CRESST REPORT 822

THE IMPACT OF SHORT-TERM SCIENCE TEACHER PROFESSIONAL DEVELOPMENT ON THE EVALUATION OF STUDENT UNDERSTANDING AND ERRORS RELATED TO NATURAL SELECTION

OCTOBER, 2012

Rebecca E. Buschang

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

The Impact of Short-Term Science Teacher Professional Development on the Evaluation of Student Understanding and Errors Related to Natural Selection

CRESST Report 822

Rebecca E. Buschang
CRESST/University of California, Los Angeles

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by grant number OPP1003019 from the Bill and Melinda Gates Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the Bill and Melinda Gates Foundation.

TABLE OF CONTENTS

THE IMPACT OF SHORT-TERM SCIENCE TEACHER PROFESSIONAL DEVELOPMENT ON THE EVALUATION OF STUDENT UNDERSTANDING AND ERRORS RELATED TO NATURAL SELECTION

Rebecca E. Buschang
CRESST/University of California, Los Angeles

Abstract

This study evaluated the effects of a short-term professional development session. Forty volunteer high school biology teachers were randomly assigned to one of two professional development conditions: (a) developing deep content knowledge (i.e., control condition) or (b) evaluating student errors and understanding in writing samples (i.e., experimental condition). A pretest of content knowledge was administered, and then the participants in both conditions watched two hours of online videos about natural selection and attended different types of professional development sessions lasting four hours. Significant differences between conditions in favor of the experimental condition were found on participant identification of critical elements of student understanding of natural selection and content knowledge related to natural selection. Results suggest that short-term professional development sessions focused on evaluating student errors and understanding can be effective at focusing a participant's evaluation of student work on particularly important elements of student understanding. Results have implications for understanding the types of knowledge necessary to effectively evaluate student work and for the design of professional development.

Background and Problem Statement

In 2001, Congress passed the No Child Left Behind Act of 2001 (NCLB). This Act required states to develop assessments of basic skills in mathematics, science, and English and administer these tests annually to students in particular grade levels for schools to receive federal funding. Additionally, certain schools designated as Title I schools (more than 40% of students receive free or reduced lunches) had to show gains in performance from year to year on the standardized tests and make Adequate Yearly Progress (AYP) to receive additional federal funding. Schools not meeting the AYP goals several years in a row were required to show evidence that improvements were being made and might also be subjected to a complete restructuring of the school or an additional reduction in funding.

More recently, the Common Core State Standards (CCSS), a set of math and English language arts standards that include reading and writing standards for Literacy in Science and Technical Subjects, have been adopted by 45 states. A national consortium of 27 states, including California, are working together as part of the SMARTER Balanced Assessment Consortium (SBAC) to develop an assessment system aligned to the new Common Core standards. An additional set of states is working with Achieve as part of the Partnership for the Assessment of Readiness for College and Careers (PARCC or Partnership) to develop an assessment program with a similar purpose to SBAC. These systems will assess students both summatively and formatively using computer-adaptive assessments and performance tasks. These types of summative assessments will replace current state tests, but will still be administered annually under current NCLB rules. Formative assessments and tools that align with the CCSS are also being developed to help teachers gather information about student progress towards learning goals. These interim formative assessments will require teachers to develop and score constructed responses and performance tasks as well as to interpret results and to determine appropriate next instructional steps. Additionally, new national K-12 content standards for science, the Next Generation Science Standards (NGSS), have been released in draft form and are expected to be considered for adoption in the fall of 2012. New state consortia-built assessments will follow.

This shift in the use of assessment data in schools, emphasizing formative assessments, has already changed the practices of teachers. Many teachers already use state assessment data, which will be replaced with summative CCSS data, as a final evaluation of students' level of competence. Summative information is not expected to remedy students' specific problems, but might instead be used to modify course plans for the following year.

Formative assessment systems such as the CCSS formative assessment system will allow teachers to gather interim performance information from students and analyze their progress to determine and execute the next, appropriate instructional steps. For example, if a student's response reflects a lack of understanding of one subcomponent of natural selection, helping the student to understand the knowledge related to that specified subcomponent should be the aim of reteaching or relearning. Or if a certain misconception is prevalent in the interim data collected, teachers can address those specific misconceptions with one or more students in ensuing instruction. Formative assessments provide an occasion to monitor student learning.

Being able to use student performance effectively in a formative way depends on the teacher's ability to diagnose student performance. Diagnosing requires a deep understanding of the content area by the teacher (Black & Wiliam, 1998; Heritage, 2007; Sadler, 1989; Shepard, 2005). For science teachers, to know the components making up the deep understanding of the content area means they must update their scientific knowledge with current information and processes because scientific knowledge is changing at such a rapid rate.

In addition to a deep understanding of the content, diagnosing student performance requires teachers to develop highly specialized knowledge related to teaching a particular subject, also called pedagogical content knowledge (Black & Wiliam, 1998; Heritage, 2007; Sadler, 1989; Shepard, 2005). Pedagogical content knowledge helps instructors identify performance elements such as specific misconceptions or misunderstandings students at a particular age or level of learning might hold. These elements help teachers analyze and assess both student understanding and errors. Moreover, good diagnoses imply a high degree of understanding of individual differences, as different students will have varied sets of errors or gaps to address.

Using information formatively to diagnose student understanding and errors also requires teachers to be able to differentiate between serious misconceptions and minor mistakes. For example, if a biology student incorrectly believes that natural selection follows a preordained path that always benefits the organism, then the student has a major misunderstanding of natural selection. On the other hand, if a student uses the word "strong" rather than "beneficial" when referring to traits, this is still an error to be dealt with, but less serious than the previous example. Teachers need to have the knowledge and necessary skill to distinguish among these errors.

Teachers are increasingly expected to diagnose student understanding and errors as part of instructional practices under the current standards-based educational reform movement (Gallagher & Worth, 2008; Pellegrino, Chudowsky, & Glaser, 2001). They rarely receive formal training on these types of assessment practices during their teacher preparation program (Stiggins, 1999). Teachers either develop these skills on their own or through professional development training. However, little is known about the best ways to develop skills related to diagnosing student performance, in part because research on this topic has been limited. Of the research literature that exists on professional development related to formative assessment, most studies measure student achievement data as the outcome measure and not the intervening teacher outcomes. In comparison to the number of studies focused on student outcomes, very few studies examine the impact of professional development on related teacher outcomes, and no peer-reviewed studies have been published that focus on the development of formative assessment skills in specific subject areas (Schneider & Randel, 2010). Therefore, few inferences can be made about the impact of these types of professional development on teacher knowledge and the most effective ways to develop desired teacher knowledge.

Moreover, the research on professional development suggests that for professional development to be effective, it should be long term (Garet, Porter, Desimone, Birman, & Yoon, 2001; Yoon, Duncan, Lee, Scarlos, & Shapley, 2007). However, time and money are increasingly limited in schools because of limited state funds (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Hill, 2009). Alternatives such as short-term professional development must be considered.

Therefore, the aim of this study was to examine the impact of short-term professional development programs focused on different factors hypothesized to affect the quality of teacher evaluation of student understanding and errors. The main research question examined differences between conditions on teacher outcome measures related to their evaluation of errors and understanding in common samples of student written work. Volunteer teachers were randomly assigned to receive one session of professional development concentrated on either (a) developing deep content knowledge (i.e., control condition) or (b) analyzing and assessing student errors and understanding in writing samples (i.e., experimental condition). Within the field of biology, the topic for this study was natural selection. Teachers were administered a pretest of natural selection. Teachers in both conditions then watched two hours of online videos about natural selection before their professional development session to set a baseline of knowledge about natural selection. The main outcome measure asked teachers to diagnose and rate the same four samples of student writing. Teachers were also administered a posttest of content knowledge related to natural selection, a demographic survey, and a survey about their professional development experience.

This study begins to fill a gap in the research literature by examining the types of knowledge that are important for teachers when making evaluations of student understanding and errors. In addition, few studies of professional development have used the methodology and outcome measures used in this study, including (a) randomized experimental methods to infer causal relationships and (b) examining teacher outcomes rather than only relying on teacher surveys or broad student outcomes to measure effectiveness of professional development. Finally, this study investigates whether short-term professional development opportunities for teachers can be effective in evaluating student work.

Literature Review

This section describes the scholarly literature relevant to this study. First, it considers the topic of formative assessment. Next, two types of teacher knowledge are reviewed that are thought to impact formative assessment skills, including content knowledge and pedagogical content knowledge. In addition, the relationships among each aspect of teacher knowledge needed to diagnose student understanding and errors are discussed. The literature relevant to effective characteristics of professional development is then reviewed. These findings are important because they are the characteristics that guided the design of the professional development program in this study. Next, the literature related to professional development of formative assessment skills is summarized. Limitations of these studies, as well as how the proposed study provides a unique contribution to the literature, are discussed. Finally, the literature related to the topic of the study context, natural selection, is outlined, including prior research related to student misconceptions.

Formative Assessment

Formative assessments are typically defined as midstream evaluations where evidence is collected from students and used to inform future instruction (Black & William, 1998; Sadler, 1989; Shepard, 2005). Examples of formative assessment include using student responses during a class discussion to change the upcoming instruction, or using student answers on a quiz to determine what topics(s) need reteaching. These types of activities differ from many typical classroom assessments used to evaluate students within a specified body of knowledge (e.g., end-of-chapter tests or quizzes, project). These assessments are called summative, suggesting they are at the end of an instructional sequence, and the topics of assessments are not revisited for instruction.

Formative assessments vary with respect to the type and amount of student performance information collected by teachers, the amount of planning, the formality of the assessment, and the nature of the feedback (Bell & Cowie, 2001; Shavelson et al., 2008). Assessments that are unplanned or are “on-the-spot” typically occur during “teachable moments” in whole class discussion or perhaps in individual teacher-student interactions. In these interactions, teachers may notice that students do not fully understand a topic and thus change their teaching spontaneously to address any perceived gaps in understanding. In contrast, formal preplanned formative assessments typically involve assessments matched to learning objectives, carrying out a lesson or lessons, administering or reviewing student work, interpreting the results of the assessment, and, finally, using the information from the assessment to refine or create new learning goals and instruction.

This study focused on the specific formative assessment practices of evaluating student understanding and identifying their errors related to natural selection. In science, the identification of student errors also requires teachers to identify the preconceptions or misconceptions students may hold about a particular topic.

Effect sizes for formative assessment practices have been estimated to be between 0.40 and 0.70, indicating a moderate to large positive impact (Black & Wiliam, 1998). However, a more recent meta-analysis suggests that effect sizes for formative assessments may be closer to 0.25. The authors cited that most of the research reviewed was flawed with uninterpretable results (Kingston & Nash, 2011). Criteria for inclusion in the revisited meta-analysis included the use of a control or comparison group, the inclusion of appropriate statistics to calculate effect sizes, a publication date after 1988, and a focus on K-12 students. Of the 300 studies reviewed, only 13 met the criteria for inclusion. Of those, the largest effect size (d = .30) was found for formative assessment studies related to professional development.

While results of this meta-analysis indicate that formative assessment practices have a smaller effect than typically reported in the literature, the author suggests that their practical significance is still important. The author further suggests that an estimated effect size of 0.30 indicates that if 20% of students are currently at or above a proficient level on standardized state tests, then formative assessment practices would lead to a 9% increase in the number of proficient students. This means that for every 100,000 students, there would be an increase to proficiency for 9,000 students. An alternative way to interpret effect sizes is that an effect size of 0.30 means that 62% of the group that did not use formative assessment practices would be below the average person in the group that did use formative assessment practices. Either way, the number of teachers using formative assessment may increase tremendously with the implementation of formative assessment systems being developed as part of the Common Core State Standards and Next Generation Science Standards. The consequential impact on student achievement may represent an important source of growth.

Types of Teacher Knowledge

Formative assessments require teachers to evaluate and analyze student work and, then, to determine appropriate next steps (Black & Wiliam, 1998; Heritage, 2007; Sadler, 1989; Shepard, 2005). To engage effectively in these practices, teachers must have sufficiently developed deep content (subject matter) knowledge and pedagogical content knowledge, such as knowledge of student misconceptions. Each of these types of knowledge impacts the quality of the judgments teachers can make about students.

Content knowledge. Content knowledge is the knowledge an individual has about a particular domain. In biology, this knowledge ranges from knowing the parts of a cell to more complex and abstract knowledge such as the relationship between genetics and evolution. Individuals with degrees in relevant subject areas, as well as those who work in fields where specialized knowledge is required, will all have relatively high degrees of content knowledge.

Because training and experience vary among teachers, the level of content knowledge on a particular topic will also vary. In fact, the literature suggests that not only is the amount of knowledge in a particular domain different between experts and novices, but the organization of that knowledge is also different. Experts tend to have principled organization of a content topic, while novices will describe a domain using superficial and fragmented features (Bruer, 1993; Chi, Feltovich, & Glaser, 1981). In other words, expert understanding of a topic can be viewed as linking subtopics to "big ideas" or "themes" of that topic while novices often see subtopics as unrelated.

This deeper knowledge is important to teaching. Hashweh (1987) examined the impact of content knowledge (obtained by card sorts, concept maps, summary statements) on different aspects of biology and physics teachers' instructional planning. Teachers were asked to plan activities given only a physics and biology textbook that included activities. Results indicated that individuals were able to modify activities and create new activities when they were working in their content area of expertise, but were not able to do so when they were working out of their content area of expertise. Teachers working in their content area of expertise were also able to suggest high-level questions not provided in the textbook, while teachers out of their content area were only able to suggest recall questions based on textbook content. Individuals with less content knowledge also had inaccuracies in their lesson plans and were unable to identify student inaccuracies and misconceptions. These results suggest that deeper content knowledge impacts the activities teachers choose, the questions they ask to determine student understanding, and their ability to diagnose student errors.

Overall, research has failed to find a large or consistent effect of content knowledge by itself on student achievement (Hattie, 2009). However, Baker et al. (1996) found a consistent effect of self-reported teacher knowledge and student performance in the same content area. Some researchers have argued that content knowledge only matters up to a certain point, after which it has no further impact on student achievement (Monk, 1994). This would mean that teachers may need a specific level of content knowledge relevant to their goals for students, whereas expert levels of content knowledge would not be relevant or result in a positive impact on student achievement.

Content knowledge is necessary to effectively perform professional activities. For example, content knowledge is necessary for judging student errors and understanding. If a teacher has only a novice’s understanding of the subject matter he or she will only be able to identify surface-level understanding by a student. If a teacher has a more expert understanding of principles and themes and how they interrelate, they will be able to recognize more complex student learning and discern important misconceptions. Content knowledge is also essential to teachers because it is the basis of pedagogical content knowledge.

Pedagogical content knowledge. The construct of pedagogical content knowledge bridges the gap between content knowledge (e.g., biology, physics, etc.) and general pedagogical knowledge (e.g., classroom management, knowledge of different teaching strategies). It focuses on the domain-specific knowledge a teacher has about teaching their specific subject area, in particular “useful forms of representations, such as the most powerful analogies, illustrations, examples, explanations, and demonstrations… that make it comprehensible to others” (Shulman, 1986, p. 9). For example, in teaching natural selection to students, effective teachers must know that a common misunderstanding of students is that natural selection is a choice or that it is based on a need or desire.

