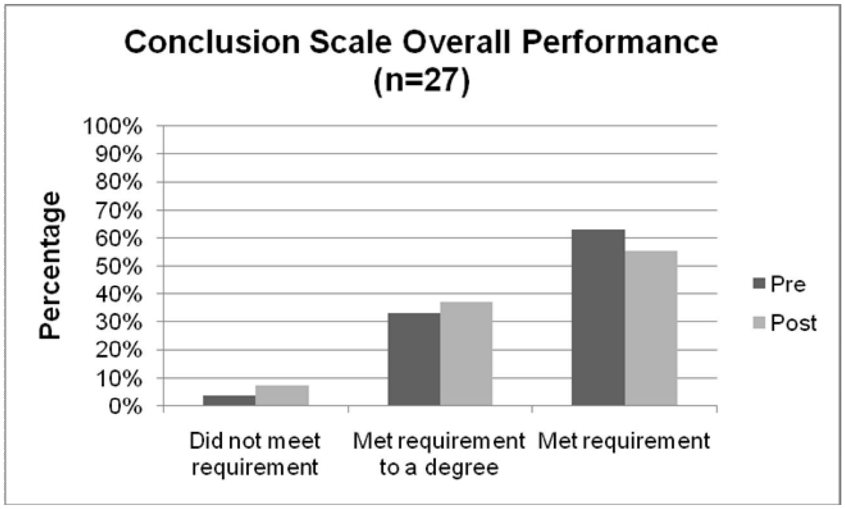

Figure 3.13. Conclusion scale grading criteria percent text change for sample CPR reports.

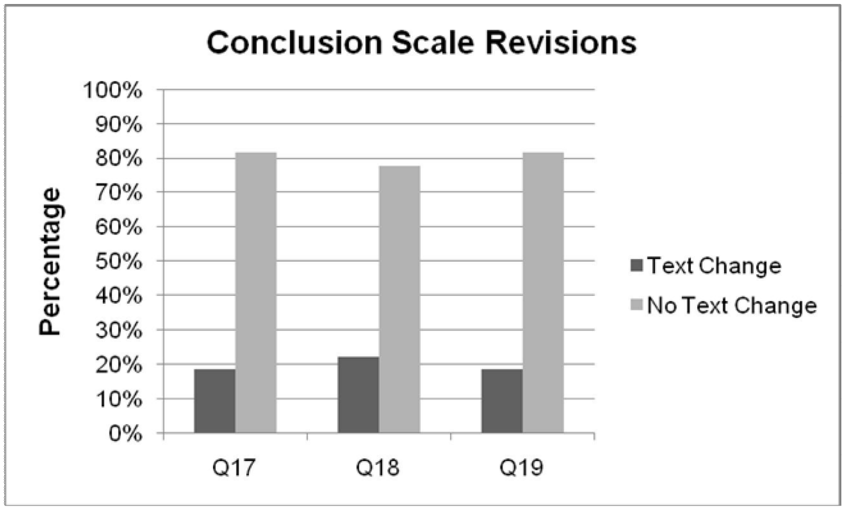

Figure 3.14. Conclusion scale grading criteria percent text change for sample CPR reports.

Text change analysis for the conclusion grading scale (Table D.11) show 7.41% of revisions improved post CPR conclusion scale scores. 11.11% of revisions did not change post CPR scores and 3.70% of text changes led to a decline in scores for the conclusion scale. Text change analysis for conclusion scale criteria (Table D.12) indicate Q18 was the only criteria in which revisions led to an increase in post CPR scores (7.41%). 11.11% of text revisions related to Q18 did not impact scores post CPR. The majority of text changes associated with Q17 and Q19 did not impact or change post CPR scores (18.52%).

Analysis of ecological inquiry reports reveal 3 of 27 sample reports had revisions to the conclusion grading criteria scale that led to an improvement in post CPR scores. Revisions to these three reports included the addition of a conclusion post CPR.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting conclusion scale requirements. These errors were not specifically related to revisions.

- Q17 (n=2) Provided no conclusion post CPR.
- Q18 (n=10) Included a conclusion that was not based on the results. Reports failed to provide a conclusion based on the results because of a un-testable hypothesis or due to no inclusion of data results.

Discussion scale grading criteria results

The discussion grading criteria scale included the following three questions for pre-post CPR text analysis.

- Q20: Is there interpretation (possible/hypothesized mechanisms or explanations) of the results?
- Q21: Is there discussion on limitations of the study?
- Q22: Is there discussion on future studies/new questions?

Post CPR, 44.44% of participants met the discussion scale requirement and 55.56% met the discussion requirement to a degree (Figure 3.15). Text change analysis of ecological inquiry reports show participants revised 22.22% of text related to Q20 and Q21 of the discussion scale (Figure 3.16). Participants revised 29.63% of text related to Q22.

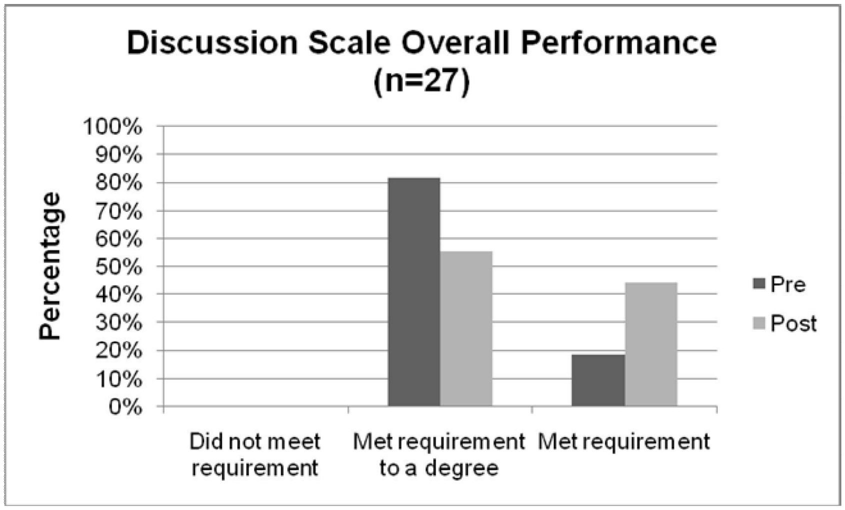

Figure 3.15. Discussion scale grading criteria percent text change for sample CPR reports.

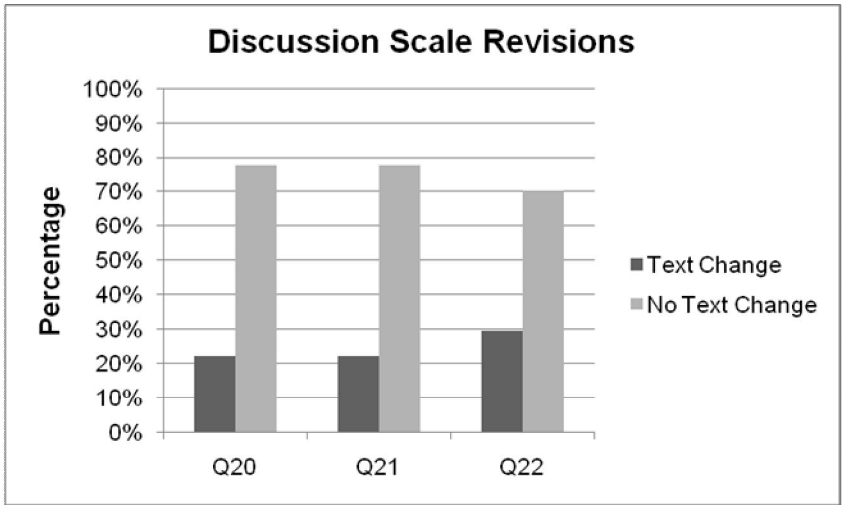

Figure 3.16. Discussion scale grading criteria percent text change for sample CPR reports.

Text change analysis for the discussion scale (Table D.13) show a total of 18.52% of revisions led to an increase in post CPR discussion scale scores. Similarly, 18.52% of revisions had no impact on post CPR scores. 3.70% of text changes resulted in a decline to post CPR discussion scale scores. Text change analysis for discussion scale criteria (Table D.14) show 7.41% text changes associated with Q20 improved post CPR scores and 7.41% of text changes resulted in no impact to scores. 7.41% of revisions related to Q20 led to a decline in discussion criteria scores post CPR. 11.11% of text revisions improved scores for Q21 while 11.11% of revisions had no impact. The majority of revisions associated with Q22 increased post CPR scores compared to 3.70% of text changes that resulted in a decline to scores.

Analysis of ecological inquiry reports shows 10 of the 27 sample reports had revisions to the discussion grading criteria scale that improved post CPR scores.

- Q20 (n=3) Revisions post CPR included interpretation or explanation of the results.
- Q21 (n=6) Included revisions that were related to participants' citing limitations to their study. 1 of the 6 reports simply included a text change of "While the study was limited by the fact…" making the study limitations easier to identify by the reader.
- Q22 (n=8) Revisions related to participants including discussion of further studies or new questions raised by their ecological inquiry studies.

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting discussion scale requirements. These errors were not specifically related to revisions.

- Q20 (n=12) Contained no text associated with the interpretation of results. 1 of the 12 reports had included an interpretation in the pre CPR report; however, text analysis of post CPR revisions revealed the participant deleted their interpretation from their report (decline in score).
- Q21 (n=8) Did not contain text associated with the inclusion of study limitations.
- Q22 (n=3) No text related to providing a description of future studies or new questions raised by participants' ecological inquiry.