There are many perspectives on the exact components of pedagogical content knowledge but most consider Shulman’s (1986) two components central: (a) knowledge related to representing a particular subject to students (e.g., analogies, examples, demonstrations, representations, etc.) and (b) knowledge of student understanding (e.g., misconceptions, learning difficulties, preconceptions, etc.) (Van Driel, Verloop, & de Vos, 1998). Pedagogical content knowledge is central to diagnosing student understanding in formative assessment because teachers need to be aware of key words or phrases that indicate student understanding or errors, what misconceptions students may have on a particular topic, and what the sources of certain errors are. Therefore, the level of knowledge an individual teacher possesses will likely influence interpretations of student errors and understanding.

Characteristics of Effective Professional Development

The literature recommends that effective professional development should (a) focus on a limited number of teaching practices, (b) address a specific content area, (c) provide opportunities for “active” learning, and (d) persist over time to increase the likelihood of positive outcomes. These characteristics also pertain to good instruction in general. For this study, a specific professional development session was developed based on three of the four recommended characteristics of professional development. The study evaluated whether a short-term professional development program was effective.

Professional development should focus on a limited number of teaching practices. Professional development opportunities often focus on general teaching strategies such as classroom management or higher order thinking skills. However, studies suggest that professional development opportunities focused on a limited number of teaching practices are more effective. For example, Desimone, Porter, Garet, Yoon, and Birman (2002) surveyed a large national sample of teachers over a three-year period about their professional development training and their classroom practices. They found that attending professional development that focused on one teaching practice increased the teacher's reported use of that practice as compared to a professional development opportunity focused on many practices.

Professional development should address a specific content area. The literature also suggests that professional development programs focused on a particular subject area have a more positive effect than those with a general focus (Blank, de las Alas, & Smith, 2007). In a survey of over 1,000 teachers who attended different Eisenhower-funded professional development opportunities, Garet et al. (2001) found that teachers who attended professional development focused on specific science or math content instead of general content were more likely to report an increase in their knowledge of the subject and skills related to teaching that subject.

Professional development should provide opportunities for "active" learning. Active learning describes activities that allow teachers to engage in learning instead of passively receiving information. Examples of active learning include the analysis of student work, lesson studies, and viewing and critiquing videos of classroom lessons. Evidence suggests that active learning is critical if teachers are to learn how to reflect on their teaching and learn how to use knowledge to improve their teaching (Ball & Cohen, 1999; Garet et al., 2001).

Recently, professional development opportunities have used artifacts to get teachers actively engaged in classroom practice. In several studies, videos of classroom lessons were used with groups of teachers. Over time, all of the studies found that teachers' discussions became more productive, analytical, and focused on student learning (Borko, Jacobs, Eiteljorg, & Pittman, 2008; Sherin & Han, 2004; van Es & Sherin, 2008). For example, at the beginning of discussions of lessons, teachers talked about what they saw in the video using surface-level descriptions. However, by the end of the period of professional development, teachers' discussion focused on deeper issues related specifically to teaching such as student mathematical thinking of certain concepts (van Es & Sherin, 2008). In one qualitative study, math teachers met once a month for a year and shared a common piece of student work at each meeting (Kazemi & Franke, 2004). These communities of teachers spent their time analyzing the student work, and researchers found that over time, there was a shift in the importance placed on student thinking rather than surface features of the products.

By getting closer to practices through the analysis of student work or watching video of classroom lessons, teachers' thinking shifted to be more student-centered. These examples also provide evidence that these shifts in teacher thinking can lead to changes in teacher practice.

Professional development sustained over time increases the likelihood of positive outcomes. In addition to focusing professional development on specific content and teaching practices, positive benefits have been found for sustaining professional development over time. For example, Yoon et al. (2007) analyzed nine studies that evaluated the impact of professional development on student achievement and also met the What Works Clearinghouse evidence standards. These studies reported between 5 and 100 contact hours with teachers. A positive and significant effect on student achievement was found for professional developments with more than 14 contact hours. For the three studies with fewer than 14 hours of contact time, no statistically significant effect on student achievement was found. No examination of mediating teacher outcomes based on contact hours was included in any of the studies.

In another study, Garet et al. (2001) evaluated the impact of contact hours on opportunities for teachers to engage in active learning, such as evaluating student work, and on the coherence of professional development. Using self-reported survey data from over 1,000 teachers, time span and total contact hours were found to have a large positive influence on opportunities for teachers to participate in active learning and on their perceptions of the coherence of the professional development workshops. They concluded that it was not the particular amount of time that was important, but instead that

> longer activities tend to include substantially more opportunities for active learning, such as the opportunity to plan for classroom implementation, observe and be observed teaching, review student work, and give presentations and demonstrations. Longer activities also tend to promote coherence including connections to a teacher's goals and experiences, alignment with standards, and professional communication with other teachers. (Garet et al., 2001, p. 933)

One area that needs further examination is the impact of contact hours on teacher outcomes. The conclusions by Garet et al. (2001) suggest that if teacher professional development programs are focused on providing active learning opportunities, shorter term programs may also impact teacher outcomes. Because restrictions on educational funding inhibit long-term professional development activity, shorter, more affordable professional development is needed. In this study, the professional development was focused on a specific content area, natural selection, and the aim was limited to examining teacher judgments about student understanding and errors from given samples of student work. Time was limited to a one-session professional development.

Professional Development Related to Formative Assessment

In general, the field lacks information on how to develop teacher formative assessment skills using effective characteristics of professional development. Only seven studies could be found that reported the impact of professional development focused on formative assessment. Four of the studies only examined student performance as an outcome measure. Quint, Sepanik, and Smith (2008) compared classrooms in 21 schools that volunteered to use a formative assessment program and received one-on-one professional development throughout the school year to classrooms in 36 schools that did not use the program. No differences were found on student performance on state assessments. Meisels et al. (2003) examined standardized test scores to evaluate the impact of a three-year implementation of an embedded performance assessment system in schools that self-selected to implement the program. This study found significantly higher standardized reading scores for the treatment group. Wiliam, Lee, Harrison, and Black (2004) also studied students of 24 teachers who volunteered to participate in a long-term professional development as compared to students of teachers who did not. Results indicate a small to medium effect size of .32 in favor of the volunteered teachers' students on national tests administered to students. Finally, Yin et al. (2008) randomly assigned 12 teachers to one of two conditions. They treated teachers with a short-term professional development session to train them how to use specific embedded assessments in their science classrooms. The control teachers received the embedded assessments, but no training on how to use them. Results showed no differences between conditions on student performance data. The inconsistencies in the results are difficult to interpret and are limited by non-random assignment into groups or small sample sizes for many studies. The lack of teacher outcome data and the minimal explanation of professional development activities in these studies make interpretation of non-significant results especially difficult.

In addition, two studies measured teacher and student outcomes. First, Phelan, Choi, Vendlinski, Baker, and Herman (2009) randomly assigned 91 teachers within schools to a treatment or control group. The treatment consisted of training for teachers on three lessons to be taught throughout the year, and focused on the math topics and the formative assessment process for each lesson. Control teachers did not receive professional development on formative assessment, but were asked to teach students the same math topics as the treatment group using their own lesson plans. In addition, experimental group teachers and researchers met as small groups after each of the three lessons to review student work and examine common student errors. When conditions were compared, results showed significant differences in favor of the treatment group on student performance of standardized math outcome measures (state assessments) and on teacher math outcomes.

In the second study measuring teacher and student outcomes, Brookhart, Moss, and Long (2007) examined the impact of teacher enrollment in a year-long professional development program. In this study, student performance on district benchmark tests ($n = 109$) in six kindergarten and first grade classes whose teachers were selected to participate in the professional development were compared to students ($n = 42$) in the same district whose teachers were not enrolled in the program. Results indicated students in both groups improved similar amounts from pretest to posttest, but that in first grade the treatment group showed significant improvement over the control group. Teacher self-reflections and researcher notes were also collected to determine the impact of the professional development on teachers in the professional development program. An examination of the teacher data collected suggested an increase in teacher knowledge about and practices that may influence formative assessment. Results of these two studies suggest that student performance was affected by the professional developments. However, non-random assignment in the Brookhart et al. (2007) study makes interpretation of these results difficult. Unlike studies that only examined student outcomes, including mediating teacher information can help explain why certain student outcomes were found and others were not.

A final formative assessment study only measured teacher outcomes related to formative assessment skills and practices. In this study, individuals who were participating in the National Board Certification process were compared to individuals who were not participating over a three-year period (Sato, Chung, & Darling-Hammond, 2008). The National Board Certification is a voluntary certification for teachers who have taught at least three years. Benefits of this certification include but are not limited to monetary incentives, reciprocity of certification among many states, and prestige. Videotapes of classroom lessons, responses to questions about videotaped lessons, interviews, student work samples, and surveys were evaluated on six dimensions of formative assessment practice to determine changes in teacher views on and use of formative assessment in their classroom. Results indicate National Board teachers had statistically higher mean scores on the six dimensions of formative assessment, used a wider variety of assessments, and used their assessments in a wider variety of ways than teachers not participating in the National Board process. Again, caution must be used in the interpretation of results due to non-random assignment into groups and differential motivation of Board-certified teachers.

To summarize, the review highlights studies examining the impact of professional development related to formative assessment practices. Very few can be found. The causal inferences that can be made about the impact of professional development are limited by the non-random assignment into condition in most of these studies. Moreover, a lack of consistent results suggests professional development programs infrequently impact student performance. In addition, because most of the studies did not measure teacher interim outcomes, the chain of interpretation is broken. It is not known whether the programs impact teacher knowledge or skill related to formative assessment. If the results were negative on student outcomes, it is not known whether the treatment had the necessary effect on the teachers. In the Handbook of Formative Assessment, Schneider and Randel (2010) recommended,

> Research on professional development in formative classroom assessments should include proximal teacher outcomes to help understand the processes or mechanisms responsible for producing any potential effects… Some measure of these proximal outcomes is necessary to begin to understand their relations with the ultimate outcome of student achievement. (p. 269)

The measurement of mediating teacher outcomes would yield a better understanding about what treatments positively impact the development of teacher knowledge of and skill in formative assessment practices. Measuring teacher outcomes would have the added benefit of creating a criterion measure for professional development programs.

Natural Selection as a Topic of Study

The topic of content of this study is natural selection. While it is not without controversy in some states, natural selection was chosen because it is a key component of evolution and is a topic taught at many levels of schooling (California State Department of Education, 1990; National Research Council, 1996; Rutherford & Ahlgren, 1989).

Additionally, there is a large research literature base related to misconceptions students hold related to natural selection (Bishop & Anderson, 1990; Brumby, 1984; Lawson & Thompson, 1988; Nehm & Reilly, 2007). Common student misconceptions related to natural selection include (Bishop & Anderson, 1990):

- Changes in traits are attributed to a need-driven adaptive process rather than random genetic mutations and sexual recombination.

- Variation in traits within a population is not identified as being responsible for natural selection.
- Differences in reproductive success are not identified as being responsible for natural selection.
- Traits are seen as gradually changing in all members of a population.

In addition, some research has been conducted to identify the sources of these misconceptions. Greene (1990) asked 322 college students in an introductory biology course to respond to the following question:

> The ancestor of the modern day bat could not fly, resembling a shrew or mouse. Assume that the bat evolved wings from the arm and paws of shrew-like ancestors. Explain how this could have happened using the idea of natural selection.

Responses were classified for (a) the type of change focus (e.g., population focus or typological focus), (b) the attribution to change (e.g., if changing environments are linked to change, if variations are attributed to change, etc.), and (c) the stated mechanism for selection (e.g., Darwinian, Lamarckian, etc.). Results indicated that there was a pattern to student misconceptions related to natural selection and that the origins of individual student misunderstanding were attributed to (a) how students view variation in a population and (b) what students attribute change to. The identification of the sources of misconceptions is important because they help provide evidence for why a student might hold a particular misconception. This type of information could be helpful to teachers evaluating student understanding and errors.

Natural selection is a topic of study that most teachers must teach, and it is a topic that is complex for students to understand. This complexity creates a need for teachers to have a deep understanding of natural selection and high levels of pedagogical content knowledge to ensure student learning and competently assess student work.

Summary of Literature

The literature reviewed suggests that if we want to improve teacher formative assessment practices, we need to create professional development interventions that help teachers develop the subject matter knowledge and pedagogical content knowledge needed to diagnose and interpret student errors and understanding.

Methods

The methods section is divided into five sections: research questions, participants, research design, measures, and data collection procedures.

Research Questions

The main research questions examined in this study were:

- Are there differences between conditions on participants' overall analyses of student work?
- Are there differences between conditions in participants' identification of critical elements of student understanding and errors in written work related to natural selection?
- Are there differences between conditions on the level of post-session content knowledge related to natural selection?

In addition, the two secondary research questions addressed in this study were:

- Are there differences between conditions on teacher perceptions of the professional development?
- Are there differences between conditions on their perceived skills related to analyzing student work?

Participants

Recruitment. Volunteer high school biology teachers in the Los Angeles area were recruited to participate in this study. The purpose of recruiting only high school biology teachers was to limit the variation among participants.

Teacher participants were recruited using an informational study flyer. The flyer was sent to online teacher list serves. Interested individuals were asked to contact the researcher and were then given an information sheet with more details about the study. They were asked to complete the online eligibility screening survey. Interested individuals were also asked to indicate which of the five professional development dates he or she could attend. Eligibility criteria included being a current or former (within two years) high school biology teacher able to attend a professional development session at UCLA. If the individual was eligible and willing to participate, he or she was placed on the participation list. Before completing the pretest, individuals consented to participate in research.

Power analysis was conducted using G*Power 3.1 (Erdfelder, Faul, & Buchner, 1996) to determine the appropriate sample size for the main study to achieve power of .80 using the dependent measure from the pilot study. Results of the power analysis indicated that for a two-group design using an Analysis of Covariance (ANCOVA) with one covariate, the following total sample sizes would be needed: 15 participants for a large anticipated effect size (f = .80), 34 participants for a medium anticipated effect size (f = .50), and 90 participants for a low anticipated effect size (f = .30). Pilot study data and effect sizes of similar studies indicated that a medium to large anticipated effect size was justified. Therefore, a goal of enrolling 90 participants in the main study was set. However, the number of recruited participants was 82.

Dropout rate. Overall, 82 teachers responded to the informational flyer sent via teacher list serves. Of those that responded, 76 were eligible for the study. Of the 76 that were eligible for the study, 13% (*n* = 9) took the pretest but did not watch the videos or attend a professional development session. One additional individual took the pretest, watched the videos, and attended a majority of the professional development session but did not take the posttests due to early departure. Therefore, the total dropout rate for individuals who were accepted into the study and completed the pretest was 20% (*n* = 10). Six were in the control condition and four in the experimental condition. If individuals who did not complete preprofessional development videos are excluded, the dropout rate was 2% (*n* = 1). The total number of eligible participants who took the pretest, watched the videos, and attended professional development was 66.

Assignment to condition. Five professional development sessions were offered: two control condition sessions and three experimental condition sessions. Twenty-six eligible participants were only able to attend specific sessions. Either they were only able to attend one professional development session or were only able to attend professional development sessions offered for one condition (e.g., they were able to attend three dates, but all three were the experimental condition). Therefore, these 26 individuals were placed non-randomly into a professional development session because of their limited availability on certain dates (control: *n* = 12, experimental, *n* = 14). The remaining 40 individuals were randomly assigned to a condition.

It was decided that the 26 non-randomly assigned individuals would not be included in the analyses. This decision was based on the examination of condition equivalence. Two sets of condition equivalence analyses were conducted. First, condition equivalence for the entire sample of 66 participants was examined. Results of this analysis indicated that the control condition had significantly more prior training on assessing students (*p* = .03). Next, condition equivalence for the 40 randomly assigned participants was examined. Results indicated no significant differences between conditions for the 40 randomly assigned participants. Because the focus of the professional development was related to assessing students and to be more conservative, it was decided that the 26 non-randomly assigned individuals would not be included in the analyses. Analyses focused only on the 40 individuals who were randomly assigned to conditions.

Payment. All teachers received $175 for completing the study.

Description of participants. A summary of participants' teaching and educational experience is presented next in Table 1 and Table 2.

Table 1

Description of Participants' Teaching Experience

	Control (*n* = 20)		Experimental (*n* = 20)	
Background variable	*M*	*SD*	*M*	*SD*
Total years teaching experience	14.70	7.51	10.45	7.59
Total years teaching high school biology	8.95	7.76	6.05	6.07
Total years teaching middle school biology/life science	2.50	4.89	2.20	2.93
Years teaching:				
Honors biology	2.15	3.36	1.05	1.82
Regular biology	6.37	5.64	5.68	6.31
Advanced Placement biology	1.32	3.38	0.95	1.90
Remedial biology	1.74	3.57	0.63	2.31
English Language Learner biology	3.74	6.40	1.26	2.10

Table 2

Description of Participants' Background Variables

	Control (*n* = 20)		Experimental (*n* = 20)	
Background variable	*n*	%	*n*	%
Undergraduate major[a]				
Biology	13	65%	15	79%
Education	3	15%	0	0%
Other field	7	35%	5	25%
Master's degree				
Biology	5	25%	1	5%
Other science field	0	0%	1	5%
Education	10	50%	8	40%
Other field	6	30%	3	15%
None	1	5%	7	35%
Credential				
Currently credentialed	15	75%	18	90%
Obtaining credential	0	0%	1	5%
Not credentialed (retired, private school, etc.)	5	25%	1	5%
Prior training on assessing students				
Completed full course	7	35%	3	15%
Completed a few days of training	9	45%	10	50%
None	4	20%	7	35%

[a]The total percentage of undergraduate majors exceeds 100% due to the inclusion of double majors by some participants.

Sixty percent of the total sample was female with an average total teaching experience of 13 years. The average number of years teaching high school biology was approximately eight. Seventy-two percent had biology undergraduate degrees, an additional 8% had education degrees, and the remaining 20% obtained undergraduate degrees in other fields such as business or art. Seventy-one percent also held Master's degrees, of which 15% were in a biology field. Two participants possessed PhDs in the following fields: biology and education. Eighty-three percent had a current California teaching credential, three participants were obtaining a credential, two taught at private schools and did not need a credential, and four additional participants were either retired or unemployed within the past year and did not have current credentials.

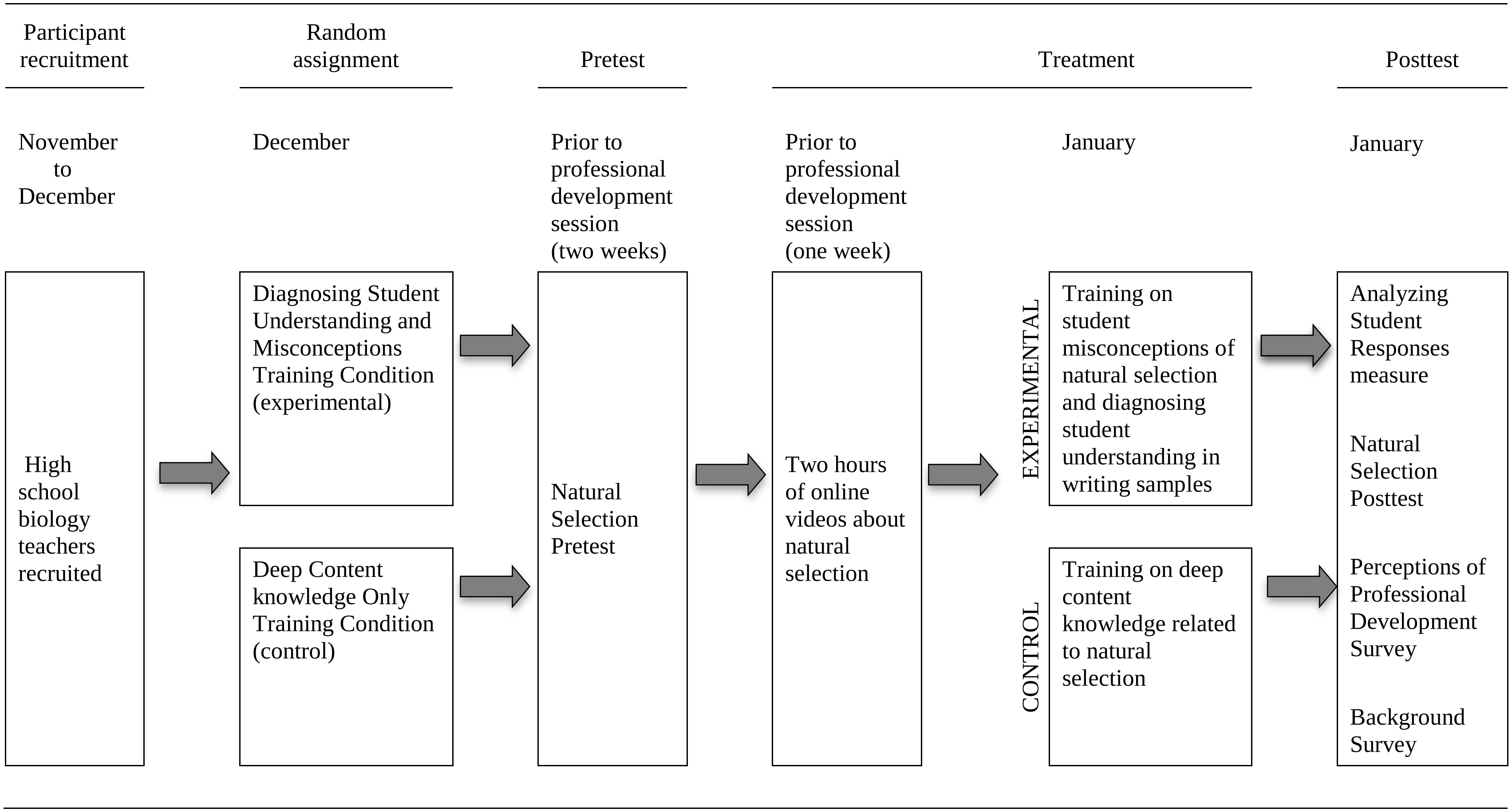
Research Design and Procedures

This study used a two-group experimental and control condition design. The study evaluated the effects of a short-term professional development session on participant skill analyzing student understanding and errors. The 40 teachers were randomly assigned to one of two conditions: (a) Diagnosing Student Understanding and Misconceptions Training Condition (i.e., experimental) or (b) Deep Content Knowledge Only Training Condition (i.e., control).

The timeline and schematic for the study is presented in Table 3. First, prior knowledge of natural selection for all participants was gathered with a Natural Selection Pretest. Next, participants in both conditions watched approximately two hours of online videos about natural selection to create a baseline of knowledge prior to the professional development sessions. The professional development session for the experimental condition focused on helping teachers understand student thinking about natural selection, including understanding the wide range of student misconceptions and errors associated with natural selection. For the experimental condition, participants spent time analyzing student errors and did not receive direct training on subject matter related to natural selection. The professional development session for the control condition focused only on the subject matter of natural selection. Control participants did not receive training on student thinking about natural selection.

Table 3

Study Timeline and Schematic



Participant recruitment	Random assignment	Pretest	Treatment		Posttest
November to December	December	Prior to professional development session (two weeks)	Prior to professional development session (one week)	January	January
High school biology teachers recruited	Diagnosing Student Understanding and Misconceptions Training Condition (experimental) Deep Content knowledge Only Training Condition (control)	Natural Selection Pretest	Two hours of online videos about natural selection	EXPERIMENTAL: Training on student misconceptions of natural selection and diagnosing student understanding in writing samples CONTROL: Training on deep content knowledge related to natural selection	Analyzing Student Responses measure Natural Selection Posttest Perceptions of Professional Development Survey Background Survey

The independent variable in this study was professional development condition. The main dependent variable analyzed was a measure, Analyzing Student Responses (ASR), which was based on four samples of student written work related to natural selection that the participants evaluated. Demographic information was obtained through a background survey administered at the end of the professional development session. In addition, a content knowledge posttest on natural selection, the Natural Selection Posttest, and a survey of participant perceptions of their professional development experience, the Perceptions of Professional Development Survey, were administered after the professional development session.

Measures

Five measures were administered to participants in this study. Table 4 presents the administration timeline and a description of each of the study measures.

Table 4

Description of Study Measures

Measure	Administration timeline	Description
Analyzing Student Responses (ASR)	Posttest	Main outcome measure: Open-ended responses and ratings of the same four samples of student work
Natural Selection Pretest	Pretest	Multiple-choice and open-ended items related to natural selection
Natural Selection Posttest	Posttest	Multiple-choice and open-ended items related to natural selection
Background Survey	Posttest	Survey: Demographic information
Perceptions of Professional Development Survey	Posttest	Survey: Self-reported perceptions of professional development experience

Main outcome measure: Analyzing Student Responses. Participants were asked to complete the ASR as a dependent posttest measure (see Appendix A). The purpose of this measure was to obtain participants' judgments about student understanding and misconceptions on the topic of natural selection. The ASR consisted of four samples of different students' writing. Participants were first asked to evaluate each sample of student writing in an open-ended format and then rate each student's understanding of natural selection and errors.

Development of the ASR first required the creation of simulated student work samples. Student samples were compiled based on anonymous samples collected from two high school environmental science classes. Simulated student responses were initially compiled to highlight a range of student misconceptions and errors, misunderstandings, and omissions related to natural selection. For example, the first simulated student response, Task A, features the common misconception that change occurs because of need or desire, the misconception that change is always directional, the misunderstanding by the student about "strong" and "weak," and an omission about variation within species.

Next, the ASR was administered to eight high school biology teachers during the pilot study to test the wording, format, and instructions, and to eliminate simulated student responses that showed a minimal range of participant ratings. The pilot tested version of the ASR consisted of eight samples of student work. Results indicated that ratings of four of the student samples showed minimal variation among teacher ratings. These student samples were eliminated from the measure leaving four samples for the final measure. Additionally, results of the pilot study indicated that the instructions were too complex and not clear. Instructions were clarified and simplified for the main study.

The final step in the ASR development was expert verification of ratings. Two experts on student thinking about natural selection were asked to agree or disagree with the ratings given for each simulated student response and explain their agreement or disagreement. The purpose of the expert review was to verify the classification accuracy of each student response and to elicit justification from experts regarding their rating. Both experts agreed with the classification on three of the four simulated student responses. Task D did not have exact agreement between experts. It was modified, and then reverified by the experts. After the second round of expert verification, experts agreed with the classification of Task D.

The final version of the ASR administered to participants consisted of two parts, as shown in Appendix A. First, participants were asked to evaluate, in open-ended format, what each of the four students understood about natural selection, what their misconceptions and errors were, and what they omitted from their written responses. Then they were asked to rate each student's overall understanding of natural selection on a scale of 1 to 4.

Participants' open-ended responses were coded for the identification of the embedded errors, understanding, and omissions within each student responses. Table 5 summarizes the components of student responses in the writing samples. For each component shown in Table 5, participants were awarded one point. Points for each student response were added together for the total embedded components identified in the ASR.

Table 5

Description of the Errors, Understandings, and Omissions Embedded Within and the Ratings of the Final Set of Simulated Student Responses for the ASR Task

Task	Rating	Shows signs of misconception(s)/ misunderstanding(s)	Shows signs of understanding(s)	Omission(s)
A	Poor	- Incorrectly believes change is directional - Incorrectly associates change with need and desire - Survival of the fittest incorrectly linked to “strong” and “weak”		- Variation within species
B	Good	- Strong and weak associated incorrectly with sight	- Differential reproduction - Variation (basic) - Inheritance (basic)	
C	Poor	- Incorrectly associates change with need and desire - Incorrectly believes all organisms in population change at once rather than individuals - Misunderstanding of role of differential reproduction in natural selection		
D	Basic		- Role of reproduction in natural selection - Variation (basic)	- Differential reproduction - Omits inheritance

Note. Bolded entries indicate critical elements of student responses.

Participants’ open-ended responses were also scored based on the number of critical elements of understanding and errors about natural selection identified within each student response. Critical elements are presented in Table 5 in boldfaced text. The critical elements are those that have been identified in the research literature as especially difficult concepts for students to understand. Additionally, the two experts used these elements as justification for their ratings of student understanding. Scores were created by awarding participants one point for the identification of each of the critical elements in each student response. For example, in the first simulated student response, Task A, participants who identified both the misconception about a desire to change and the misconception that change is directional were given two points, participants who identified only one of those misconceptions in the student response were given one point, and participants who did not identify either were given zero points. Scores for individual students were added together to create a Total Critical Element ASR Score.

Next, scores for the ASR ratings of student understanding of natural selection (1 to 4 Likert rating of each student's understanding of natural selection) were created. Participants were awarded one point for the correct rating, half a point for a rating that was one level from the correct rating, and zero points for a rating more than one level away from the correct rating. A total rating score was created by adding the scores for individual ratings of student understanding together for a maximum of four points.

Measures of teacher content knowledge. A pretest and a posttest of natural selection were administered to all participants. Information about each measure is described below.

Natural Selection Pretest. Participants were administered a pretest online within two weeks prior to their professional development session (see Appendix B). The purpose of this task was to measure participants' prior knowledge of natural selection to determine condition equivalence and to be used as a covariate in analyses. Parallel forms of some items from this pretest were used on the Natural Selection Posttest and are described in the next section.

Items for the Natural Selection Pretest were selected from a previously used measure (Shtulman, 2006). This instrument was initially created and used to measure understanding of natural selection with college students. The original instrument was divided into six subtopics related to natural selection: variation, inheritance, adaptation, domestication, speciation, and extinction. Each subtopic had between three and five questions. Shtulman used qualitative methods to compare experts (i.e., college professors and graduate students) to novices (i.e., undergraduate students), and found results that suggested that this measure distinguishes between individuals with less knowledge and those with deeper understanding of the topic.

Item selection for this study consisted of several stages. First, items from the initial Shtulman (2006) instrument were eliminated if they were considered too basic for a biology teacher or only indirectly related to natural selection. The remaining items were pilot tested with high school biology teachers to determine if there was a sufficient range of responses and if instructions and wording were appropriate for teachers. Results of the pilot test indicated that one item showed a ceiling effect, with all participants answering correctly. The item was eliminated from the measure. Remaining items were administered to two biology experts (i.e., biology professors) for review and expert ratings. Expert review showed that there was a discrepancy between experts on two items. It was concluded that the wording was confusing on this item, and it was excluded on the final version of the Natural Selection Pretest and the Posttest.

The final version of the Natural Selection Pretest administered in this study consisted of four multiple-choice items where participants explained their selections, one item where they ranked the relevance of certain factors related to natural selection, and one open-ended item. The other two survey items related to natural selection (e.g., Do you agree, partially agree, or disagree with the following statement, “Natural selection is not ‘survival of the fittest,’ but rather ‘reproduction of the fittest’”) were also included. These survey items were used to prepare for the professional development sessions rather than to determine participant understanding of natural selection. Participants were administered the pretest online, given unlimited time, and asked not to use outside resources to answer these items.