Content placement grading criteria results

The content placement grading criteria scale included the following four questions for the text analysis of the content placement grading criteria scale.

- Q23: Is the report organized in three sections (Introduction & Hypothesis, Methods, Results & Discussion)?
- Q24: Is the Introduction & Hypothesis section free of content belonging to the Methods/Results & Discussion?
- Q25: Is the Methods section free of content belonging to the Introduction & Hypothesis/Results & Discussion?
- Q26: Is the Results/Discussion section free of content belonging to the Introduction& Hypothesis/Methods?

Post CPR, 62.96% of participants met content placement scale requirements and 33.33% met the requirements to a degree (Figure 3.17). Text change analysis of ecological inquiry reports show participants revised 11.11% of text related to Q23 and Q24 of the content placement scale (Figure 3.18). Participants revised 14.81% of text related to Q25 and 7.41% of text related to Q26.

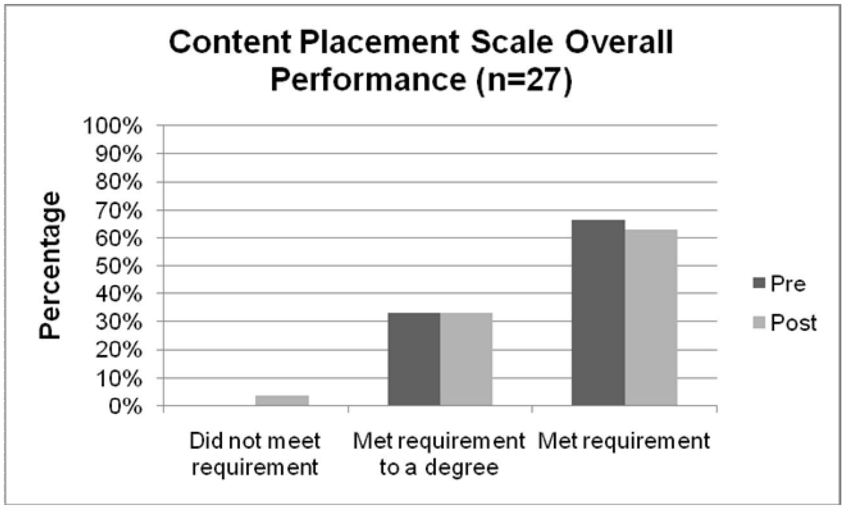

Figure 3.17. Content placement scale grading criteria percent text change for sample CPR reports.

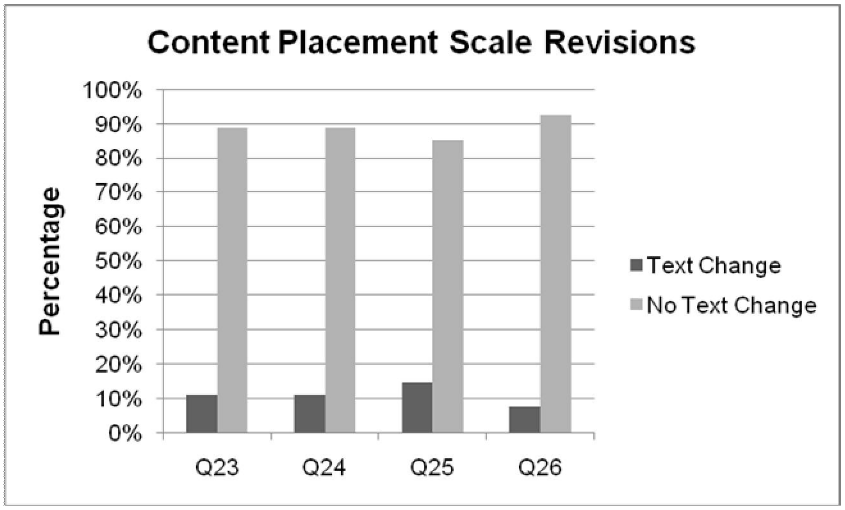

Figure 3.18. Content placement scale grading criteria percent text change for sample CPR reports.

Text change analysis for the content placement grading scale (Table D.15) show the majority of text revisions led to a decline in post CPR scores (7.41%). 3.70% of revisions increased post CPR scores for the content placement scale while 3.70% of revisions did not impact or change scores. Text change analysis for content placement scale criteria (Table D.16) show 7.41% of text changes associated with Q24 improved post CPR scores. Revisions related to Q23, Q25 and Q26 did not impact or change scores. 3.70% of text changes associated with Q23, Q24 and Q26 led to a decline in post CPR scores and 7.40% of revisions led to a decline for Q25 scores.

Analysis of ecological inquiry reports shows 2 of the 27 sample reports had revisions to the content placement grading criteria scale that changed post CPR scores.

- Q24 (n=1) Contained a text change that included a study interpretation in the Introduction of the report (decrease in post score). The student provided an interpretation of study results as reasoning for testing the hypothesis.
- Q24 (n=1) Introduction of pre-report contained interpretation; however the interpretation was deleted from the introduction in the post CPR report (increase in score).

Analysis of the overall quality of ecological inquiry reports revealed the following common errors which prevented participants from meeting content placement scale requirements. These errors were not specifically related to revisions.

- Q24 (n=2) Study findings provided in the ecological inquiry report introduction. Participants' used findings as reasoning for testing hypothesis.
- Q25 (n=2) Contained text related to study interpretation in methods section of report.

Overall percent text changes for grading criteria scales

Text-analysis shows overall percent text change for sample reports pre-post CPR was 17.70 % (n=27, Range: 0.0 to 93%). The average percent text change or revisions of post CPR reports for low, average and high performers was 31.78%, 17.33% and 4% respectively. Low performers made more revisions to post CPR reports than average and high performers. Overall grades of sample reports changed very little pre to post

CPR (refer to Table D.17 in Appendix D). Average grades for post CPR sample reports for low, average and high groups were 67.95%, 70.94% and 84.62% respectively (Figure 3.19). Low performer grades for post CPR reports were the lowest and therefore it would be expected that they would make more revisions to post reports than average and high performers. Study results are not reported by individual performance groups (low, average and high).

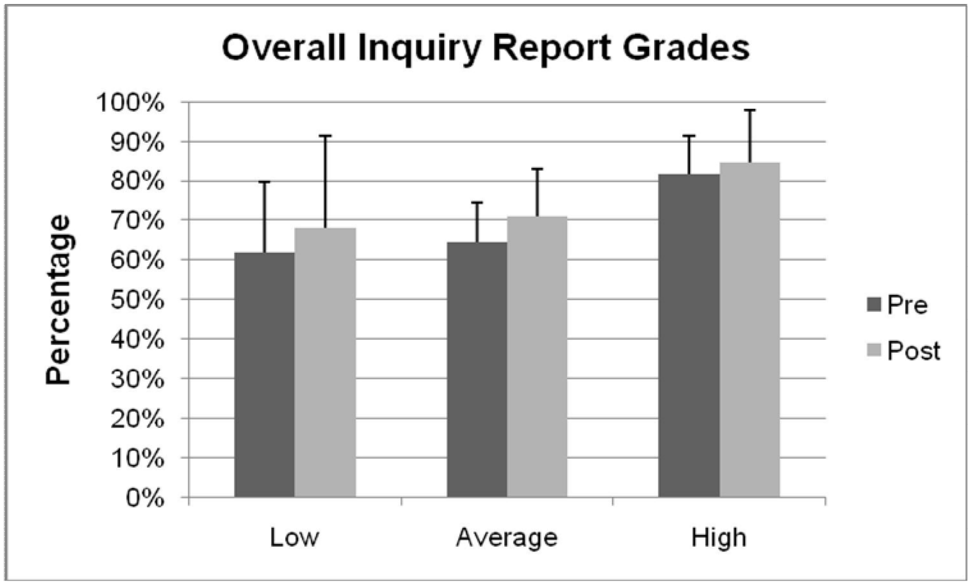

Figure 3.19. Overall ecological inquiry report grades for low, average and high performers in study sample.

Discussion

Type, distribution and impact of revisions

The ecological inquiry report components with the most revisions that improved post CPR scores were the objective, sampling and discussion scales. Most text changes to objective criteria dealt with participants including text that made the objective easier to identify. The majority of sampling scale revisions that led to increased scores was related to Q8 and Q9 (evaluated reporting of sample size and sufficiency of sample size). Nearly all revisions that led to increased discussion scale scores dealt with Q21 and Q22. Text changes for Q21 and Q22 detailed study limitations or future questions related to ecological study results.

Text changes to some ecological inquiry report components showed a high percentage of revisions but had little impact on scores. For example, Q4 of the hypothesis scale had the most revisions. However, the majority of revisions were simple additions of "My hypothesis is…" and not text changes resulting in improved scores. The data collection/analysis scale had a high percentage of revisions but little change in scores because post CPR revisions still lacked appropriate variables to test hypotheses or sufficient description of data collection/analysis. Revisions to improve hypothesis and data collection/analysis criteria would have been more challenging and might have required the student to re-design and/or re-do the data collection and analysis – basically requiring the participant to conduct another ecological inquiry. Conducting another inquiry might not be feasible considering participants had only one week to revise and re-submit their ecological inquiry reports. Therefore, failure of participants to improve in these ecological inquiry report components may be partially due to time constraints.