Multiple-choice items were scored as correct or incorrect. One point was awarded for each correct multiple-choice item. Two raters scored open-ended items for all measures. Before scoring, each rather was trained on how to use and interpret the three-part rubric shown in Appendix C. This rubric examines participant descriptions of variation within populations, differential reproduction, and inheritance. After the rubric was explained, each rater scored approximately 10% of the open-ended responses. Discrepancies in scores were discussed and resolved. This training process continued until rater agreement was high and raters were comfortable with the rubric. Raters then scored all open-ended items independently. Raters had an average exact agreement of 96%. Participants were awarded a maximum of one point for each aspect of the rubric for a total of three points for the open-ended items. Open-ended items were given more weight due to a higher difficulty.

Total scores for each participant were created by adding the number of correct multiple-choice items plus the open-ended scores. Cronbach’s alpha for the Natural Selection Pretest was .61. An internal consistency below .70 is often considered questionable. Therefore, some caution was used when interpreting results related to the Natural Selection Pretest.

Natural Selection Posttest. Participants were administered the Natural Selection Posttest shown in Appendix D after the professional development session. This task consisted of three multiple-choice and two open-ended items. The three multiple-choice items and one open-ended item were parallel to items from the Natural Selection Pretest. Ten survey questions related to the difficulty of the items were administered at the end of the posttest.

Items on the Natural Selection Posttest were scored using the same method as the pretest. Multiple-choice items were scored as correct or incorrect. Open-ended items were scored by the two raters who scored the Natural Selection Pretest. Exact agreement between raters for the open-ended items was 95%. Total scores for each participant were created by adding the number of correct multiple-choice items plus the open-ended scores. Cronbach's alpha for the Natural Selection Posttest was .65.

Survey measures. Two surveys were administered to all participants after their professional development session.

Background survey. All participants were administered a background survey after their professional development session (see Appendix E). This survey included typical demographic information such as age, ethnicity, and gender. It also included questions specific to teaching experience and education, such as the total number of years teaching, college degree(s) held, and self-rating scales of typical teaching activities. The background survey was administered during the pilot study to determine if wording and answer choices were appropriate for high school biology teachers. Minor changes were made to the formatting and wording of items based on pilot testing. Additional items were added to capture a fuller range of background information from participants.

Perceptions of Professional Development Survey. All participants were administered a 10-question survey at the end of their professional development session (see Appendix F). The purpose of this survey was to elicit participants' perceptions of (a) the effectiveness of their professional development experience, (b) whether goals of the professional development were met, (c) whether participants took the professional development seriously, and (d) the strengths and weaknesses of the professional development session. Space was also provided at the end of the survey for open-ended comments about their experience.

A scale for the Perceptions of Professional Development Survey was created by adding the Likert scale items together. Negative items were reversed before adding them in the scale. The internal consistency for this 10-item scale was .87 (Cronbach's alpha).

Pilot Testing

A pilot study was conducted with eight teachers. The primary purpose was to assess the wording of items, instructions, and range of responses on measures. Professional development materials and protocols were also examined. The pilot study was a two-day professional development on consecutive Saturdays. The first day was focused on content knowledge related to natural selection and the second day was focused on understanding student misconceptions and errors about natural selection. Participants were paid $200 for their participation in the pilot study. Several changes were made to the measures and professional development materials based on the pilot study as described earlier.

Data Collection Procedures

The pretest was administered online to all participants two weeks prior to the start of the professional development sessions using an online survey website. One week prior to their professional development date, participants were sent the links to two online videos.

- Video #1: 2010 COSEE-West: Evolution Presentation by Dr. Patrick Krug, California State University, Los Angeles (first 36 minutes; http://www.usc.edu/org/cosee-west/onlineworkshops.html)
- Video #2: Adaptive Evolution: Natural Selection by Stephen C. Stearns, Yale University (40 minutes; http://academicearth.org/lectures/adaptive-evolution-natural-selection)

Participants were asked to watch both videos and complete the seven-question quiz and survey on each video presented in Appendix G before attending their professional development session. The purpose of participants watching the videos was to create a baseline of knowledge about natural selection before the professional development session. The purpose of the quizzes was to verify that participants watched each video. In addition, there were six Likert-scale survey items (e.g., "I found the video interesting," "I learned something from the video," "I would show this video to my students") for each video. The quiz and survey took no more than 15 minutes to complete.

Five dates were available for professional development sessions. Two dates were devoted to the control condition and three dates were devoted to the experimental condition. For each professional development session, the same two researchers were present, the same PowerPoint presentations were used, and notes were taken in each professional development session to minimize differences between professional development sessions on the same topic. The overall time between conditions was identical. The location was different among conditions due to availability of rooms.

The Diagnosing Student Understanding and Misconceptions Training (i.e., experimental) emphasized understanding student thinking about natural selection and evaluating student writing samples for misconceptions and understanding of natural selection. No direct instruction on the content of natural selection was given. However, indirectly the content of natural selection was discussed. The experimental condition had the following schedule (9:00 a.m. – 4:30 p.m.):

1. *Introductions (30 minutes).* Participants were introduced to the researchers, provided background on the study, and given the agenda for the day.

2. *Student understanding about natural selection (1 hour).* Participants were presented with research related to student misconceptions of natural selection. A list of common student misconceptions related to natural selection was given to participants and the sources of these misconceptions were discussed as a group. Misconceptions were also broken down into three categories, including misconceptions about (a) variation within species, (b) differential reproduction, and (c) inheritance/DNA.
3. *Analyzing student work (3 hours).* Participants were shown five samples of student work of varying degrees of understanding. Student samples focused on (a) variation within species, (b) differential reproduction, and (c) inheritance/DNA. Participants discussed in small groups and then debriefed in the whole group about what each student understood about natural selection, what misconceptions or errors they had, and what was omitted from each response.
4. *Posttest measures and surveys (1 to 2 hours).* Participants were administered the ASR, the Natural Selection Posttest, the Perceptions of Professional Development Survey, and a background survey.

The Deep Content Knowledge Training (i.e., control) emphasized the subject matter of natural selection and understanding the connections between different subtopics of natural selection, including variation, differential reproduction, and inheritance. No instruction or materials on student thinking about natural selection were given to the control condition.

The control condition had the following schedule (9:00 a.m. – 4:30 p.m.):

1. *Introductions (30 minutes).* Participants were introduced to the researchers, provided background on the study, and given the agenda for the day.
2. *Research on expertise (1 hour)*. Participants were presented with research related to the development of expertise. The focus of this discussion was on characteristics of expert thinking.
3. *Natural selection (3 hours).* Participants were given news articles and watched video clips related to three subtopics of natural selection, including (a) variation within species, (b) differential reproduction, and (c) inheritance/DNA to discuss in small groups and then whole groups. Knowledge maps (also called concept maps) about each subtopic of natural selection were drawn by participants and discussed as a whole group.
4. *Posttest measures and surveys (1 to 2 hours).* Participants were administered the ASR, the Natural Selection Posttest, the Perceptions of Professional Development Survey, and a background survey.

It should be noted that breaks and lunch are not indicated in the above schedules. Several breaks were taken throughout the professional development sessions in addition to a 30-minute lunch. Break timing and length were comparable between professional development conditions.

Summary of Methods

Of the 76 eligible individuals who volunteered for the study, 40 were randomly assigned to one of two professional development conditions: (a) deep content knowledge/control condition (*n* = 20) or (b) student understanding/experimental condition (*n* = 20). All participants completed the online Natural Selection Pretest and then watched two online videos on the topic of natural selection prior to their professional development session. After completing approximately four hours of professional development, participants were administered two posttest measures consisting of a measure that focused on analyzing student work, the ASR, and the Natural Selection Posttest. Participants also completed the background survey and the Perceptions of Professional Development Survey.

Data Analysis and Results

In this section, a description of the data analyses and results are presented. First, data analysis procedures are presented. Then descriptive statistics related to background variables and results of the analyses of condition equivalence are presented. Next, descriptive statistics of dependent measures are presented. Finally, results of the analyses related to the main research question and secondary research questions are reported. A summary of results is provided at the end of this section.

Data Analysis Procedures

Descriptive analysis of background variables and condition equivalence. Before analyses of research questions were conducted, a descriptive analysis of background variables was conducted and condition equivalence was examined. The descriptive analysis consisted of the frequencies of categorical background data including the type of undergraduate degrees, if participants had Master’s degrees, teaching credential type, and the amount of prior training on assessing students. Means and standard deviations for teaching experience and Natural Selection Pretest scores and items were also examined.

Condition equivalence was examined by comparing differences between conditions on the Natural Selection Pretest and demographic information such as type of undergraduate degree, years teaching experience including years teaching different levels of biology (i.e., Advanced Placement biology), prior amount of training on assessing students, and the type and proportion of participants with Master’s degrees and PhD degrees. Condition differences on continuous variables such as pretest scores and years teaching were examined using independent samples *t* tests. Condition differences on categorical variables such as prior assessment course status and awarded Master’s degrees were examined using a chi-square test.

Descriptive analysis of dependent measures. Next, a descriptive analysis of scores and individual items from all study measures were examined by condition. This includes means and standard deviations of scores for the ASR, Natural Selection Pretest and Posttest, and continuous background variables such as the number of years teaching. Additionally, the frequencies of categorical data from study measures were examined including the Perceptions of Professional Development Survey and categorical background data such as the type of undergraduate degree. Finally, a correlational analysis of the ASR, Natural Selection Pretest and Posttest, Perceptions of Professional Development Survey, and background characteristics was conducted using Pearson's *r* to determine the relationship between variables for the whole group and for each condition separately.

Main research questions. For each of the main research questions below, the rationale, hypothesis, and analysis procedure are presented.

Main Research Question 1: Are there differences between conditions on participants' overall analyses of student work? The purpose of this research question was to determine if there were differences between conditions on the ASR scores. Results of this analysis were the main indicator of effectiveness of the experimental condition professional development session. It was hypothesized that there would be a statistically significant difference between conditions and that individuals in the experimental professional development session that focused on analyzing student errors and understanding would evaluate and rate student responses more accurately.

To determine if differences existed between conditions, the two factors of the ASR were examined separately. First, the total number of embedded components identified across all four students was examined using an independent samples *t* test. An independent samples *t* test was determined to be appropriate because samples were independent of each other, the dependent variable was continuous, the dependent variable had a normal distribution with similar variance in each condition, and no covariates were used. Next, four separate independent samples *t* tests were conducted to determine if there were differences on the number of embedded components identified in each student response. Cohen's *d* was also calculated to determine the effect size of the embedded components identified. Cohen's *d* is a measure of the magnitude of the treatment effect and does not take into account sample size. Effect sizes below 0.30 or less are typically considered small, near 0.50 are considered medium, and 0.80 and above are considered large.

The second factor in the ASR measure was the ratings of student understanding. Participants rated each student response on a 1 to 4 Likert scale (i.e., strong, proficient, basic, poor). Total scores of these ratings and ratings for individual students (1 point for correct rating, half a point for one level away, zero points for more than one level away) were analyzed using a chi-square test. Cohen’s *d* was also calculated to determine the effect size.

Main Research Question 2: Are there differences between conditions in participants’ identification of critical elements of student understanding and errors in written work related to natural selection? The purpose of this research question was to examine the open-ended responses from the ASR measure to determine if there was a differential impact of the experimental professional development session on participant identification of critical elements when evaluating student understanding and errors. It was hypothesized that participants in the experimental condition would identify critical elements of student understanding more frequently than those in the control condition.

To determine if one condition was more successful than the other at identifying critical elements of student understanding and errors, a related independent samples *t* test was conducted and Cohen’s *d* was calculated for both the overall number of critical elements identified and separately for individual student responses.

Additionally, the proportion of errors, understanding, and omissions identified were examined separately across all students to determine if differences existed between conditions on the type of information being identified using independent samples *t* tests and Cohen’s *d* estimate of effect size. This analysis was conducted to determine if one type of information (e.g., errors, understandings, or omissions) was more likely to be identified by one condition.

Main Research Question 3: Are there differences between conditions on the level of post-session content knowledge related to natural selection? The final main research question explores differences in performance on the Natural Selection Posttest between the two professional development conditions. It was expected that there would be significant differences in performance on the Natural Selection Posttest and that individuals in the control condition that focused on content knowledge would score higher due to the focus on deep content knowledge in the control condition’s professional development session.

Prior to examining condition differences, the homogeneity of regression assumption was tested by examining the interaction between each covariate and the independent factor. Next, a one-way ANCOVA was conducted using Natural Selection Posttest as the dependent variable, condition as the independent variable, and pretest as covariate. Cohen’s *d* was also calculated based on the unadjusted means.

Secondary research questions. For each of the secondary research questions below, the rationale, hypothesis, and analysis procedure are presented.

Secondary Research Question 1: Are there differences between conditions on teacher perceptions of the professional development? This research question determined if there were differences between conditions on their perceived skills related to analyzing student work. This question examined the impact of the professional development session on their perceived skill on activities related to diagnosing student understanding and errors. This question was measured by survey questions administered on the background survey after the professional development session. It was expected that participants in the experimental condition would consider themselves as more highly skilled due to the practice they received during the professional development session. To determine if conditions differed statistically on their self-perceptions, an independent samples *t* test was conducted on each survey item to determine if a statistical difference existed between conditions and Cohen's *d* was calculated.

Secondary Research Question 2: Are there differences between conditions on their perceived skills related to analyzing student work? This secondary research question explored the differences between conditions reported on the Perceptions of Professional Development Survey. Because there were two professional development conditions, it was important to determine if there was a difference between participant perceptions of the professional development sessions. It was expected that there would be no difference between professional development conditions on the perception of the professional development sessions. To determine if perceptions of conditions differed statistically, an independent samples *t* test was conducted and Cohen's *d* was calculated.

Results

Descriptive statistics of background variables and condition equivalence. Descriptive analyses of teacher background and results of condition equivalence are reported in this section. Table 6 reports categorical data from the Background Survey. A majority of teachers in both conditions had biology undergraduate degrees, many had Education Master's degrees, and some had biology Master's degrees or Master's degrees from other fields. Over three fourths of teachers in both conditions were currently credentialed in California. There were more females in the control condition and equal numbers of males and females in the experimental condition. There was a wide range of prior training on assessing students, with a majority of participants in both conditions only having a few days of training on this topic.

Table 6

Comparison of Condition Equivalence for Categorical Variables Related to Teacher Background

	Control (*n* = 20)		Experimental (*n* = 20)				
Background variable	*n*	%	*n*	%	*df*	χ^2	*p*
Undergraduate major					2	3.12[a]	.21
Biology	13	65%	15	79%			
Education	3	15%	0	0%			
Other field	7	35%	5	25%			
Master's degree					2	4.39[a]	.11
Biology	5	25%	2	10%			
Education	10	50%	8	40%			
Other field	6	30%	3	15%			
Credential type					2	3.93[a]	.14
Currently credentialed	15	75%	18	90%			
Obtaining credential	0	0%	1	5%			
Not credentialed (retired, private school, etc.)	5	25%	1	5%			
Prior training on assessing students					2	2.47	.29
Completed full course	7	35%	3	15%			
Completed a few days of training	9	45%	10	50%			
None	4	20%	7	35%			
Gender					1	1.67	.20
Female	14	70%	10	50%			
Male	6	30%	10	50%			

[a]Expected cell size is less than 5 in more than 20% of cells.

Examination of condition differences on the variables presented in Table 6 indicated that conditions did not differ significantly on these variables. Since the cell size was less than five in more than 20% of the cells for three of the five variables analyzed. Caution must be taken when interpreting these results.