Few examples were found through text analysis of individual post CPR reports that justify a decline in scores. It was likely that declines in scores were due to grader (researcher) disagreement. The researcher blindly graded pre and post reports; and therefore, inconsistent grading could have led to lower scores in post CPR reports rather than revisions.

Errors present in post CPR ecological inquiry reports

Only the result, conclusion, and content placement scales showed a high number of participants that met scale grading criteria. A larger proportion of the participants met scale requirements 'to a degree' for the objective, hypothesis, sampling, data collection/analysis and discussion scales. There were likely three common reasons for low grades for ecological inquiry components. First, the high number of participants who failed to develop a hypothesis within the limits of the Bear Cam still picture data invariably had lower scores in many other scales. Second, participants often failed to provide sufficient descriptions for samples selection, variables selection, and data analysis for ecological inquiry reports. Providing students with more explicit instructions on the importance of providing sufficient details in the methods section to allow others to repeat the study may improve performance related to the method scale. Participants may have less experience in writing methods of a scientific paper considering most introductory science labs require students to carry-out experiments with prescribed or "recipe" instructions and only report results. Third, the tasks related to some of the scales or items were more difficult in nature and required the use of higher order thinking skills which participants may need more practice or guidance to develop. Additional structured learning activities, practice and feedback on interpretation of sample findings associated with this inquiry project as well as activities to develop higher-order thinking skills in other contexts of the course would likely result in improved performance related to these scales.

Possible instructional approaches to improve performance

Several instructional strategies may have potential to improve student performance through the CPR and revision process. One strategy is to offer more targeted learning resources and activities on key elements of the report with which many students tend to have difficulties. For example, supplemental learning resources such as more examples of testable and non-testable hypotheses based on the available data and in-class discussions and on-line exercises or quizzes based on these examples may improve students' ability in formulating research questions. A second strategy is possibly to assign higher weights to grading criteria of a higher difficulty level (i.e. interpretation) which may encourage students to pay closer attention to these aspects during initial writing and revision post CPR. In addition to insufficient guidance to relative efforts required for different elements of the ecological inquiry and report, having all items of grading criteria with equal point designations as we did also may not reflect the overall quality of individual ecological inquiry appropriately. Recognizing that the learning process is incremental we probably cannot expect large gains or improvement on the more difficult elements of the ecological inquiry and report writing. A third strategy could be to design and implement multiple small CPR projects focused on one or two difficult elements of ecological inquiry report writing to help students develop better skills through repeated practice in different contexts.

Recommendations for future research

Future CPR research should continue to study how revisions directly impact student writing. Most importantly, researchers should continue to develop research methods associated with qualitatively evaluating the effects of CPR. Qualitative research would allow a closer examination of the direct impact of CPR and revision on student writing in order to provide more direct evidence of student learning gains. By studying the mechanisms of revision post CPR, instructors could improve the quality and design of CPR assignments that promote student acquisition of scientific writing skills. Additionally, studies should examine how CPR grading criteria instruments influence post CPR revisions. Further study of how grading criteria instruments influence the quality of student writing can help researchers provide guidelines for designing CPR grading instruments that assist students in develop higher-order thinking and writing skills. Furthermore, researchers investigating student learning gains as a result of student use of CPR should report instructor-set CPR scoring for writing assignments. Reporting of CPR instructor set scoring in research articles could help future instructors and researchers assess the applicability of findings to future CPR assignments and studies.

Conclusion

Calibrated Peer Review was developed to promote anonymous peer review of writing, a critical component to the scientific inquiry process. As an educational tool CPR can offer instructors in any discipline a way to author and manage peer review writing assignments. Most studies on impacts of CPR on student writing and reviewing skills have relied on examining changes in CPR generated scores which are based on grading of a small number of peers and dependent on the specific instructor-set CPR scoring weight regimes. Due to the varying quality of peer grading and inconsistency among scoring weight regimes, it is difficult to compare and generalize the findings of these studies which also offer limited insights to the revision process and how post CPR revisions directly impact writing. This study examined the revision process associated with CPR through analyzing the distribution of text changes and their impacts on the improvement of ecological inquiry report components independent of CPR scoring measures. This approach allowed the exploration of the revision process by identifying specific problems that limited improvement through revision, and also helped to formulate intentional instructional modifications to effectively enhance student learning. As instructors continue to design and implement CPR assignments, it is critical to evaluate the impact of CPR in a manner that facilitates comparison among CPR studies and to explore the revision process through approaches like text analysis in order to understand how to better design and implement CPR assignments to improve essential writing and reviewing skills of students. Findings associated with this study indicate that when students are provided with a revision stage post CPR they are likely to make changes to criteria that are easier and require less higher-order thinking skills (i.e. interpretation). Instructors interested in developing higher-order thinking skills through writing should consider implementing grading criteria instruments containing appropriate weight designations that indicate the level of difficulty for specific criteria.

4. SUMMARY AND CONCLUSIONS

Calibrated Peer Review (CPR) has increasingly been adopted by instructors to promote student writing and peer review skills in science. Limited research has evaluated CPR impacts on student learning; specifically scientific writing skills. Recent CPR research has focused on examining changes in CPR generated scores over a semester in order to report student gains in writing and reviewing skills. CPR generated scores are influenced by instructor-set scoring, the clarity of CPR grading criteria instruments, the level of student participation during CPR, and the quality of peer grading. Therefore, findings generated from examining changes in CPR generated scores may be difficult to compare across CPR research studies. This study provided participants with a revision period post CPR, assessed changes in pre-post CPR writing independent of CPR holistic scoring measures and incorporated text analysis to examine the mechanisms of the revision process post CPR as well as its impact on the quality of student writing.

CPR impact to student writing of ecological inquiry reports was evaluated using a 28-item grading rubric (9 scales). Ecological inquiry reports (pre-post CPR) were blindly graded by the researcher using the rubric and 8 pairwise t-tests were conducted to compare pre-post CPR scores for individual grading criteria components. Results show that significant gains from pre- to post-CPR were related to the objective, sampling and discussion grading criteria.

Analysis using the cohen’s d effect size statistic explored how participant achievement level, gender, ethnicity and major/non-major status affected the impact of CPR and revision on the quality of ecological inquiry report components. Low performers had more moderate to large gains in post CPR scores compared to average and high performers. In addition, non-majors made more moderate to large improvements in post CPR scores compared to majors. Data trends also reveal limited but differential pattern of gains by gender pre-post CPR. Hispanic participants showed gains for the objective and sampling scale while White participants showed gains for the discussion scale. Interestingly, results also indicated non-majors showed more gains pre-post CPR compared to majors.

Text analysis was used to examine post CPR revisions to ecological inquiry report components. Revisions which improved post CPR scores included making the objective easily identifiable, reporting of sample size, discussion of study limitations and future questions related to ecological study results. Many ecological inquiry report components showed high percentage of revisions but no improvement in scores. Overall, participants only met grading scale criteria to a degree for the majority of the ecological inquiry components. There are three common causes for these intermediate overall grades. First, the high number of participants who failed to develop a hypothesis within the limits of the Bear Cam still picture data. Second, participants failed to provide sufficient descriptions of sample selection criteria, variables collected, and data analysis conducted. The third reason for low overall grades was that participants made most changes or revisions related to easier criteria that required less higher order thinking skills. Grading criteria associated with making the hypothesis logical or testable, basing conclusions solely on the results, interpreting results may require more higher order thinking skills which participants need more practice in developing.

Future research on the impact of CPR on student scientific writing needs to assess how grading criteria developed for CPR assignments affect student writing outcomes. In this study, participants made the most changes to the objective, sampling and discussion scales; and meeting criteria for these scales may have largely been easier than meeting criteria for some of the other scales. The equal weight designation of points among grading criteria could have led students to focus on revising easier criteria in order to reach an adequate grade. Targeted instruction on higher order thinking criteria can improve the overall quality of ecological inquiry components.

REFERENCES

Carlson, P.A., and Berry, F.C. (2003). Calibrated peer review and assessing learning outcomes. In: Engineering as a Human Endeavor: Partnering Community, Academia, Government, and Industry. ed. EP Innovations, Inc., Champaign (IL): Stipes Publishing L.L.C., 1-6.

Carlson, P.A., and Berry, F.C. (2008). Using computer-mediated peer review in an engineering design course. IEEE T Prof Commun. *51*, 264-279.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Ed and Psych Meas. *20*, 37-46.

Gerdeman, R.D., Russell, A.A., Worden, K.J. (2007). Web-based student writing and reviewing in a large biology lecture course. J Coll Sci Teach. *36*, 46-52.

Gragson, D.E., and Hagen, J.P. (2010). Developing technical writing skills in the physical chemistry laboratory: A progressive approach employing peer review. J Chem Educ. *87*, 62-65.

Griffing, L. (2007). Secret lives of the brown bears of McNeil River. http://griffing.tamu.edu/Site/McNeil/index.htm (accessed 10 January 2009).

Gunersel, A.B., Simpson, N.J., Aufderheide, K.J., Wang, L. (2008). Effectiveness of calibrated peer review for improving writing and critical thinking skills in biology undergraduate students. JoSoTL. *8*, 25-37.

Gunersel, A.B., and Simpson, N. (2009). Improvement in writing and reviewing skills with calibrated peer review. IJ-SoTL. *3*, 1-14.

Hartberg, Y., Gunersel, A.B., Simpson, N.J., Balester, V. (2008). Development of student writing in biochemistry using calibrated peer review. JoSoTL. *2*, 29-44.

Kennicutt, W.K., Gunersel, A.B., Simpson, N. (2008). Overcoming student resistence to a teaching innovation. IJ-SoTL. *2*, 1-26.

Landis, R.J., Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics. *33*, 159-174.

Margerum, L.D., Gulsrud, M., Manlapez, R., Rebong, R., Love, A. (2007). Application of Calibrated Peer Review (CPR) writing assignments to enhance experiements with an environmental chemistry focus. J Chem Educ. *84*, 292-295.

McCarty, T., Parkes, M.V., Anderson, T.T., Mines, J., Skipper, B.J., Grebosky, J. (2005). Improved patient notes from medical students during web-based teaching using faculty-calibrated peer review and self-assessment. Acad Med. *80*, 67-70.

Prichard, R.J. (2005). Writing to learn: An evaluation of the calibrated peer review program in two neuroscience courses. JUNE. *4*, 34-39.

Reynolds, J., and Moskovitz, C. (2008). Calibrated peer review assignments in science courses: Are they designed to promote critical thinking and writing skills? J Coll Sci Teach. *38*, 60-66.

Robinson, R. (2001). An application to increase student reading & writing skills. Am Biol Teach. *63*, 474-480.

Walvoord, M.E., Hoefnagels, M.H., Gaffin, D.D., Chumchal, M.M., Long, D.A. (2008). An analysis of calibrated peer review (CPR) in a science lecture classroom. J Coll Sci Teach. *37*, 66-73.

APPENDIX A

Table A.1 CPR generated scores defined

Score	Description
Text Rating (TR)	Average weighted score of peer reviews (1 to 10 points).
Text Score[a]	Percent weighted score of total Text Points.
Overall Grade	Total text, calibration, peer review and self-assessment scores (100 points).
Calibration % Style	Style calibration questions correctly answered.
Calibrations % Content	Content calibration questions correctly answered.
Calibrations Average Deviation	Student average deviation from instructor scoring of benchmark writing samples used in calibration.
Calibrations Score[a]	Percentage of total calibration points (Style + Content + Retake+ Average Deviation).
Reviewer Competency Index (RCI)	Algorithm that applies a weight to a student's peer review in the CPR system (1 to 6 points).
Reviews Average Deviation	Student review compared to average of other reviewers. (Summation of three reviews).
Reviews Score[a]	Weighted Score converted to percentage of total PEER REVIEW points.
Self-Assessment Deviation	Self-assigned holistic score compared to average of classmates' ratings
Self-Assessment Score[a]	Weighted Score converted to a percentage of total SELF REVIEW points.
a. Points set by instructor prior to each CPR assignment.	
Annotated table information from conference paper (Carlson and Berry 2008).	

Table A.2 Alaska Grizzly Bear web-based ecological inquiry project grading criteria

2007	Pts	2008 and 2009	Pts
Participation in online discussion	30	Participation in online discussion	30
Revised report (Graded by TA)	30	Revised report (Graded by TA)	30
Participation in CPR	30	Participation in CPR	30
Feedback on Grizzly Bear inquiry project (pre and post survey).	10	Feedback on the inquiry project (pre and post survey), and independent evaluation of group members' discussion participation.	10
Total Points:	100	Total Points:	100

Table A.3 CPR scoring criteria for 2007, 2008 and 2009

Workspace (CPR stage)	Points
Text quality	0
Calibrations	40
Peer Reviews	40
Self-assessment	20
Total Points	100

Table A.4 CPR calibrations score criteria

	2007	2008 and 2009
Calibrations Mastery	Students must answer 0 of 1 (0 %) style questions correctly, 4 of 9 (44.44%) content questions correctly, and not deviate more than 3 points from a calibration text rating.	Students must answer 1 of 3 (33.33%) style questions correctly, 13 of 18 (72.22%) content questions correctly, and not deviate more than 3 points from a calibration text rating.
Reviews Mastery	Students must not deviate more than 3 points from the average rating of the reviewed text.	Students must not deviate more than 3 points from the average rating of the reviewed text.
Self-Assessment Mastery	To receive full credit students must not deviate more than 3 points from the average rating of their text. To receive half credit students must not deviate more than 4.5 points from the average rating of their text.	To receive full credit students must not deviate more than 3 points from the average rating of their text. To receive half credit students must not deviate more than 4 points from the average rating of their text.

Table A.5 CPR grading criteria for 2008 and 2009

Objective	1. Is the objective (what one is attempting to investigate) clearly stated? 2. Is the objective clear? No objective = 0 3. Is the objective reasonably specific given the study?
Hypothesis	4. Is hypothesis presented? 5. Is the hypothesis logical? 6. Is entire hypothesis testable with available data? 7. **Is part of the hypothesis testable with available data?
Sampling Methods	8. Is the number of samples (pictures) reported? 9. Is the number of samples sufficient (>=30; or the maximum available if <30, but no less than 5)? 10. Is there sufficient description for sample selection (how)?

Table A.5 continued

Data collection/ Analysis Methods	11. Are the data (variables) collected appropriate for testing the hypothesis? 12. Is there sufficient description for data collection (variables and how collected)? 13. Is there sufficient description for data analysis (i.e. frequency, count, average, etc)?
Results	14. Are results presented? 15. Are results specific (sums, averages, ratios)? 16. Do results correspond to the variables specified in the Methods section? No results = 0
Conclusions	17. Are conclusions presented? 18. Are the conclusions based solely on the results? 19. Do the conclusions correspond to the hypothesis?
Discussion	20. Is there interpretation (possible/hypothesized mechanisms or explanations) of the results? 21. Is there discussion on limitations of the study? 22. Is there discussion on future studies/new questions?
Content Placement	23. **Is the report organized in three sections (Introduction & Hypothesis, Methods, Results & Discussion)? 24. Is the Introduction & Hypothesis section free of content belonging to the Methods/Results & Discussion? 25. Is the Methods section free of content belonging to the Introduction & Hypothesis/Results & Discussion? 26. Is the Results/Discussion section free of content belonging to the Introduction& Hypothesis/Methods?
Writing	27. ** Are >=90% of the report with clearly and correctly written sentences? 28. ** Are >50% but <90% of the report with clearly and correctly written sentences?
**	Question not included in scale grading criteria t-test analysis.

Table A.6 CPR grading criteria for 2007

Introduction and Hypothesis	1. Is the objective of the inquiry (what one is attempting to investigate) clearly stated? 2. Is the hypothesis logical and clearly presented? 3. Is the hypothesis testable based on data that can be collected from the pictures?
Methods	4. Are the data (variables) collected appropriate for testing the hypothesis? 5. Is the amount of data collected sufficient (30 samples; or the maximum possible if <30, but no less 6. Is the description of the procedure (data collection and analysis) sufficient and clear so others can repeat the study?
Results and Discussion	7. Are the results presented clearly? 8. Are the conclusions stated clearly and adequately, and are based on the results? 9. Are meaningful additional interpretations/discussion presented (further explanations of the results, additional hypothesis on the mechanisms for the pattern tested, new questions raised, limitation of your inquiry, future studies needed, etc.)?
Writing, Clarity and Neatness	10. Is the report logically organized and with clearly and correctly written sentences?
The answer to each question should be Yes, No or To a degree, depending on how well the student met the requirement.	

Table A.7 Demographic characteristics for study samples by year

Gender	2007	2008	2009
Female	35	32	27
Male	54	42	39
College			
Architecture-1	9	10	8
Agriculture and Life Sciences-2	54	35	30
Engineering-3	0	1	0
General Academic Programs-4	18	18	23
Geosciences-5	1	1	0
Liberal Arts-6	6	5	1
Mays Business School-7	1	3	0
Science-8	0	0	0
Veterinary Medicine and Biosciences-9	0	0	4
Ethnicity			
American Indian or Alaskan Native -1	2	0	0
Asian/Pacific Islander/Oriental-2	2	2	1
Black-Non-Hispanic-3	4	1	1
Hispanic-4	13	8	17
White-Non-Hispanic-5	67	62	47
Unknown -6	1	1	0

Table A.8 Inter-rater reliability scores for grading criteria scales

Grading Criteria Scale	Kappa Score
Objective	0.846
Hypothesis	0.500
Sample Methods	0.571
Data Collection Analysis Methods	0.484
Results	1.000
Conclusion	0.5
Discussion	0.711
Content Placement	0.714
Writing	1.000

Table A.9 Pairwise t-tests comparing pre-post CPR scores for 2007 participants

Year	Scale	N	Paired Differences			t-value	P-value
			Mean	95% CI		(2-tailed)	
				Lower	Upper		
2007	Objective	89	0.090	-0.157	0.337	0.722	0.472
2007	Hypothesis	89	0.000	-0.156	0.156	0.000	1.000
2007	Sampling	89	0.202	0.017	0.388	2.165	**0.033
2007	Data Collection/Analysis	89	-0.011	-0.195	0.173	-0.121	0.904
2007	Results	89	-0.101	-0.260	0.058	-1.264	0.209
2007	Conclusion	89	0.056	-0.107	0.219	0.685	0.495
2007	Discussion	89	0.247	0.091	0.404	3.139	**0.002
2007	Content Placement	89	0.135	-0.059	0.328	1.384	0.170