Next, continuous variables from the background survey are reported in Table 7. These indicate that the total number of years of teaching experience varied among participants. The mean number of years of teaching experience was approximately 10 years, mostly teaching high-school-level biology. Only a few participants had middle school life science teaching experience. Regular level biology was the most taught level of biology. There was a relatively low average of years of experience teaching English Language Learners, Honors, Advanced Placement, and remedial biology. Examination of condition equivalence as presented in Table 7 indicates that the conditions did not differ significantly on these variables.

Table 7

Comparison of Condition Equivalence on Continuous Variables Related to Teaching Experience

	Control (*n* = 20)		Experimental (*n* = 20)				
Variable	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*
Total years teaching experience	14.70	7.51	10.45	7.59	38	1.78	.08
Total years teaching high school biology	8.95	7.76	6.05	6.07	38	1.32	.20
Total years teaching middle school biology/life science	2.50	4.89	2.20	2.93	38	-1.27	.20
Years teaching:							
Honors biology	2.15	3.36	1.05	1.82	36	1.29	.21
Regular biology	6.37	5.64	5.68	6.31	36	0.41	.72
Advanced Placement biology	1.32	3.38	0.95	1.90	36	0.41	.68
Remedial biology	1.74	3.57	0.63	2.31	36	1.60	.27
English Language Learner biology	3.74	6.40	1.26	2.10	36	1.60	.12

Finally, the Natural Selection Pretest scores and item level data are presented in Table 8. Data indicate that participants in both conditions achieved approximately 60% correct. There were no differences between conditions on the Natural Selection Pretest.

Table 8

Comparison of Condition Equivalence on Participants’ Natural Selection Pretest Scores

	Control (*n* = 20)		Experimental (*n* = 20)				
Score	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*
Total score	4.86	1.49	4.69	1.61	38	0.34	.74
Percent correct	.60	.19	.59	.20			

Note. Maximum score of 8.

Descriptive statistics of dependent measures. Results of the descriptive analysis of the study measures are presented next. First, results of the ASR measure indicate that scores were relatively low in both conditions (see Table 9). Participants in the experimental condition scored higher than those in the control condition on the number of embedded components identified, the total rating score, and the total number of critical elements identified. Item-level descriptive statistics for the ASR are found in Appendix H.

Table 9

Description of the Scores for the ASR Measure

		Control (*n* = 20)		Experimental (*n* = 20)	
ASR score	Maximum score	*M*	*SD*	*M*	*SD*
Total number of embedded components identified	15	4.80	2.42	6.10	1.97
Total ratings of student understanding score	4	1.75	0.50	1.85	0.73
Total number of critical elements identified	8	2.50	1.47	3.95	1.47

Next, descriptive statistics for the Natural Selection Posttest scores are presented in Table 10. Data indicate that participants in the control condition achieved 55% of the items correct and those in the experimental condition achieved 62% correct. Item-level descriptive statistics for the Natural Selection Posttest are found in Appendix H.

Table 10

Description of the Natural Selection Posttest Scores

	Control (*n* = 20)		Experimental (*n* = 20)	
	M	*SD*	*M*	*SD*
Total score	4.92	1.71	5.64	1.20
Total percent correct	.55	.19	.62	.13

Note. Maximum score of 9.

Results of the descriptive analysis of the two survey measures are presented next. Results of the Perceptions of the Professional Development Survey indicate that overall perceptions were very positive on individual items as presented in Table 11.

Table 11

Description of the Perceptions of Professional Development Survey Items Between Conditions

	Control (*n* = 20)				Experimental (*n* = 20)			
	Agree		Disagree		Agree		Disagree	
Survey items	*n*	%	*n*	%	*n*	%	*n*	%
This training was…								
interesting	20	100%	0	0%	20	100%	0	0%
organized	20	100%	0	0%	20	100%	0	0%
helpful	19	95%	1	5%	19	95%	1	5%
boring[a]	3	15%	17	85%	3	15%	17	85%
informative	19	95%	1	5%	19	95%	1	5%
I learned from this training	16	80%	4	20%	16	80%	4	20%
This training was not worth the amount of time it took[a]	1	5%	19	95%	3	15%	17	85%
I would recommend this training	17	85%	3	15%	18	90%	2	10%
I will use examples	20	100%	0	0%	19	95%	1	5%
I enjoyed the training	19	95%	1	5%	19	95%	1	5%

[a]Indicates negative perception item. Scale was reversed for analyses.

Table 12 presents the reported perceptions of skill conducting different classroom activities, measured after the professional development session. Results indicate that participant perceptions were relatively high. Most participants rated themselves as skilled rather than highly skilled or unskilled in all categories. No participants rated themselves as highly unskilled in any category.

Table 12

Description of Teacher Perception of Skill Performing Specific Classroom Activities

	Control (*n* = 20)		Experimental (*n* = 20)	
Skill level	*n*	%	*n*	%
Giving feedback				
Highly unskilled	0	0%	0	0%
Unskilled	3	15%	1	5%
Skilled	12	60%	15	75%
Highly skilled	5	25%	4	20%
Creating formative assessments				
Highly unskilled	0	0%	0	0%
Unskilled	5	25%	3	15%
Skilled	11	55%	13	65%
Highly skilled	4	20%	4	20%
Analyzing student misconceptions				
Highly unskilled	0	0%	0	0%
Unskilled	2	10%	4	20%
Skilled	14	70%	13	65%
Highly skilled	4	20%	3	15%
Diagnosing level of student understanding				
Highly unskilled	0	0%	0	0%
Unskilled	4	20%	1	5%
Skilled	10	50%	14	70%
Highly skilled	5	25%	5	25%

The frequency of different professional activities as reported by participants after the professional development session is presented in Table 13. A majority of participants attended professional development outside of the school-provided professional development a few times a year. While there was a wide range in the frequencies, a majority of participants in both conditions reported providing feedback to students between a few times a month to once a week and assigning long written assignments a few times a semester or less, and over 65% of participants in each condition reported analyzing student misconceptions at least once a month.

Table 13

Description of Frequencies Conducting Professional Activities

	Control (*n* = 20)		Experimental (*n* = 20)	
Skill level	*n*	% correct	*n*	% correct
Attend professional development (not school sponsored)				
Never	2	10%	1	5%
Few times a year	12	60%	16	80%
Few times a semester	4	20%	3	15%
Few times a month	2	10%	0	0%
About once a week	0	0%	0	0%
Daily or almost daily	0	0%	0	0%
Provide feedback to students				
Never	1	5%	0	0%
Few times a year	2	10%	1	5%
Few times a semester	4	20%	2	10%
Few times a month	7	35%	7	35%
About once a week	3	15%	5	25%
Daily or almost daily	3	15%	4	20%
Assign long written assignments				
Never	6	30%	3	15%
Few times a year	5	25%	8	40%
Few times a semester	4	20%	6	30%
Few times a month	3	15%	1	5%
About once a week	1	5%	1	5%
Daily or almost daily	1	5%	0	0%
Analyze student misconceptions				
Never	1	5%	2	10%
Few times a year	3	15%	3	15%
Few times a semester	2	10%	1	5%
Few times a month	4	20%	2	10%
About once a week	5	25%	4	20%
Daily or almost daily	5	25%	7	35%

Correlational analyses (Pearson's *r*) for all study measures for the entire sample are reported on Table 14, the control condition in Table 15, and the experimental condition in Table 16. A statistically significant high positive correlation was found between the Natural Selection Pretest and Posttest scores, and the total number of years teaching and the number of years teaching biology. These are as expected. A high positive association was found among the total number of critical elements identified in the ASR and the total embedded components identified. A moderate positive correlation was found among the critical elements identified in the ASR, condition, and self-perceived skill diagnosing student understanding. Statistically significant but moderate positive correlations were found between participants' self-reported ability to diagnose student understanding and identify misconceptions, their teaching credential type, and their undergraduate major. Biology majors were more likely to have current teaching credentials.

For the control condition, a statistically significant but moderate positive correlation was observed between the Natural Selection Pretest score and the total ASR embedded components identified. A positive correlation was found between self-perceived ability to identify misconceptions and their undergraduate major with those who had biology majors rating themselves with higher ability.

For the experimental condition, a statistically significant but moderate relationship was found between total number of years teaching and their self-reported ability to identify misconceptions. Remaining correlations show little or no relationship among variables.

Table 14

Correlation Matrix for the Entire Sample

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Condition	–														
2. Total ASR ratings score	.08	–													
3. Total ASR embedded components identified score	.29	-.04	–												
4. Total critical elements identified	.46**	.06	.79**	–											
5. Natural Selection Pretest score	-.06	-.22	.17	.25	–										
6. Natural Selection Posttest score	.24	-.23	.14	.29	.62**	–									
7. Perceptions of the Professional Development score	.00	.30	-.15	.00	-.25	-.10	–								
8. Prior training on assessing students	-.24	.07	-.27	-.26	.16	.01	.13	–							
9. Self-perception of ability to diagnose student understanding	.24	.00	-.10	-.09	-.04	.08	-.01	.10	–						
10. Self-perception of ability to identify misconceptions	-.14	.26	-.32*	-.47**	-.11	-.17	-.07	.18	.48**	–					
11 Years teaching biology	-.21	-.02	-.23	-.28	-.01	.14	.12	.15	-.11	.26	–				

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
12. Total years teaching	-.28	-.02	-.35*	-.35*	-.13	.04	.21	.04	.00	.40*	.71**	–			
13. Teaching credential type	-.26	-.01	-.10	-.06	-.17	-.14	.02	.02	.10	.12	-.01	.09	–		
14. Undergraduate major	-.09	-.03	-.05	.00	-.03	.10	.14	.00	-.10	-.22	-.04	.00	.50**	–	
15. Type of Master's degree	.04	.07	.18	.13	.00	.13	.17	.06	.10	.05	.01	-.07	-.03	-.14	–

**p* < .05. ***p* < .01.

Table 15

Correlation Matrix for the Control Condition

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Total ASR ratings score	–													
2. Total ASR embedded components identified score	-.04	–												
3. Total critical elements identified	-.04	.86**	–											
4. Natural Selection Pretest score	-.15	.45*	.45*	–										
5. Natural Selection Posttest score	-.26	.12	.18	.76**	–									
6. Perceptions of the Professional Development score	.37	-.09	-.11	-.26	-.19	–								
7. Prior training on assessing students	.11	-.27	-.31	.15	.07	.08	–							
8. Self-perception of ability to diagnose student understanding	.12	-.22	-.19	-.02	.00	-.05	.11	–						
9. Self-perception of ability to identify misconceptions	.00	-.38	-.35	-.01	-.09	-.04	.30	.66**	–					
10. Years teaching biology	-.23	-.06	-.09	.12	.33	-.04	.14	-.20	.22	–				
11. Total years teaching	-.25	-.26	-.26	-.07	.22	.09	.03	-.05	.26	.69**	–			
12. Teaching credential type	.06	-.20	-.04	-.32	-.18	.20	.04	.31	.19	-.15	-.07	–		
13. Undergraduate major	.05	-.01	.01	-.15	.09	.34	.15	-.17	-.48*	-.22	-.25	.58**	–	
14. Type of Master’s degree	-.12	.22	.15	.27	.29	.01	.29	.21	.15	-.22	-.31	-.13	-.01	–

**p* < .05. ***p* < .01.

Table 16

Correlation Matrix for the Experimental Condition

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Total ASR ratings score	–													
2. Total ASR embedded components identified score	-.08	–												
3. Total critical elements identified	.07	.67**	–											
4. Natural Selection Pretest score	-.26	-.10	.17	–										
5. Natural Selection Posttest score	-.26	.00	.25	.55*	–									
6. Perceptions of the Professional Development score	.25	-.27	.09	-.23	-.02	–								
7. Prior training on assessing students	.09	-.14	-.01	.15	.09	.21	–							
8. Self-perception of ability to diagnose student understanding	.21	-.11	-.35	-.05	.13	.06	-.07	–						
9. Self-perception of ability to identify misconceptions	.42	-.23	-.57**	-.19	-.22	-.07	.04	.46*	–					
10. Years teaching biology	.21	-.38	-.39	-.19	.01	.36	.05	.23	.28	–				
11. Total years teaching	.18	-.34	-.35	-.22	-.03	.37	-.08	.29	.47*	.71**	–			
12. Teaching credential type	-.04	.37	.28	.00	.17	-.24	-.24	-.27	-.04	-.19	.19	–		
13. Undergraduate major	-.05	-.04	.09	.07	.19	-.05	-.22	.12	-.05	.15	.20	.40	–	
14. Type of Master’s degree	.20	.10	.09	-.26	-.09	.32	-.18	-.14	-.02	.36	.21	.18	-.28	–

**p* < .05. ***p* < .01.

Main research questions. Results for the three analyses of the main research questions are presented below.

Main Research Question 1: Are there differences between conditions on participants' overall analyses of student work? The first research question examined in this study investigated the impact of the experimental treatment on participants' ability to evaluate student responses and rate student understanding.

First, results of condition differences on the number of embedded components identified on the ASR are presented. Results of the independent *t* test on the total number of identified errors, understandings, and omissions across all student responses indicate that the participants in the experimental condition identified more (*M* = 6.10, *SD* = 1.97) embedded components than those in the control condition (*M* = 4.80, *SD* = 2.42). This difference was not statistically significant, *t*(38) = 1.86, *p* = .07. Cohen's *d* was estimated at 0.59. The maximum number of embedded components possible to identify was 15. Results for the number of embedded components identified by each condition in individual student responses are presented in Appendix I.

Next, results of participant ratings of student understanding were examined. Results indicate that participant total rating scores did not differ significantly between conditions, *t*(38) = .51, *p* = .62. Cohen's *d* was 0.16. Results of ratings for individual student responses by condition are presented in Appendix I.

Main Research Question 2: Are there differences between conditions in participants' identification of critical elements of student understanding and errors in written work related to natural selection? The second research question examined in the study investigated the impact of the experimental treatment on the participants' ability to identify critical elements of student understanding.

First, results indicate that participants in the experimental condition identified significantly more of the critical elements across all student responses (*M* = 3.95, *SD* = 1.40) than those assigned to the control condition (*M* = 2.50, *SD* = 1.47), *t*(38) = 3.20, *p* = .003. Cohen's *d* was estimated at 1.01. The maximum score for the total critical elements identified was eight. Results for the number of critical components identified by each condition in individual student responses are presented in Appendix I.

Furthermore, data were analyzed to determine if one condition was more proficient at identifying different types of information across all student responses including errors, understandings, or omissions. Results in Table 17 indicate that the experimental condition identified a statistically higher proportion of errors (*M* = .47, *SD* = .16) than the control condition (*M* = .33, *SD* = .22), *t*(38) = 2.97, *p* = .004. The effect size for the proportion of errors identified was 0.73. No differences and low effect size were found for the proportion of understandings and omissions identified across student responses.

Table 17

Differences on the Proportion of the Types of Embedded Components Identified Across Students Between Conditions

	Control (*n* = 20)		Experimental (*n* = 20)					
Type of component identified	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Proportion of errors	.33	.22	.47	.16	38	2.97	.004	0.73
Proportion of understandings	.38	.26	.40	.26	38	0.32	.896	0.08
Proportion of omissions	.39	.34	.40	.26	38	0.24	.862	0.03

It should be noted that while significant differences were found between conditions in these analyses, participants in both conditions identified less than half of the total critical elements. Additionally, the proportion of errors, understandings, and omissions identified was never more than 50% of the total possible to identify.