*significant at the level of α=0.10
** significant at the level of α=0.05

Table A.10 Pairwise t-tests comparing pre-post CPR scores for 2008 participants

Year	Scale	N	Paired Differences			t-value (2-tailed)	P-value
			Mean	95% CI			
				Lower	Upper		
2008	Objective	74	0.351	0.071	0.632	2.498	**0.015
2008	Hypothesis	74	0.068	-0.056	0.191	1.093	0.278
2008	Sampling	74	0.378	0.146	0.611	3.246	**0.002
2008	Data Collection/Analysis	74	0.108	-0.017	0.233	1.728	*0.088
2008	Results	74	0.108	-0.038	0.254	1.472	0.145
2008	Conclusion	74	0.149	-0.026	0.323	1.699	*0.094
2008	Discussion	74	0.486	0.294	0.679	5.032	**0.000
2008	Content Placement	74	-0.027	-0.185	0.131	-0.341	0.734

*significant at the level of α=0.10
** significant at the level of α=0.05

Table A.11 Pairwise t-tests comparing pre-post CPR scores for 2009 participants

Year	Scale	N	Paired Differences			t-value (2-tailed)	P-value
			Mean	95% CI			
				Lower	Upper		
2009	Objective	66	0.409	0.101	0.717	2.654	**0.010
2009	Hypothesis	66	0.030	-0.137	0.197	0.363	0.718
2009	Sampling	66	0.242	0.014	0.471	2.120	**0.038
2009	Data Collection/Analysis	66	0.061	-0.154	0.276	0.563	0.576
2009	Results	66	0.061	-0.082	0.203	0.851	0.398
2009	Conclusion	66	0.061	-0.122	0.243	0.664	0.509
2009	Discussion	66	0.227	0.067	0.387	2.834	**0.006
2009	Content Placement	66	-0.091	-0.239	0.057	-1.229	0.223

*significant at the level of α=0.10
** significant at the level of α=0.05

Table A.12 Cohen's d analysis comparisons of pre-post CPR objective, sampling and discussion scores for low, average and high performers.

Performance Level	Objective		Sampling			Discussion		
	2008	2009	2007	2008	2009	2007	2008	2009
Low	0.518	0.750	0.164	0.829	0.971	-0.111	0.560	0.263
Average	0.220	0.244	0.183	0.158	0.430	0.431	0.640	0.331
High	0.261	0.429	0.307	0.366	-0.671	0.302	0.565	0.389

Table A.13 Cohen's d analysis comparisons of pre-post CPR objective, sampling and discussion scores for female and male participants.

Gender	Objective		Sampling			Discussions		
	2008	2009	2007	2008	2009	2007	2008	2009
Female	0.222	0.300	0.168	0.508	0.403	0.442	0.445	0.312
Male	0.370	0.340	0.235	0.292	0.244	0.155	0.668	0.339

Table A.14 Cohen's d analysis comparisons of pre-post CPR objective, sampling and discussion scores for Hispanic and White participants.

Ethnicity	Objective		Sampling			Discussions		
	2008	2009	2007	2008	2009	2007	2008	2009
Hispanic	0.594	0.373	0.283	0.800	0.153	0.318	0.338	0.439
White	0.254	0.327	0.179	0.297	0.362	0.234	0.551	0.298

Table A.15 Cohen's d analysis comparisons of pre-post CPR objective, sampling and discussion scores for majors and non-majors.

College	Objective		Sampling			Discussions		
	2008	2009	2007	2008	2009	2007	2008	2009
Major	0.049	0.353	0.196	0.281	0.403	0.166	0.710	0.300
Non-major	0.522	0.306	0.234	0.488	0.227	0.455	0.473	0.351

APPENDIX B

INQUIRY PROJECT USING THE BEAR CAM PROJECT DIRECTIONS

1. Observation Immersion Experience:

   Study the still pictures of the bears and search for interesting patterns in bear behavior and spatial distribution. Some possible aspects may be: Where are the bears? Are there more bears in some locations than in others? Does the pattern vary with time of day? What are the bears doing and does their behavior vary with time of day or places (rapids, rock, shore, etc.)? Are there other organisms around and do they interact with the bears (when, where and how)? Don't limit yourself to the above questions. Be curious and creative.

   6 days of Bear Web cam pictures are provided in a flash file format. You will need to download the latest version of Adobe Flash Player. It can be downloaded at http://www.adobe.com/products/flashplayer/

   The date when the web cam pictures were taken can be found on the first slide of each flash player file. Each picture is labeled with the time in the following format. For example, "12_57_38 PM_PRST_3" was a picture taken at 12:57 (and 38 seconds) PM.

2. Develop testable hypothesis:

   Formulate your hypothesis in terms of specific predictions of the pattern you believe exists. Here are a few helpful questions to ask yourself: Is the hypothesis structured so that it can be tested and proven right or wrong? Can the hypothesis be tested with data that you can possibly collect from the still pictures, or how can you go about testing the hypothesis using the pictures? Is the hypothesis narrow or focused enough so it is feasible to collect sufficient data in the short period of time you have? Is the hypothesis interesting to you and possibly to others?

   You need to

   (1) post your hypothesis to your discussion group for peer feedback and

   (2) make at least two postings responding to others postings by ...

3. Design and conduct your data collection to test the hypothesis:

   What data do you need to test the hypothesis? Can they be collected from the still pictures? If not, modify the hypothesis or select another one. Design the procedure for collecting the data. For example, tally the number of bears at different locations (rock, rapid, deeper water, shore, etc.) and/or different times of a day; measure the distance to the nearest neighboring bear of females with cubs vs. other bears; etc. You need to collect sufficient data - the rule of thumb is at least 30 samples (stills), but do as many as possible if the picture set does not allow 30 samples for your particular question. If you are measuring distances, a map of the study area on a grid-sheet is provided along with an aerial photo of the study area in separate files. You can use it to record locations of bears and estimate distances between bears.

Record the data, by each still picture (date and time) used, in an Excel or Word file.

Submit your data file in Vista using the assignment tool titled 'Data file submission' before 11 pm on ...

4. Analyze and interpret your data and write report:

   Summarize the data using means (and standard deviations if appropriate) for different location, times, or groups of bears, as appropriate based on your hypothesis.

   Interpret the results (summarized data) and write your research report. Guidelines for the report are provided in a separate file.

   The report is due by 11:00 pm on... Submission procedure (to CPR) will be provided later.

5. Participate in group discussion on Vista

   A minimum of 6 meaningful and descriptive postings is required. *You should have made at least 3 postings by 11:00 pm on Friday… and have made at least 6 postings by 11:00 pm on ….* Please strive to post meaningful and constructive responses and avoid doing them very late so they will be helpful to other students in the group. We will have peer evaluations, by all students in your group, of the quality and timeliness of your postings as part of your grade.

6. Discuss observations and the hypotheses you made on bear behavior, data collection, analysis, and interpretations, and respond to postings of peers from your discussion group. The purpose of this activity is to allow everyone the opportunity to provide and receive peer feedback, which will help each of you to develop a more complete and well thought-out project. Directions for posting to your discussion group are provided on Vista. Make sure to post under the topic for your Inquiry Project Discussion Group, not the Default Topic that would go to the whole class. Directions on how to post are provided in a separate file.

APPENDIX C

Last Updated: October 18, 2010

Grading Criteria for Applying Rubric

1. Is the objective (what one is attempting to investigate) stated?

This question is not evaluating the quality of the objective at all. This question is merely evaluating whether students attempted to include an objective.

Many times students will begin to state the objective by writing: "The objective of this study is…" or "This study is attempting to investigate…" or "The purpose of this investigation…" As a grader you can use these keywords to identify the objective. These keywords will not always be used; and therefore the sentences used to state the objective may be a bit ambiguous.

When keywords are not present you have to be more careful and tough with your grading. Only give a 1 if the student has unmistakably provided description of the investigation taking place.

2. Is the objective clear? No objective=0

How easy did the student make it for you to find the objective? This question is ONLY evaluating whether the student clearly communicated the intention or purpose of their investigation. This question is NOT evaluating specificity nor sentence structure or grammar.

If student does not provide an objective then this question will automatically be given a zero.

Full credit will be given if student clearly stated, "The purpose of investigation," or "The objective of the study," or "I am attempting to investigate…."

At times when reading the introduction it may seem students provided more than one objective and you may be unsure how to grade that. In this circumstance, the student may be attempting to communicate that there are several layers to their objective. See example below:

Student writes: “I am attempting to *investigate* if bears are independent or social animals. The *purpose of this study* is to determine whether bears fish in groups or alone.”

(NOTICE THAT STUDENT USES SEVERAL KEY WORDS (*keywords in italic*)!!)

Rather the student may have stated, “I am attempting to investigate whether bears are social or solitary animals by observing whether bears fish alone or in groups.”

This is just a reminder to not be entirely strict with this question. If you can easily infer what the students’ investigation entails then you may decide to give credit.

3. Is the objective reasonably specific given the study? No objective=0

In order to evaluate this question you may want to review the hypothesis. Reviewing the hypothesis will assist you in determining whether students were too broad in their objective. Many times students will state a general objective such as, “I am attempting to investigate bear behavior at McNeil River Falls.” By reviewing the students’ hypothesis, you can derive the objective is too broad (see below).

“My hypothesis is bears prefer to fish in the rapids rather than from the rock in the middle of river or shore.”

By reviewing the hypothesis you may realize that the student could have been much more specific when stating the purpose of their investigation. Instead of stating, “I am attempting to investigate bear behavior at McNeil River falls,” a better objective would have been, “I am attempting to investigate bears’ fishing habits at McNeil River Falls.