Main Research Question 3: Are there differences between conditions on the level of post session content knowledge related to natural selection? The final research question examined in this study investigated the impact of professional development session on Natural Selection Posttest scores. To determine if there was a positive impact from the professional development session on the experimental condition participants' content knowledge of natural selection, first homogeneity of regression assumption was tested and then a one-way ANCOVA was conducted. Homogeneity of regression assumption was tested by examining the interaction between the covariate pretest and the independent factor condition. Results indicate no significant interaction between the covariate and the independent variable, F(1, 36) = 1.99, p = .11.

Next, Natural Selection Pretest scores, Natural Selection Posttest scores, and the adjusted Natural Selection Posttest scores based on the one-way ANCOVA using pretest as a covariate are presented in Table 18.

Table 18

Means, Standard Deviations, and Adjusted Means of Natural Selection Pretest and Natural Selection Posttest

	Control (*n* = 20)			Experimental (*n* = 20)		
	M	*SD*	*SE*	*M*	*SD*	*SE*
Natural Selection Pretest Scores (covariate) (maximum score 8)	4.86	1.49	---	4.69	1.61	---
Natural Selection Posttest Scores (maximum score 9)	4.92	1.71	---	5.64	1.20	---
Adjusted Natural Selection Posttest Scores using pretest as a covariate (maximum score 9)	4.86	---	.26	5.70	---	.23

Results of the one-way ANCOVA conducted using Natural Selection Posttest as the dependent variable, condition as the independent variable, and pretest as a covariate are presented in Table 19. Results indicate that the experimental condition scored significantly higher on the Natural Selection Posttest than the control condition after controlling for pretest score, F(1, 36) = 5.36, *p* = .03. Cohen's *d* for unadjusted means was 0.49.

Table 19

Analysis of Covariance of Natural Selection Posttest Score as a Function of Condition with Natural Selection Pretest Scores as Covariate

	df	*SS*	*MS*	*F*	*p*	η^2
Natural Selection Pretest Score (covariate)	1	34.51	34.51	26.99	.00	.43
Condition	1	6.85	6.85	5.36	.03	.13
Error	36	46.04	1.28			
Total	39	1177.24				

Secondary research questions. Results of the two secondary research questions are presented below.

Secondary Research Question 1: Are there differences between conditions on their perceived skills related to analyzing student work? This research question examined the impact of the professional development sessions on participants' perceived skill on activities related to diagnosing student understanding and errors on survey questions administered after the professional development session. Results presented in Table 20 indicate that no differences were found between participants in the experimental and control conditions on their perceived skill on activities related to diagnosing student understanding and errors.

Table 20

Differences on the Self-Perceived Skill Conducting Classroom Activities Related to Assessing Student Work Between Conditions

	Control (*n* = 20)		Experimental (*n* = 20)					
Perceived skill	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Giving feedback	3.10	.64	3.15	.49	38	.28	.78	0.09
Creating formative assessments	2.95	.69	3.05	.61	38	.49	.62	0.15
Analyzing student misconceptions	3.10	.55	2.95	.14	38	-.82	.42	-0.37
Diagnosing level of student understanding	3.05	.71	3.20	.52	37	.74	.46	0.24

Secondary Research Question 2: Are there differences between conditions on teacher perceptions of the professional development? This research question examined the Perceptions of Professional Development Survey to determine if a statistical difference could be found between participants' views of the professional development sessions. Results of the independent samples *t* test indicate that the two conditions did not differ significantly on their perceptions of the training, *t*(38) = .33, *p* = .75. The experimental condition had a total Perceptions of Professional Development score of 34.10 (4.20), and the control condition had the same total score of 34.10 (4.89). Cohen's *d* was 0.

Summary of Results

Analyses of condition differences indicate that the short-term professional development session focused on evaluating student understanding of natural selection had a moderate effect but insufficient power to establish significant group differences. Results also indicate that there was a significant positive impact of the experimental professional development session on identifying the critical elements of student understanding, the number of errors identified across students, and content knowledge related to natural selection. No differences were found between conditions on their self-perceived skill conducting activities related to evaluating student work or their perceptions of the training.

Discussion and Implications

> We continue to know relatively little about what teachers *learn* from professional development, let alone what students learn as a result of changed teaching practices, which is the ultimate measure of standards-based reform efforts. To create excellent programs of professional development, it is necessary to build an empirical knowledge base that links different forms of professional development to both teacher and student learning outcomes. (Fishman, Marx, Best, & Tal, 2003, p. 643)

This study was an empirical examination of two professional development conditions that focused on factors hypothesized to impact teacher outcomes as related to judging student understanding of natural selection, content knowledge of natural selection, and pedagogical content knowledge specifically related to teacher evaluation of student understanding and errors. This research is timely given the increased focus on diagnostic information from formative assessments as part of reform initiatives (Gallagher & Worth, 2008; Pellegrino, Chudowsky, & Glaser, 2001).

It was expected that a focus on students' misconceptions of natural selection and practice analyzing students' work in the experimental condition would increase teacher skill evaluating student understanding and errors related to natural selection. Results indicate that the experimental condition was effective at significantly increasing teacher knowledge and skill diagnosing critical elements related to student understanding of natural selection and their content knowledge related to natural selection, but was not effective at significantly increasing the overall number of embedded components identified and the ratings of student understanding. Additionally, no differences were found between self-perceived ability of skills related to analyzing and assessing student understanding or participants' perceptions of their professional development session.

Interpretation of the Descriptive Statistics

The descriptive statistics indicate that the sample showed a wide range of teaching experience in biology, with only a few participants having middle school teaching experience. Natural Selection Pretest scores suggest that participants were prepared to teach the subject matter. In addition, a majority had received Master's degrees, and attended professional development outside of the school-provided sessions regularly. This suggests that this group was motivated to acquire additional professional knowledge. However, few participants reported having more than a few days of training on the topic of student assessment practices prior to this professional development, indicating that most had limited to no formal training on the topic of this study.

Interpretation of the Inferential Statistics

Key Finding 1: Professional development focused on evaluating student errors and understanding can increase participant expertise evaluating student work. One of the key findings was that there was limited evidence of differences between conditions on the total number of embedded components identified across all students. However, substantial evidence of differences between conditions on the number of critical elements was identified. These differences were often associated with the identification of misconceptions. Critical elements were a subset of the total embedded components and consist mostly of errors (e.g., misconceptions). They were the elements that had been used as justification by experts for their ratings of student understanding. The lack of difference on the total number of embedded components but a significant difference on the identification of critical elements suggest that the experimental condition helped participants develop expertise on how to focus on and distinguish among the importance of elements found within student responses. The research literature on expertise indicates that experts view their domain of expertise differently than non-experts (Chi et al., 1981). In studies of teachers, expert teachers are also better able to weigh the importance of information over non-experts (Carter, Crushing, Sabers, Stein, & Berliner, 1988; Carter, Sabers, Cushing, Pinnegar, & Berliner, 1987

The participants in the experimental condition probably developed these skills because one of the aims of their professional development session was to increase their knowledge of the wide range of student misconceptions. There are many misconceptions related to natural selection, and some misconceptions are more common than others (Bishop & Anderson, 1990; Brumby, 1984; Lawson & Thompson, 1988; Nehm & Reilly, 2007). However, teachers are not always aware of all the misconceptions. Therefore, time at the beginning of the experimental professional development session was spent developing teachers' knowledge of these misconceptions, including the sources of different misconceptions and talking about which misconceptions are the ones most widely held by students. These misconceptions were revisited throughout the rest of the professional development session as participants practiced identifying and discussing student responses. This information and practice may have provided the experimental teachers with valuable insight about important misconceptions found within student responses and may have led to the differences found between conditions.

Moreover, the focus on misconceptions may have helped teachers to confront, alter, or correct their own knowledge of natural selection. Research has suggested that many biology teachers hold misconceptions specifically related to natural selection (Nehm & Schonfeld, 2007). Misconceptions held by teachers would be a major impediment to diagnosing student understanding accurately. For example, if a teacher had a certain misconception, then they would regard that information as correct rather than incorrect. Therefore, a better understanding of natural selection by the experimental teachers as indicated on the Posttest of Natural Selection may have resulted in a deeper insight into the nuances of evaluating student understanding of natural selection.

While significant differences were found between groups on the number of critical elements identified, neither condition identified more than half of the total possible. This suggests that while the experimental condition may have been more effective than the control condition, the experimental condition was not able to develop the skills of participants to the maximum level. This may be the result of the short length of the professional development session. A professional development program sustained over time may lead to a higher skill level of participants in the experimental condition.

Key Finding 2: *Deep analysis of student work can provide opportunities to increase content knowledge.* A second important finding is that the participants in the experimental professional development session had significantly higher Natural Selection Posttest scores than those in the control condition after controlling for the Natural Selection Pretest. The control condition may have been less effective than the experimental condition because the profession development session was similar to lessons the teachers might conduct in their own classrooms. The control professional development session may also have reflected information they learned in college. In other words, control condition participants may have already known the information provided in the control professional development session.

Deeper and more complex content knowledge is often associated with higher order thinking skills (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Therefore, an alternative explanation for the lower content knowledge of the control condition is that the professional development session may not have provided participants the types of activities that would have helped them deepen their content knowledge. Alternatively, in the experimental condition, the process of actively engaging with the content area through the evaluation of student work may have helped participants deepen their knowledge related to the content area. This suggests that professional developments where teachers analyze student work might help to provide the opportunities needed for increased content knowledge.

Key Finding 3: Measuring teacher outcomes related to professional development are important. The research literature advises that professional development should be long-term to be effective (Garet et al., 2001; Yoon et al., 2007). However, these recommendations are typically related to student outcome measures. This study focused on teacher outcome measures.

Results from this study suggest that both short-term professional development sessions were well received by teachers. Evidence from this study also suggested that there was an impact of the experimental professional development session on specific aspects of teacher knowledge and skill related to evaluating student work in less than one full day. Specifically that the number of critical elements participants identified and that the Natural Selection Posttest scores were higher for participants in the experimental condition. If only student outcomes had been measured and no effect found, these effects on the teacher would not have been known. This study is an example of the first step in Desimone's model of professional development (2009). She suggests that first teacher outcomes must be measured to determine if the professional development alters teacher knowledge, beliefs, or practice, and then changes in practice that influence student achievement should be measured. This model is suggested as a way to gain a more comprehensive understanding of the effectiveness of different professional development programs.

Because teacher classroom practice or student outcomes were not measured as part of this study, it is not known whether the professional development had an impact on these aspects. It is possible that a longer term or different professional development would be needed to impact teacher practice or student outcomes.

Limitations

The generalizability of the findings of this study is limited by the sample. Participants in this study were volunteers and received payment for their participation. In addition, professional development sessions were only offered on specific weekends and holidays and participants were required to drive to the location, which limited the number of teachers who were able to volunteer for the study. Evidence from the Perceptions of Professional Development Survey results also suggests that participants in the study were highly enthusiastic about the professional development sessions. These factors may have resulted in a sample that was more motivated to attend professional development than the total population of Los Angeles area high school biology teachers. Therefore, generalizations of results are limited to the sample studied.

Furthermore, findings from this study are limited by the small sample size. Sample size was further reduced when individuals who were non-randomly assigned to a condition were eliminated from analyses. Moderate effect sizes, but non-significant differences, suggest that the sample size could have affected results. Effect size quantifies the magnitude of the difference in means between two conditions, and *p* values indicate the likelihood that the difference found between conditions was due to random chance of sampling. *p* values depend partially upon sample size. Therefore, it is possible to find a statistically significant difference between conditions and a small effect size if the sample is very large. Conversely, it is possible to find non-significant differences between conditions with moderate or large effect sizes if the sample is small. In this study, the sample size was relatively small, but effect size indicates moderate effect sizes. Power analysis indicates that this study may not have had the necessary power to find statistical differences between conditions. Future studies should include larger sample sizes.

Implications for Future Research

Two areas of future research are suggested based on the results of this study. First, results of this study suggest that teachers in the experimental condition gained knowledge and skill related to evaluating student work. However, since the scope of this study was limited to short-term teacher outcomes rather than long-term effects, it is not known if the changes seen in the experimental condition would impact teacher practice. Therefore, one suggestion for future studies is to offer the same professional development sessions, but instead of only measuring immediate teacher outcomes, also measure changes in classroom practice. Changes in classroom practice could be measured with classroom observations, interviews, or surveys both prior to and after the professional development session. If changes in classroom practice were found, a next step could be to determine the impact of the professional development on student achievement. This series of studies offers a way to monitor the impact of the professional development on different outcomes and might help to reveal necessary changes to the professional development if no impact is found on certain outcomes.

A second area of future study is to examine the impact of level of content knowledge on the evaluation of student work. The sample in this study was limited to high school biology teachers. This group had a moderately high level of content knowledge related to natural selection. Investigating middle or elementary school teachers or preservice teachers who have not yet completed a Bachelor’s degree in biology may provide interesting insight into how content knowledge impacts an individual’s ability to evaluate student work. This type of study may also provide information about the minimum levels of content knowledge required to effectively evaluate student work or if short-term professional development would even be effective if prior content knowledge was not high enough.

Implications for the Design of Professional Development

Findings from this study also have implications for the design of professional development. Supporting the existing literature, evidence from this study suggests that professional developments related to evaluating student work must be content specific and focused on limited teaching practices (Desimone et al., 2002; Garet et al., 2001). A very narrow content focus should be taken if the professional development session is especially short. Results also suggest that the time devoted by science teachers to developing their understanding of student misconceptions may be important. For professional development aimed at helping teachers learn how to evaluate student work, it is also important to help teachers learn how to distinguish between important and unimportant errors, which will help them develop more expert-like skills.

Summary

Overall, despite the small sample size, the results of this study provide relatively strong evidence that a short-term professional development for high school biology teachers focused on evaluating student understanding and errors in natural selection can increase both teachers' ability to focus on and identify critical elements of students' understanding and their content knowledge. Findings from this study begin to fill some gaps in the literature on teacher outcomes. Continued research in this area will help to provide a better understanding of the type of knowledge and applications needed to evaluate student errors and understanding. More rigorous studies of professional development will help to determine the most effective ways to provide training to teachers, including approaches to update their knowledge in rapidly changing fields.

References

Baker, E. L., Niemi, D., Herl, H., Aguirre-Munoz, A., Staley, L., & Linn, R. L. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco: Jossey Bass.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*, 536-553.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, *27*, 415-427.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74. doi: 10.1080/0969595980050102

Blank, R. K., de las Alas, N., & Smith, C. (2007). *Analysis of the quality of professional development programs for mathematics and science teachers: Findings from a cross-state study*. Washington, DC: Council of Chief State School Officers.

Bloom, B., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: Longman.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24,* 417-436.

Brookhart, S. M., Moss, C. M., & Long, B.A. (2007). *A cross-case analysis of teacher inquiry into formative assessment practices in six Title I reading classrooms* (CASTL Technical Report Series No. 1-07). Retrieved March 15, 2011 from http://www.duq.edu/castl/_pdf/CASTL_Technical_Report_1_07.pdf

Bruer, J. T. (1993). *Schools for thought.* Cambridge, MA: MIT Press.

Brumby, M. N. (1984). Misconceptions about the concept of natural selection by medical biology students. *Science Education*, *68*, 493-503. doi: 0.1002/sce.3730680412

California State Department of Education. (1990). *Science framework for California public schools.* Sacramento, CA: California Department of Education.

Carter, K., Sabers, D. S., Cushing, K. S., Pinnegar, S., & Berliner, D. C. (1987). Processing and using information about students: A study of expert, novice and postulant teachers. *Teaching and Teacher Education*, *3*, 147-157.