4. **Is a hypothesis presented?**

This question is NOT evaluating quality of the students' hypotheses. It is evaluating how easy the student made it for the reader to find the hypothesis? If student does not provide a hypothesis then this question will automatically be given a zero.

This question is ONLY evaluating whether the student clearly communicated their research question or hypothesis. That is they clearly stated, "My hypothesis," or "My Research Question."

5. **Is hypothesis logical?**

This question is somewhat evaluating the quality of the hypothesis. It is NOT EVALUATING WHETHER THE HYPOTHESIS IS TESTABLE.

Students should not receive full credit if when stating their hypothesis they use obscure language or use terms not clearly defined in the introduction. Consider obscure language to be 1) language/terms that have more than one interpretation/meaning and 2) there is no way of knowing what student meant by using the term. Also, when language/text included does not pertain to 1) explanation or 2) prediction of hypothesis then this question should receive a zero (see below).

Basically a hypothesis should consist of two parts: 1) stating of an observed bear behavior/pattern and 2) presenting a prediction or explanation (should align with pattern observed by student). The student does not necessarily have to present these elements of the hypothesis in one sentence. Your grading can be flexible when evaluating the second part of the hypothesis; that is, the student may receive full credit if just presenting the first part of the hypothesis (presenting a prediction or explanation). In order to receive full credit student needs to have the part one (prediction or explanation of bear behavior).

6. **Is entire hypothesis testable with available data?**

This question is related to whether students considered the limitations of the data they were using. Students have access to 6 days worth of picture stills collected from a web camera set up at McNeil River Falls. Students are provided with the date and time the pictures were taken.

Students at times want to test hypotheses related to the success of bears fishing. This would be impossible to do with picture data because there is no way to obtain counts of fish caught. In addition, another common un-testable hypothesis used by students is an indirect attempt to test the abundance of fish at a location. For example, "My hypothesis is that bears prefer the rapids because there is more fish than near the shore." Another related hypothesis is, "My hypothesis is that bears prefer to fish in the rapids because it is easier to catch fish than near the shore or on the rock." Obviously, students cannot test the ease of catching fish or how abundant the fish are with the available data. Part of their hypothesis would be testable. Students could count the numbers of bears at each location to see if there is a preference for fishing (see next question).

7. **Is part of the hypothesis testable with available data?**

See explanation and notes for question above. Students will automatically receive a 1 if they received a one for item 6.

8. Is the number of samples (pictures) reported?

GIVE ZERO for vagueness. For example, a student may state, "I went through all the data available to find pictures where all the falls were shown in the sample." In this example although the student went through ALL the pictures they failed to mention how many pictures they found showing all the falls (number of sample). The only time the word all is acceptable for reporting sample size is for example, I used all the pictures on July 7, 2005 (very specific).
Students should also not receive credit if they report observations rather than sample size. For example, some students may state, "I collected pictures with cubs in them, such as on July 7th at 11 am, and July 3 at 3 pm."

Ask, do students state how much data they collected? For example, the student may state, "I collected 50 bear pictures for my analysis." Sometimes students will say, "I used all the pictures from July 7, 2005." Assume in this case that the student used 100 percent of the pictures provided for that day. You can give full credit considering that they are indirectly telling you that they used all 86 pictures provided for July 7, 2005. Below are the picture counts for each day. You can also give full credit if a student states I randomly chose 5 pictures from each day (5 random pictures X 6 days = 30 pictures used for sample). Again, the student is indirectly reporting how much data was used. **Also, if the student provides a range you can give them credit as well. If students give a range, however they will not receive credit for question 10.**

7/7/2005 – 86 pictures
7/8/2006 – 65 pictures
7/9/2005 – 34 pictures
7/11/2005 – 177 pictures

712/2005 – 215 pictures
7/13/2005 – 204 pictures

9. Is the number of samples sufficient (>=30 samples; or the maximum available if <30, but no less than 5)? If student does not specify number of samples used give a zero.

Students were required to use a sample of AT LEAST 30 pictures OR THE MAXIMUM POSSIBLE (sample included all pictures with specific attributes). Again, you can give full credit if the student indirectly tells you how much data was collected. For example, "I randomly chose 10 pictures from each day" or "I used ALL the pictures with bear cubs (specific attribute) in them." Although you yourself do not know how many bear cub pictures are in the complete dataset of 781 pictures you can assume they used the MAXIMUM AMOUNT OF PICTURE SAMPLES AVAILABLE. HOWEVER, DO NOT IGNORE THE STIPULATION THAT STUDENTS NEEDED AT LEAST 5 PICTURES IF THEY ATTEMPTED TO USE THE MAXIMUM POSSIBLE. You may have to look at the results to ascertain this. BE CAREFUL! If students say in their results they did not have enough pictures to do an analysis then ASSUME THEY DID NOT MEET THE AT LEAST 5 PICTURE REQUIREMENT (unless otherwise stated by student) and give them a zero for this rubric item. Keep in mind that this is not the same as the student stating, "I did not have enough data to disprove or support my hypothesis."
7/7/2005 – 86 pictures
7/8/2006 – 65 pictures
7/9/2005 – 34 pictures
7/11/2005 – 177 pictures
712/2005 – 215 pictures
7/13/2005 – 204 pictures

When students collect less than 30 picture samples and have written that they only reviewed certain days of pictures excluding some of the data made available they should not receive credit for this item. The logic behind this is that they had more data they could have reviewed to see if it matched criteria for sample selection.

10. **Is there sufficient description for sample selection (how)? Give credit if repeatable.**

Students need to provide a clear step by step sequential description of how sample was collected. If the description is confusing or contradictory in any manner you can give a zero in this category. For example, students may state, "I collected 30 random samples from July 7, 2010 that included bears eating and hunting." This statement is a bit contradictory in that it suggests the student first selected pictures with only bears eating and hunting and then randomly chose 30 from those pictures found of bears eating and hunting.

This question is evaluating repeatability of the methods section. You should ask can I repeat the methods to re-test the student's hypothesis. If there is any doubt then the student should not receive full credit for this rubric item.

NO CREDIT: IF STUDENTS GIVE A RANGE FOR SAMPLE (# OF PICTURES USED) THEY WILL AUTOMATICALLY RECEIVE ZERO FOR THIS CATEGORY!

Below are some common examples of when a student DOES NOT receives full credit:

NO CREDIT: I randomly collected 5 pictures from 3 days of available picture stills. (Student failed to mention which days were used.)

Full CREDIT: I randomly collected 5 pictures from June 7, 8, and 9, 2005.

NO CREDIT: I used 30 pictures to test my hypothesis. (Simply stating the number used is not sufficient to get credit for this item.)

FULL CREDIT: I used only the first 30 pictures on 7/9/2005. (Although this is not good practice we are only evaluating repeatability).

11. **Are the data (variables) collected appropriate for testing the hypothesis (the portion that is testable)?**

    IF STUDENTS FAIL TO REPORT ALL THE VARIABLES NECESSARY TO TEST THEIR HYPOTHESIS (THE PORTION THAT IS TESTABLE) THEY WILL AUTOMATICALLY RECEIVE ZERO.

    You must review the hypothesis mentioned in the Introduction. Ask whether the student is collecting variables directly related to the testing of the hypothesis in the introduction? Many times students will collect many variables even ones unrelated to their hypotheses. You can ignore the students' excess data collection IF they collect data for variables related to testing of their hypothesis.

12. **Is there sufficient description for data collection (variables and how collected)? Give credit if repeatable**.

    WE ARE EVALUATING THE METHODS SECTION ONLY. If students fail to provide a description of variables collected in the methods we will not evaluate their results in the discussion section of their paper to infer what variables were collected (i.e. The average of bears on the falls was 11, the average of bears on shore was 5, etc). That is, we can infer students counted

the number of bears per picture (var 1) located in the falls (var 2) and then the shore (var 3). However, no credit will be given because student failed to explicitly provide a description of how variables were collected in the methods section.

If student provided a description of variables collected in the results/discussion section we will give credit for question 12 but assume they misplaced part of the methods in the results/discussion section.

Similar to the last question we are evaluating repeatability. If student has ambiguous variables or does not provide a good enough description of data they collected from the pictures then a zero is given. See example below:

NO CREDIT: "I counted the number of bears on the rock location in the morning, mid-day and afternoon."

FULL CREDIT: I counted the number of bears on the rock location in the early morning (6 am-11 am), midday (11 am -3pm) and afternoon (3pm – 9pm).

13. **Is there sufficient description for data analysis (i.e. frequency, count, average, etc)?**

WE ARE EVALUATING THE METHODS SECTION ONLY. Again, like question 12 if students fail to provide a sufficient description of data analysis in the methods we will not evaluate their results in the discussion section to infer what exact computations/analysis was performed (frequency, count, average). Elaborating on the example for question 12, "The average of bears on the falls was 11, the average of bears on shore was 5, etc." Thus, we can infer students counted the number of bears per picture (var 1) located in the falls (var 2) and then the shore (var 3) and then took an

average of those counts per picture. However, no credit will be given because student failed to explicitly provide a description of data analysis in the methods section. READ BOLD PRINT NEXT PAGE.

If student provided a description of analysis performed in the results/discussion section we will give credit for question 13 but assume they misplaced part of the methods in the results/discussion section.

Similar to the last two questions we are again evaluating repeatability. Can we repeat the student's data analysis? See example:

FULL CREDIT:" I averaged, "or "I summed," or "I calculated."

14. Are results presented?

With this question, students are not expected to quantify results (see next question). They may describe trends in their data rather than provided percentages, frequency, etc.