Carter, K., Cushing, K. S., Sabers, D. S., Stein, P., & Berliner, D. C. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 3*, 29-31.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). State of the profession: Study measures status of professional development. *Journal of Staff Development, 30*(2), 42-44.

Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38,* 181–199.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24,* 81–112.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28,* 1-11.

Fishman, B., Marx, R., Best, S., & Tal, R. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education*, *19*(6), 643-658.

Gallagher, C., & Worth, P. (2008). *Formative assessment policies, programs, and practices in the Southwest Region* (Research Report No. 041). Retrieved from the U.S. Department of Education, Institute of Education Sciences website: http://ies.ed.gov/ncee/edlabs

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915-945.

Greene, E. D., Jr. (1990). The logic of university students' misunderstanding of natural selection. *Journal of Research in Science Teaching*, *27,* 875-885.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Hashweh, M. Z. (1987). Effects of subject-matter knowledge in the teaching of biology and physics. *Teaching and Teacher Education, 3*, 109–120.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, *89,* 140-145.

Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan*, *90*(7), 470-476.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. *Journal of Mathematics Teacher Education, 7*, 203-235.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and call for research. *Educational Measurement: Issues and Practice, 30,* 28-37.

Lawson, A. E., & Thompson, L. D. (1988). Formal reasoning ability and misconceptions concerning genetics and natural selection. *Journal of Research in Science Teaching, 25*, 733–746.

Meisels, S., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement scores. *Educational Policy Analysis Archives, 11*(9). Retrieved March 24, 2011, from http://epaa.asu.edu/epaa/v11n9/

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125-145.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *BioScience*, *57*, 262-272.

Nehm, R. H., & Schonfeld, I. S. (2007). Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? *Journal of Science Teacher Education, 18*, 699-723.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington, DC: National Academy Press.

Phelan, J., Choi, K., Vendlinski, T., Baker, E. L., & Herman, J. L. (2009). *The effects of POWERSOURCE intervention on student understanding of basic mathematical principles* (CRESST Report No. 763). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Quint, J., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) program in Boston elementary schools.* New York: MDRC.

Rutherford, J. F., & Ahlgren, A. (1989). *Science for all Americans*. Washington, DC: Oxford University Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sato, M., Chung, R. R., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal, 45,* 669-700.

Schneider , M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skill in formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 251-276). New York: Routledge.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., … Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*, 295-314. doi:10.1080/08957340802347647

Shepard, L. (2005, October). *Formative assessment: Caveat emptor*. Paper presented at the ETS Invitational Conference, New York, NY.

Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, *20*, 163-183.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4-14.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, *52*, 170-194.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18,* 23–27. doi: 10.1111/j.1745-3992.1999.tb00004.x

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education*, *11*(1), 49-65.

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, *35*, 673-695.

van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education*, *24*, 244-276.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarlos, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement. issues & answers* (REL 2007-No. 033). Retrieved from Regional Education Laboratory website: http://eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED498548&ERICExtSearch_SearchType_0=no&accno=ED498548

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., … Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, *21*, 335-359.

APPENDIX A:

ASR TASK

CRESST
National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Analyzing Student Responses

Directions

In this task, you are provided with four student responses to writing prompts. For each student lettered A through D in the top right corner of the page, you will answer five questions. The first three questions ask you to identify (a) the correct concepts or ideas, (b) the misconceptions or incorrect concepts or idea, and (c) the omitted concepts or ideas in each response about natural selection. The last two are multiple-choice questions about each response.

On the final page of the task, you will rank the responses given by Students A-D in order from most scientifically correct response to least correct. You will also be asked four general questions about this task.

Notes

- Some writing prompts are different from others, so please be sure to carefully read the entire prompt and response.
- For questions 1-3 of each student's response please use bullet points or numbers to clearly distinguish multiple entries.
- Please write legibly and avoid abbreviations in your responses.

Time

Please take no more than 45 minutes to complete this task.

Student A

Writing prompt:

Blackfin tunas are a fast swimming predatory fish only found in the western Atlantic from Cape Cod to Brazil. They are able to swim up to 60 miles per hour when chasing after prey. How would a biologist explain how the ability to swim fast evolved in Blackfin tunas, assuming their ancestor could only swim 20 miles per hour?

Student response:

Blackfin tunas are able to swim faster than their ancestors because of survival of the fittest. Survival of the fittest is when the strong ones live and the weak ones die. Here, maybe the prey got faster than 20 mph, and the tunas couldn't catch anything to eat, so then they had babies with stronger fin muscles so the babies could catch the prey. Then the prey got faster again, so the tunas had babies with even stronger fins. This happened over and over until now they can swim over 60 mph. If the tunas hadn't of had babies with stronger fins, they would have died because they couldn't have caught the prey.

1. Based on the response, what concepts **related to natural selection** does this student **understand**? *If none write "none" below.*

2. What **misconceptions or scientifically inaccurate concepts** **<u>related to natural selection</u>** are reflected in this student's response? *If none write "none" below.*

3. Sometimes student responses are incomplete. It is not that they are necessarily incorrect, but they have omitted essential information. Are there any **omitted concepts** **<u>related to natural selection</u>** that you would have liked to have seen in this response? *If none write "none" below*

MULTIPLE-CHOICE QUESTIONS

4. **What is your overall rating of this student response?** ***(select/mark only one answer)***
   - □ Strong: Strong understanding of natural selection evident. Misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection very minor or nonexistent.
   - □ Good: Good understanding of natural selection evident. Mostly correct conceptions related to natural selection identified. Some evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also identified.
   - □ Basic: Basic understanding of natural selection evident. Some understanding of natural selection identified. However, significant evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also present.
   - □ Poor: Poor understanding of natural selection evident. Major misconceptions, inaccuracies, and/or omitted concepts about natural selection are identified. Evidence of correct conceptions about natural selection very limited or nonexistent.

5. **This student requires:** ***(select/mark only one answer)***
   - □ substantial instruction on all/most concepts related to natural selection.
   - □ substantial instruction on certain concepts related to natural selection
   - □ minor instruction on concepts related to natural selection.
   - □ no instruction on concepts related to natural selection.

Please write other comments about this student's response below (optional).

Continue to next student response

Student B

Writing prompt:

There is a species of salamander that lives deep inside caves in Texas. These cave salamanders are all blind. How would a biologist explain how the blind cave salamanders evolved from ancestors that could see?

Student response to prompt:

I think that maybe the cave salamanders in Texas are blind even though their ancestors could see because a long, long time ago, the salamanders that could see (strong ones) did not have any children salamanders which meant that only the blind salamanders (weak ones) were making babies. Because only blind salamanders were making babies, all the babies were blind too. And being blind wouldn't matter anyways if you live in a cave because a cave is already dark, so even if you could see, it wouldn't matter because it is dark in the cave and you couldn't see anything anyways.

1. Based on the response, what concepts **<u>related to natural selection</u>** does this student **understand**? *If none write "none" below.*

2. What **misconceptions or scientifically inaccurate concepts** **<u>related to natural selection</u>** are reflected in this student's response? *If none write "none" below.*

3. Sometimes student responses are incomplete. It is not that they are necessarily incorrect, but they have omitted essential information. Are there any **omitted concepts** **<u>related to natural selection</u>** that you would have liked to have seen in this response? *If none write "none" below.*

MULTIPLE-CHOICE QUESTIONS

4. **What is your overall rating of this student response?** ***(select/mark only one answer)***
   - □ Strong: Strong understanding of natural selection evident. Misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection very minor or nonexistent.
   - □ Good: Good understanding of natural selection evident. Mostly correct conceptions related to natural selection identified. Some evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also identified.
   - □ Basic: Basic understanding of natural selection evident. Some understanding of natural selection identified. However, significant evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also present.
   - □ Poor: Poor understanding of natural selection evident. Major misconceptions, inaccuracies, and/or omitted concepts about natural selection are identified. Evidence of correct conceptions about natural selection very limited or nonexistent.

5. **This student requires:** ***(select/mark only one answer)***
   - □ substantial instruction on all/most concepts related to natural selection.
   - □ substantial instruction on certain concepts related to natural selection
   - □ minor instruction on concepts related to natural selection.
   - □ no instruction on concepts related to natural selection.

Please write other comments about this student's response below (optional).

Continue to next student response

Student C

Writing prompt:

Today there are many species of finches on the Galapagos Islands that descended from the same common ancestor. One of the differences between finches is their beak size. Beak size varies between islands. Different islands can also have different environments and food sources. Explain why the finches on different islands might have different beaks.

Student response to prompt:

Finches on the different islands have different beaks because the environments and food to eat are so different on each island. Because animals have to adapt to their environment to survive, the finches had to evolve to the habitat on their island to survive. For example, on some islands they grew stronger beaks so they could eat the really hard nuts. On other islands they all grew long skinny beaks so they could eat from tube flowers. If a family of finches moved from one island to an island that was different to the island they came from, they would evolve to the new island so they wouldn't die.

1. Based on the response, what concepts **<u>related to natural selection</u>** does this student **understand**? *If none write "none" below.*

2. What **misconceptions or scientifically inaccurate concepts <u>related to natural selection</u>** are reflected in this student's response? *If none write "none" below.*

3. Sometimes student responses are incomplete. It is not that they are necessarily incorrect, but they have omitted essential information. Are there any **omitted concepts <u>related to natural selection</u>** that you would have liked to have seen in this response? *If none write "none" below.*

MULTIPLE-CHOICE QUESTIONS

4. **What is your overall rating of this student response?** ***(select/mark only one answer)***
   - □ Strong: Strong understanding of natural selection evident. Misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection very minor or nonexistent.
   - □ Good: Good understanding of natural selection evident. Mostly correct conceptions related to natural selection identified. Some evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also identified.
   - □ Basic: Basic understanding of natural selection evident. Some understanding of natural selection identified. However, significant evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also present.
   - □ Poor: Poor understanding of natural selection evident. Major misconceptions, inaccuracies, and/or omitted concepts about natural selection are identified. Evidence of correct conceptions about natural selection very limited or nonexistent.

5. **This student requires:** ***(select/mark only one answer)***
   - □ substantial instruction on all/most concepts related to natural selection.
   - □ substantial instruction on certain concepts related to natural selection
   - □ minor instruction on concepts related to natural selection.
   - □ no instruction on concepts related to natural selection.

Please write other comments about this student's response below (optional).

Continue to next
student response

Student D

Writing prompt:

Cheetahs are able to run faster than 60 miles per hour when chasing prey. How would a biologist explain how the ability to run fast evolved in cheetahs, assuming their ancestor could run only 20 miles per hour?

Student response to prompt:

Natural selection is about fitting with your environment so you can eat and have babies. Slow baby cheetahs, probably had a hard time catching food and died. Well, actually the slow baby cheetahs were probably okay while if they were getting food from their moms but didn't live long after they had to leave their moms. The faster cheetahs were able to catch enough food and live on their own without their mom, so they lived longer than the slow cheetahs. So, cheetahs have evolved stronger legs overtime that let them run faster, catch more food, and not die. And stronger legs made them more fit to their environment.

1. Based on the response, what concepts **related to natural selection** does this student **understand**? *If none write "none" below.*

2. What **misconceptions or scientifically inaccurate concepts** **<u>related to natural selection</u>** are reflected in this student's response? *If none write "none" below.*

3. Sometimes student responses are incomplete. It is not that they are necessarily incorrect, but they have omitted essential information. Are there any **omitted concepts** **<u>related to natural selection</u>** that you would have liked to have seen in this response? *If none write "none" below.*

MULTIPLE-CHOICE QUESTIONS

4. **What is your overall rating of this student response?** ***(select/mark only one answer)***

   - □ Strong: Strong understanding of natural selection evident. Misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection very minor or nonexistent.
   - □ Good: Good understanding of natural selection evident. Mostly correct conceptions related to natural selection identified. Some evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also identified.
   - □ Basic: Basic understanding of natural selection evident. Some understanding of natural selection identified. However, significant evidence of misconceptions, scientifically inaccurate, and/or omitted concepts related to natural selection also present.
   - □ Poor: Poor understanding of natural selection evident. Major misconceptions, inaccuracies, and/or omitted concepts about natural selection are identified. Evidence of correct conceptions about natural selection very limited or nonexistent.

5. **This student requires:** ***(select/mark only one answer)***

   - □ substantial instruction on all/most concepts related to natural selection.
   - □ substantial instruction on certain concepts related to natural selection
   - □ minor instruction on concepts related to natural selection.
   - □ no instruction on concepts related to natural selection.

Please write other comments about this student's response below (optional).

Continue to next page

STUDENT RANKINGS

Please rank the student responses with “1” being the highest/most scientifically correct response and “4” being the lowest/least scientifically correct response. You can only give one student each ranking (e.g. you cannot have two students with a rank of “3”).

_____ Student A

_____ Student B

_____ Student C

_____ Student D

Please explain why you ranked the students in this order.

SURVEY QUESTIONS

(please select/mark only one answer per question)

1. **As compared to how I review my own students' understanding of topics, this was:**

   □ Very similar

   □ Similar

   □ Different

   □ Very different

2. **Completing this packet was:**

   □ Very difficult

   □ Difficult

   □ Easy

   □ Very easy

3. **I see value in reviewing student work in this way.**

   □ Strongly agree

   □ Agree

   □ Disagree

   □ Strongly disagree

4. **I do not have time to review student work in this way.**

   □ Strongly agree

   □ Agree

   □ Disagree

   □ Strongly disagree

Comments (optional)

APPENDIX B:

NATURAL SELECTION PRETEST

1. Imagine that biologists discover a new species of woodpecker that lives in isolation on some secluded island. These woodpeckers have, on average, a 1.0 inch beak, and their only food source is a tree-dwelling insect that lives, on average, 1.5 inches under the tree bark. Compared to its parents, the offspring of any two woodpeckers will grow to have:

A) A longer beak.

B) A shorter beak.

C) Either a longer or shorter beak; neither outcome is more likely.

Why? Please explain your answer below

2. The biologists clip the wing feathers of some of the birds, rendering them unable to fly. Compared to the offspring of the other woodpeckers, the offspring of those with clipped wings will be born with:

A) Longer wing feathers.

B) Shorter wing feathers.

C) Either longer or shorter wing feathers; neither outcome is more likely.

Why? Please explain your answer below

3. Suppose that a pair of woodpeckers migrates to a different island with fewer trees and more wind. As a consequence of flying in a windier environment, both woodpeckers develop stronger wing muscles. Compared to the offspring of the woodpeckers on the original island, the offspring of these two woodpeckers will be born with:

A) Stronger wing muscles.

B) Weaker wing muscles.

C) Either stronger or weaker wing muscles; neither outcome is more likely.

Why? Please explain your answer below

4. Corn you buy in the store is an entirely artificial food. Over a period of thousands of years, Native Americans purposefully transformed maize through special cultivation techniques, modifying corn from a wild grass (Teosinte) which grew in Central America 7,000 years ago. In contrast to modern maize, which yields hundreds of plump kernels per cob, each Teosinte plant yielded a handful of small, hard kernels.

Please rank the following 6 factors on the degree of relevance to domestication. Each rating may only be used once.

A) The degree of similarity among plants of the same generation.

B) The average amount of time each plant was exposed to direct sunlight.

C) The preferences of those who decided which kernels to plant.

D) The fertility of the soil in which the kernels were planted.

E) The average rainfall per year.

F) The percentage of each crop used to breed the next generation.