FULL CREDIT: "A majority of bears were observed fishing from the rock compared to very few on the shore and none in the rapids." This describes very well a trend observed in the data.

FULL CREDIT: All bears were observed fishing from the rock compared to the shore. When students use the word ALL, interpret this as all the cases observed (100%).

15. Are results specific (sums, averages, ratios)? No results=0

With this question we are specifically evaluating whether students quantified the trends observed in their data. The only exception to this is again when students use the word ALL. We should interpret this as 100%.

NO CREDIT: A majority of the bears were observed fishing from the rock compared to very few on the shore and none in the rapids." Although in this example students allude to zero observations of bears fishing in the rapids they fail to quantify what constituted a majority on the rock and very few on the shore.

FULL CREDIT: 95 percent of bears were observed from the rock compared to 5 percent total for all other locations observed.

16. Do the results correspond to the variables specified in the Methods section? No results=0

Simply ask, "Are the results presented related to the variables outlined in the Methods section." Sometimes students will provide a laundry list of variables in the Methods section and will only mention a few in the results. They should receive full credit if the results are related to variables in the laundry list. IF STUDENTS FAIL TO PROVIDE RESULTS FOR ALL VARIALBES THEY CAN STILL GET FULL CREDIT.

17. Are conclusions presented?

This question is not evaluating the quality or correctness of the conclusion. This question is merely evaluating whether the student drew any conclusion(s) from conducting their study. IF STUDENT did NOT provide results you can still evaluate whether they concluded ANYTHING from their study.

In addition, examples of students making conclusions can include: student stated whether the results support or disprove their hypothesis, or if analysis/findings provided any clear evidence for testing their hypothesis.

Keywords students use are, In conclusion or sometimes students will incorrectly state, I proved my hypothesis true or incorrect. These keywords will not always be used. This item is NOT evaluating whether students interpreted their results (see question 20).

Only evaluate whether the student considers how their findings/results provide a better understanding of the hypothesis/question/problem under review.

18. Are the conclusions based solely on the results? No conclusion=0; conclusion w/out results=0

If the students provide conclusions NOT SUPPORTED by their findings/results then they will not receive full credit.

We see examples of this when students are attempting to test the following hypothesis, "My hypothesis is that bears prefer to fish from the rock rather than any other location in the river because the fish are easier to catch." At times students collect data showing that the rock may be the preferred location for fishing; and therefore improperly state the fish are more abundant in that location. Whether or not the fish are abundant is un-testable and therefore if students state that fish are more abundant in that location then they WILL NOT RECEIVE FULL CREDIT FOR THIS ITEM.

19. Do the conclusions correspond to the hypothesis?

This rubric item is assessing only if conclusion is related to the hypothesis originally mentioned in the introduction. Many times because the hypothesis is not testable students will collect data and conclude something totally unrelated to their hypothesis. These students should not receive full credit for this item.

HOWEVER, students can receive full credit if the conclusion is related to the part of their hypothesis that was testable! STUDENTS MAY ALSO RECEIVE CREDIT IF THEY MAKE CONCLUSIONS RELATED TO PART OF THE HYPOTHESIS THAT WAS UNTESTABLE.

20. Is there interpretation (possible/hypothesized mechanisms or explanations) of the results?

Evaluate whether the student provides an explanation of what contributed to the trends in the findings or results.

21. Is there discussion on limitations of the study?

Easy to evaluate, do students mention any confounding factors that may have skewed their findings/results or made it difficult to test their hypothesis.

If students say there are no limitations they will automatically get a zero for this item!

22. Is there discussion on future studies/new questions?

Do students mention any future studies/new questions raised or related to their findings/results. You may give full credit if student mentions future studies/new questions UNRELATED to their original hypothesis or findings.

23. Is the report organized in three sections (Introduction & Hypothesis, Methods, and Results & Discussions)? If students receive zero for item 23, they will also receive zero for 24, 25 and 26.

Do students have three sections? Sometimes students failed to provide new paragraphs and spaces between sections. DO NOT TAKE OFF CREDIT IF THERE ARE NOT HARD RETURNS AFTER EACH SECTION THIS COULD BE A PROBLEM ASSOCIATED WITH UPLOADING OR DOWNLOADING

OF REPORTS. If report is presented in one big paragraph but the content reflects three sections you may give full credit for this rubric item.

24. Is the Introduction & Hypothesis section free of content belonging to the Methods/Results & Discussion?

This question is ONLY evaluating if the student included material in the *Introduction* that belongs in the *Methods* or *Results & Discussion* section of their report.

If students mention findings/results (*Results & Discussion*) related to their hypothesis in the *Introduction* a zero for this item will be given.

For example, at times after students state their hypothesis in their *Introduction* they may follow with a sentence stating, "My analysis found this hypothesis to be true." In this case students would NOT RECEIVE CREDIT for this item. Evidence of whether their hypothesis was supported or disproved should only be mentioned in the *Results & Discussion* section of reports.

25. Is the Methods section free of content belonging to the Introduction & Hypothesis/Results & Discussion?

This question is ONLY evaluating if the student included material/content in the *Methods* that belongs in the *Introduction* or *Results & Discussion* sections of their report.

For example, if a student presents their results (averages, sums or ratios) in the *Methods* and not in the *Results/Discussion* then a ZERO will be given for this rubric item.

26. Is the Results/Discussion section free of contents belonging to the Introduction & Hypothesis/Methods?

If a student includes any material or content that belongs in the *Methods* section then they will not receive credit for this item. For example, if students mention in the *Results/Discussion* how they chose to test their hypothesis (*METHODS*) they would get a zero for this rubric item.

27. Are >=90% of the report with clearly and correctly written sentences?

This is only evaluating whether a majority of the sentences were clearly and correctly written. Again, if students have problems writing more than likely it will be apparent and you should keep track of unclear or problematic sentences. If students receive full credit for this item they will automatically receive a zero for Question 28. The logic behind this is that if 90% of the report was written with clear and correct sentences then this cannot mean 50% but <90%.

28. Are >50% but <90% of the report with clearly and correctly written sentences?

You are evaluating one page reports so you will have to keep track of how much information is not clearly and correctly written within the report while reading. An independent judgment will have to be made whether 50-90 percent of report was written with clear and correct sentences. Sometimes it is helpful to highlight sentences with a unique pen/highlighter to help you make a decision. Students CAN NOT receive a 1 (full credit) for both Question 27 and Question 28.

APPENDIX D

Table D.1 Objective scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (33.33%)	Improved	Met requirement (0-1)	3.70%
		Met requirement to a degree (0.5-1)	7.41%
		Met requirement to a degree (0-0.5)	14.81%
	No change	Did not meet requirement (0)	7.41%
		Met requirement (1)	0.00%
		Met requirement to a degree (0.5)	0.00%
	Declined	Requirement met to a degree (1-0.5)	0.00%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (66.67%)	No change	Met requirement (1)	25.93%
		Met Requirement to a degree (0.5)	25.93%
		Did not meet requirement (0)	14.81%

Table D.2 Objective grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q1	Q2	Q3
Made Revision	Improvement in score	Met requirement (0-1)	18.52%	18.52%	11.11%
	Score did not change	Did not meet requirement (0)	7.41%	7.41%	11.11%
		Met requirement (1)	3.70%	0.00%	3.70%
	Decline in score	Requirement not met (1-0)	0.00%	0.00%	0.00%
No Revision	CPR post score	Met requirement (1)	55.56%	48.15%	33.33%
		Did not meet requirement (0)	14.81%	25.93%	40.74%

Q1: Is the objective (what one is attempting to investigate) clearly stated?
Q2: Is the objective clear? No objective = 0
Q3: Is the objective reasonably specific given the study?

Table D.3 Hypothesis scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (40.74%)	Improved	Met requirement (0-1)	0.00%
		Met requirement to a degree (0.5-1)	3.70%
		Met requirement to a degree (0-0.5)	3.70%
	No change	Did not meet requirement (0)	0.00%
		Met requirement (1)	11.11%
		Met requirement to a degree (0.5)	18.52%
	Declined	Requirement met to a degree (1-0.5)	3.70%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (59.26%)	No change	Met requirement (1)	37.04%
		Met Requirement to a degree (0.5)	22.22%
		Did not meet requirement (0)	0.00%

Table D.4 Hypothesis grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)			
			Q4	Q5	Q6	Q7
Made Revision	Improvement in score	Met requirement (0-1)	3.70%	11.11%	3.70%	3.70%
	Score did not change	Did not meet requirement (0)	0.00%	0.00%	7.41%	3.70%
		Met requirement (1)	37.04%	7.41%	3.70%	7.41%
	Decline in score	Requirement not met (1-0)	0.00%	3.70%	0.00%	0.00%
No Revision	CPR post score	Met requirement (1)	59.26%	55.56%	59.26%	77.78%
		Did not meet requirement (0)	0.00%	22.22%	25.93%	7.41%

Q4: Is hypothesis presented?
Q5: Is the hypothesis logical?
Q6: Is entire hypothesis testable with available data?
Q7: Is part of the hypothesis testable with available data?

Table D.5 Sampling scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (37.04%)	Improved	Met requirement (0-1)	3.70%
		Met requirement to a degree (0.5-1)	0.00%
		Met requirement to a degree (0-0.5)	7.41%
	No change	Did not meet requirement (0)	0.00%
		Met requirement (1)	7.41%
		Met requirement to a degree (0.5)	18.52%
	Declined	Requirement met to a degree (1-0.5)	0.00%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (62.96%)	No change	Met requirement (1)	18.52%
		Met Requirement to a degree (0.5)	33.33%
		Did not meet requirement (0)	11.11%

Table D.6 Sampling grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q8	Q9	Q10
Made Revision	Improvement in score	Met requirement (0-1)	18.52%	18.52%	11.11%
	Score did not change	Did not meet requirement (0)	0.00%	0.00%	14.81%
		Met requirement (1)	7.41%	7.41%	0.00%
	Decline in score	Requirement not met (1-0)	0.00%	0.00%	0.00%
No Revision	CPR post score	Met requirement (1)	59.26%	51.85%	25.93%
		Did not meet requirement (0)	14.81%	22.22%	48.15%

Q 8: Is the number of samples (pictures) reported?
Q9: Is the number of samples sufficient (>=30; or the maximum available if <30, but no less than 5)?
Q10: Is there sufficient description for sample selection (how)?

Table D.7 Data collection/analysis scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (33.33%)	Improved	Met requirement (0-1)	3.70%
		Met requirement to a degree (0.5-1)	0.00%
		Met requirement to a degree (0-0.5)	7.41%
	No change	Did not meet requirement (0)	3.70%
		Met requirement (1)	3.70%
		Met requirement to a degree (0.5)	11.11%
	Declined	Requirement met to a degree (1-0.5)	3.70%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (66.67%)	No change	Met requirement (1)	22.22%
		Met Requirement to a degree (0.5)	33.33%
		Did not meet requirement (0)	11.11%

Table D.8 Data collection/analysis grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q11	Q12	Q13
Made Revision	Improvement in score	Met requirement (0-1)	3.70%	3.70%	11.11%
	Score did not change	Did not meet requirement (0)	14.81%	14.81%	7.41%
		Met requirement (1)	3.70%	7.41%	3.70%
	Decline in score	Requirement not met (1-0)	0.00%	0.00%	0.00%
No Revision	CPR post score	Met requirement (1)	55.56%	40.74%	37.04%
		Did not meet requirement (0)	22.22%	33.33%	40.74%
Q11: Are the data (variables) collected appropriate for testing the hypothesis? *Q12: Is there sufficient description for data collection (variables and how collected)?* *Q13: Is there sufficient description for data analysis (i.e. frequency, count, average, etc)?*					

Table D.9 Results scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (22.22%)	Improved	Met requirement (0-1)	0.00%
		Met requirement to a degree (0.5-1)	11.11%
		Met requirement to a degree (0-0.5)	3.70%
	No change	Did not meet requirement (0)	0.00%
		Met requirement (1)	3.70%
		Met requirement to a degree (0.5)	3.70%
	Declined	Requirement met to a degree (1-0.5)	0.00%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (77.78%)	No change	Met requirement (1)	55.56%
		Met Requirement to a degree (0.5)	7.41%
		Did not meet requirement (0)	14.81%

Table D.10 Results grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q14	Q15	Q16
Made Revision	Improvement in score	Met requirement (0-1)	3.70%	11.11%	3.70%
	Score did not change	Did not meet requirement (0)	0.00%	0.00%	3.70%
		Met requirement (1)	14.81%	11.11%	7.41%
	Decline in score	Requirement not met (1-0)	0.00%	0.00%	0.00%
No Revision	CPR post score	Met requirement (1)	66.67%	59.26%	66.67%
		Did not meet requirement (0)	14.81%	18.52%	18.52%
Q14: Are results presented? *Q15: Are results specific (sums, averages, ratios)?* *Q16: Do results correspond to the variables specified in the Methods section?* *No results = 0*					

Table D.11 Conclusion scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (22.22%)	Improved	Met requirement (0-1)	0.00%
		Met requirement to a degree (0.5-1)	7.41%
		Met requirement to a degree (0-0.5)	0.00%
	No change	Did not meet requirement (0)	3.70%
		Met requirement (1)	7.41%
		Met requirement to a degree (0.5)	0.00%
	Declined	Requirement met to a degree (1-0.5)	3.70%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (77.78%)	No change	Met requirement (1)	44.44%
		Met Requirement to a degree (0.5)	29.63%
		Did not meet requirement (0)	3.70%

Table D.12 Conclusion grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q17	Q18	Q19
Made Revision	Improvement in score	Met requirement (0-1)	0.00%	7.41%	0.00%
	Score did not change	Did not meet requirement (0)	3.70%	3.70%	3.70%
		Met requirement (1)	14.81%	7.41%	14.81%
	Decline in score	Requirement not met (1-0)	0.00%	3.70%	0.00%
No Revision	CPR post score	Met requirement (1)	77.78%	44.44%	74.07%
		Did not meet requirement (0)	3.70%	33.33%	7.41%

Q17: Are conclusions presented?
Q18: Are the conclusions based solely on the results?
Q19: Do the conclusions correspond to the hypothesis?

Table D.13 Discussion scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (40.74%)	Improved	Met requirement (0-1)	7.41%
		Met requirement to a degree (0.5-1)	11.11%
		Met requirement to a degree (0-0.5)	0.00%
	No change	Did not meet requirement (0)	0.00%
		Met requirement (1)	0.00%
		Met requirement to a degree (0.5)	18.52%
	Declined	Requirement met to a degree (1-0.5)	3.70%
		Requirement not met (0.5-0)	0.00%
		Requirement not met (1-0)	0.00%
No Revision (59.26%)	No change	Met requirement (1)	22.22%
		Met Requirement to a degree (0.5)	37.04%
		Did not meet requirement (0)	0.00%

Table D.14 Discussion grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)		
			Q20	Q21	Q22
Made Revision	Improvement in score	Met requirement (0-1)	7.41%	11.11%	25.93%
	Score did not change	Did not meet requirement (0)	3.70%	7.41%	0.00%
		Met requirement (1)	3.70%	3.70%	0.00%
	Decline in score	Requirement not met (1-0)	7.41%	0.00%	3.70%
No Revision	CPR post score	Met requirement (1)	55.56%	51.85%	51.85%
		Did not meet requirement (0)	22.22%	25.93%	18.52%
Q20: Is there interpretation (possible/hypothesized mechanisms or explanations) of the results? *Q21: Is there discussion on limitations of the study?* *Q22: Is there discussion on future studies/new questions?*					

Table D.15 Content placement scale text change frequency and impact to ecological inquiry report scores for all groups and individual groups (low, average and high).

Text Revision	Impact on Score		% Student (N=27)
Made Revision (14.81%)	Improved	Met requirement (0-1)	0.00%
		Met requirement to a degree (0.5-1)	3.70%
		Met requirement to a degree (0-0.5)	0.00%
	No change	Did not meet requirement (0)	0.00%
		Met requirement (1)	0.00%
		Met requirement to a degree (0.5)	3.70%
	Declined	Requirement met to a degree (1-0.5)	0.00%
		Requirement not met (0.5-0)	3.70%
		Requirement not met (1-0)	3.70%
No Revision (85.19%)	No change	Met requirement (1)	59.26%
		Met Requirement to a degree (0.5)	25.93%
		Did not meet requirement (0)	0.00%

Table D.16 Content Placement grading criteria text change frequency and impact to ecological inquiry report scores

Text Revision	Impact on Score		% Students (N=27)			
			Q23	Q24	Q25	Q26
Made Revision	Improvement in score	Met requirement (0-1)	0.00%	7.41%	0.00%	0.00%
	Score did not change	Did not meet requirement (0)	0.00%	0.00%	3.70%	0.00%
		Met requirement (1)	7.41%	0.00%	3.70%	3.70%
	Decline in score	Requirement not met (1-0)	3.70%	3.70%	7.41%	3.70%
No Revision	CPR post score	Met requirement (1)	88.89%	70.37%	77.78%	92.59%
		Did not meet requirement (0)	0.00%	18.52%	7.41%	0.00%

Q23: Is the report organized in three sections (Introduction & Hypothesis, Methods, Results & Discussion)?
Q24: Is the Introduction & Hypothesis section free of content belonging to the Methods/Results & Discussion?
Q25: Is the Methods section free of content belonging to the Introduction & Hypothesis/Results & Discussion?
Q26: Is the Results/Discussion section free of content belonging to the Introduction& Hypothesis/Methods?

Table D.17 Overall grades for pre and post CPR sample reports

Group	N	Pre	Range	SD	Post	Range	SD
Low	9	61.97%	30.77-84.62%	0.179	67.95%	23.08-92.31%	0.233
Avg	9	64.53%	50.00%-84.62%	0.100	70.94%	57.69-92.31%	0.119
High	9	81.62%	65.38-92.31%	0.098	84.62%	61.54-100.00%	0.133

VITA

Denise Celeste Robledo received her Bachelor of Arts degree in Telecommunication Media Studies from the Department of Journalism at Texas A&M University in 2002. Following graduation, she was accepted into the Department of Educational Psychology and was awarded a Masters of Education degree in Educational Technology in August 2006. She began pursuing her Ph.D. in Rangeland Ecology and Management in August of 2006 under the advisement of Dr. X. Ben Wu and graduated in May 2011. Her research interests include ecology education, science inquiry, scientific writing, distance education learning environments, instructional technology, curriculum design and assessment of online learning environments.

Ms. Denise Celeste Robledo may be reached at 2138 TAMU, College Station, TX 77840-2138. Her email is denise.c.robledo@gmail.com.