Answer here- write letter of factors from above next to rank number below:

Rank 1 (most relevant) =

Rank 2 =

Rank 3 =

Rank 4=

Rank 5=

Rank 6 (least relevant) =

5. Would it be possible to cultivate corn back into a plant like Teosinte?

A) Yes

B) No

Why or why not?

6. Imagine you live in a developing country in an area increasingly plagued by pollution. Every year the trees in your part of the country have become darker from the soot and ash that has gathered on their bark. Interestingly, you have noticed that some of the beetles who live on those trees are a dark color as well.

Assuming that this darker coloration is now adaptive to the beetles, how might a change in their environment have brought about a change in their color?

Now imagine that you are a biologist intent on studying these beetles. You and your colleagues by chance gathered a random sample of these beetles in the year 2000. If a random sample of beetles is gathered every 25 years over the course of a century, what range of coloration would you expect to find at each point in time assuming the pollution does not get any better?

INSTRUCTIONS FOR QUESTION BELOW: Three choices are given below. Select the choice that most closely represents the range of coloration you could expect to find over 100 years.

Note: There are 5 possible colors of beetles represented below. 1 = lightest 5= darkest

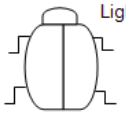
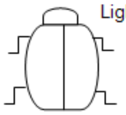
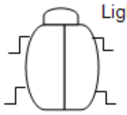
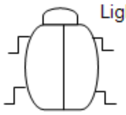
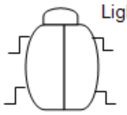
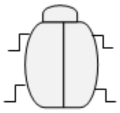
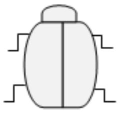
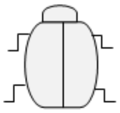
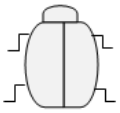
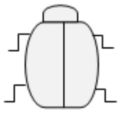
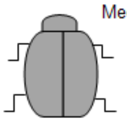
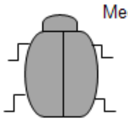
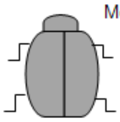
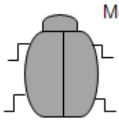
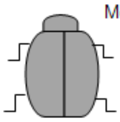
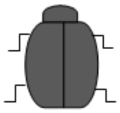
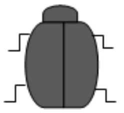
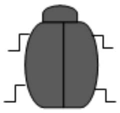
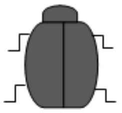
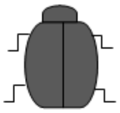
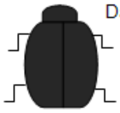
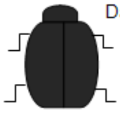
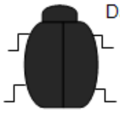
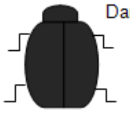
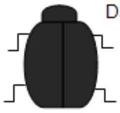
7. Which of the answer choices BELOW (see pictures below) best matches the range of coloration that you would expect to find if you collected a random sample of beetles every 25 years in this environment?

A) Answer choice A

B) Answer choice B

C) Answer choice C

Answer choice A

2000	Lightest (1) 	Lightest (1) 	Lightest (1) 	Lightest (1) 	Lightest (1) 
2025	Light (2) 	Light (2) 	Light (2) 	Light (2) 	Light (2) 
2050	Medium (3) 	Medium (3) 	Medium (3) 	Medium (3) 	Medium (3) 
2075	Dark (4) 	Dark (4) 	Dark (4) 	Dark (4) 	Dark (4) 
2100	Darkest (5) 	Darkest (5) 	Darkest (5) 	Darkest (5) 	Darkest (5) 

Answer choice B

2000	Lightest (1)	Lightest (1)	Lightest (1)	Lightest (1)	Darkest (5)
2025	Lightest (1)	Lightest (1)	Lightest (1)	Darkest (5)	Darkest (5)
2050	Lightest (1)	Lightest (1)	Darkest (5)	Darkest (5)	Darkest (5)
2075	Lightest (1)	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)
2100	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)

Answer choice C

2000	Lightest (1)	Lightest (1)	Light (2)	Medium (3)	Medium (3)
2025	Lightest (1)	Light (2)	Medium (3)	Medium (3)	Dark (4)
2050	Light (2)	Medium (3)	Medium (3)	Dark (4)	Dark (4)
2075	Medium (3)	Medium (3)	Dark (4)	Dark (4)	Darkest (5)
2100	Medium (3)	Dark (4)	Dark (4)	Darkest (5)	Darkest (5)

8. Please explain your answer choice on the beetle question.

***9. Please explain why you did not select the other answer choices.**

***10. Pronghorn antelopes are often cited as the second-fastest land animals on Earth. They are able to run faster than 50 miles per hour when running from predators and can sustain high speeds for long periods of times. How would you explain how the ability to run fast evolved in pronghorn antelopes, assuming their ancestor ran much slower?**

***12. If there is variation, differential reproduction, and heredity, there will be evolution by natural selection as an outcome. It is a simple as that.**

A) I agree with the statement above

B) I partially agree with the statement above

C) I disagree with the statement above

***12. Please explain why you agree, partially agree, or disagree with the statement above.**

***13. Natural selection is not "survival of the fittest," but rather "reproduction of the fittest."**

A) I agree with the statement above

B) I partially agree with the statement above

C) I disagree with the statement above

***14. Please explain why you agree, partially agree, or disagree with the statement above.**

END OF SURVEY- THANK YOU!

APPENDIX C:

CONTENT KNOWLEDGE OPEN-ENDED QUESTIONS SCORING GUIDE

Understanding of variation within species (subtopic 1 of 3)

Topic	Explanation	Scoring	Example Text	
Variation within species	People either talk about variation as existing within a population (correct) or not (incorrect). This is often associated with a need/want to change (incorrect) or all individuals changing together.	0- Does not talk about variation existing within a population. Sometimes talks about all the organisms having the same trait/phenotype.	0	“The **birds** had strong and thick beaks” (but often doesn’t even say this much, just doesn’t mention it)
		1- Implicitly/indirectly talks about or implies variation existing within a population without specifically identifying both groups. Usually they will only identify one group.	1	“Birds with **longer beaks**…” (but no mention of shorter beaks) OR “Over time, the birds with **shorter beaks**…” (but no mention of other size of beaks)
		2- Explicitly talks about the variation that exists within a population or talks about both groups. Identifies both traits/groups or that there is a range.	2	“Some birds have **long beaks** and some have a **short beak**.” OR “There must have been a **wide range of beak size to begin with**. Those with long beaks….” OR “The birds with **long beaks** will be able to get the food deep in the wood, but the ones with the **shorter beaks** will not”
		3- Talks about ONE variation that exist AND describes the source of variation as mutations and/or genetic recombination.	3	“Some birds have **long beaks** is likely due to multiple **mutations over time**.” OR “ **Genetic recombination** caused there to be **long beaks**.”
		4- Talks about BOTH variations that exist AND describes the source of variation as mutations and/or genetic recombination.	4	“Some birds have **long beaks** and some have a **short beak.** This variation is likely due to multiple **mutations over time**.” OR “ **Genetic recombination** caused there to be a **wide range of beak size**.”

Understanding of differential reproduction (subtopic 2 of 3)

Topic	Explanation	Scoring	Example Text	
Differential reproduction	Differential reproduction rather than survival is responsible for natural selection. Just because an organism survives doesn’t mean that the organism will reproduce and pass on its genes.	0- Does not attribute cause of natural selection to survival or differential reproduction.	0	
		1- Attribute natural selection to survival rather than reproduction/heredity or implies trait helped with survival.	1	“Because the birds with thicker beaks were able to eat more, they **survived**” OR “The birds with longer wings were faster and able to **fly away from predators** better than shorter wing birds” (implies survival)
		2- Attributes cause of natural selection to differential reproduction, but only talks about one variation.	2	“Because the birds with thicker beaks were **more likely to survive long enough to reproduce**, so that trait got passed on.” (they don’t necessarily have to say survive, but only reproduce) OR “One bird had a variation that made it have longer wings. Overtime, **birds with faster speeds** due to longer wings were surviving better and **reproducing because they could out fly predators.**”
		3- Attributes cause of natural selection to differential reproduction, talks about impact on both/multiple variations.	3	“Because the birds with **thicker beaks were** more likely to survive long enough to reproduce, so that trait got passed on **and the thinner beak** didn’t survive, so it was more likely to die before reproducing.” “**Birds with shorter wings** will more likely be caught by predators and their genes will be removed from the gene pool. **Birds with longer wings** are more likely to outrun predators and pass on those genes to their offspring.”

Understanding of results of inheritance within the population (subtopic 3 of 3)

Topic	Explanation	Scoring	Example Text	
Results of inheritance within the population	It is important to notice how people talk about the results of inheriting traits and how that trait can change in the population over time. Many times people do not talk much about how the traits of a population change.	0- Does not talk about the results of passing on traits or uses misconception about it.	0	“The hard beak birds reproduced to passed on traits to the next generation” (no info about results of passing on traits) “Needing to fly faster resulted in birds developing longer wings” (misconception)
		1- Talks about the results of one pair passing on traits/genes in one generation. NOTE: if they only talk about passing on traits, then it is a “0”. They must talk about the results	1	“The hard beak birds survived and passed on their genes to their offspring **contributing more offspring to the next generation”**” OR “The birds with longer wings were able to outfly the predators and reproduce. **Resulting in this trait being more frequent in the the next generation.”**
		2- Talks about the results of passing on traits/genes in the population, and vaguely talks about how it would be over multiple generations.	2	“The hard beak birds survived and passed on their genes to their offspring. **Eventually, this would result in a population with longer wings**” “The birds with longer wings were able to out fly the predators and this made this trait more frequent in the gene pool for the next generation. **Over generations, the birds with longer wings would be more common**.”
		3- Talks about the results of passing on traits/genes in the population over multiple generations. Is specific about how that happens.	3	“The hard beak birds survived and passed on their genes to their offspring. **Over time, the number of hard beak birds in the population increased because those were the birds surviving and reproducing more than the other birds. Those birds would then reproduce more often if they were selected for. Eventually, the percentage of hard beak birds would be greater than soft beak birds”** OR “The birds with longer wings were able to out fly the predators and this made this trait more frequent in the gene pool for the next generation. **Among each generation the faster birds with the longest wings were more likely to pass on their traits and the short wing birds likely got eaten by predators. Over many generations, the birds wings got longer as a result.**

APPENDIX D:

NATURAL SELECTION POSTTEST

CRESST

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Natural Selection Posttest

Directions

Please respond to all of the questions in this booklet and write legibly.

Time

Please take no more than 25 minutes to complete this task.

MULTIPLE-CHOICE QUESTIONS (Please select/mark only one answer per question)

1. **A new species of fish has been found living near the bottom of freshwater lakes in Central Africa. This species of fish has a snout on average of 1.0 inches long. Their only food source lives in the mud at the bottom of the lake that is, on average, 1.5 inches under a layer of pebbles. Compared to its parents, the offspring of any two of fish will grow to have:**

   □ A longer snout.
   □ A shorter snout.
   □ Either a longer or shorter snout; neither outcome is more likely.

Please explain your answer choice.

2. **To study these fish, biologists clip the tail fins of some of the fish. This causes the fish with clipped tails to swim slower than fish with unclipped tails. Compared to the offspring of fish with unclipped tail fins, the offspring of fish with clipped fins will be born with:**

   □ Longer tails.
   □ Shorter tails.
   □ Either a longer or shorter tail; neither outcome is more likely.

Please explain your answer choice.

3. **Suppose that a pair of these fish becomes isolated in a section of the lake that has a stronger current than the other areas of the lake. As a consequence of swimming in an area with a stronger current, this pair of fish develops stronger muscles. Compared to the offspring of fish in the other areas of the lake, the offspring of the pair of fish in the section of the lake with a strong current will be born with:**

   □ Stronger muscles than fish in the other areas of the lake.
   □ Weaker muscles than fish in the other areas of the lake.
   □ Either stronger or weaker muscles than fish in the other areas of the lake; neither outcome is more likely.

Please explain your answer choice.

Continue to next page

OPEN-ENDED QUESTIONS

4. There is a species of fish that lives deep inside freshwater caves of northern Mexico. These cave fish are all blind. **Explain how the blind cave fish could have evolved through natural selection from ancestors that could see.**

5. The peregrine falcon is the fastest living creature, reaching speeds of at least 124 mph and possibly as much as 168 mph when swooping from great heights during territorial displays or while catching prey (birds) in midair. **How would a biologist explain how the ability to fly so fast evolved in peregrine falcons through natural selection, assuming their ancestor could not fly nearly as fast?**

6. Slash and burn techniques are increasingly being used in South American forests to clear land and create fields. This technique increased the amount of pollution in the area and causes soot and ash to land on the surrounding areas.

The Manuripa Amazon National Park, a small protected area in eastern Peru, borders on one of the largest slash and burn areas in South America. Biologists have noticed that in the last 10 years, many trees and plants in Manuripa are becoming covered in soot and darker over time. Biologists have also noticed that butterflies in this area, on average, have become darker in the last 10 years as well.

Assuming that this darker coloration is now adaptive to the butterflies, how might a change in their environment have brought about a change in their color?

7. Now imagine that you are a biologist intent on studying the butterflies described on the previous page. You and your colleagues by chance gathered a random sample of these butterflies in the year 2000. If a random sample of butterflies is gathered every 25 years over the course of a century, what range of coloration would you expect to find at each time point assuming the pollution does not get any better or worse?

Note: There are 5 possible colors of butterflies represented below.
1 = lightest
5= darkest

Which of the answer choices (below and on next page) best matches the range of coloration that you would expect to find if you collected a random sample of butterflies every 25 years in this environment?

□ Answer choice A
□ Answer choice B
□ Answer choice C

Answer choice A

2000	Lightest (1)	Lightest (1)	Light (2)	Medium (3)	Medium (3)
2025	Lightest (1)	Light (2)	Medium (3)	Medium (3)	Dark (4)
2050	Light (2)	Medium (3)	Medium (3)	Dark (4)	Dark (4)
2075	Medium (3)	Medium (3)	Dark (4)	Dark (4)	Darkest (5)
2100	Medium (3)	Dark (4)	Dark (4)	Darkest (5)	Darkest (5)

Answer choice B

2000	Lightest (1)	Lightest (1)	Lightest (1)	Lightest (1)	Lightest (1)
2025	Light (2)	Light (2)	Light (2)	Light (2)	Light (2)
2050	Medium (3)	Medium (3)	Medium (3)	Medium (3)	Medium (3)
2075	Dark (4)	Dark (4)	Dark (4)	Dark (4)	Dark (4)
2100	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)

Answer choice C

2000	Lightest (1)	Lightest (1)	Lightest (1)	Lightest (1)	Darkest (5)
2025	Lightest (1)	Lightest (1)	Lightest (1)	Darkest (5)	Darkest (5)
2050	Lightest (1)	Lightest (1)	Darkest (5)	Darkest (5)	Darkest (5)
2075	Lightest (1)	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)
2100	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)	Darkest (5)

Explain why you selected the answer you did for question 7 (on previous pages)

Explain why the other responses were wrong for question 7 (on previous pages)

SURVEY QUESTIONS

Rate your agreement/disagreement to the following statements:
(circle the appropriate number for each statement)

	Strongly agree	**Agree**	**Disagree**	**Strongly disagree**	**N/A**
1. These questions would be appropriate for regular biology students.	1	2	3	4	
2. These questions would be appropriate for Honors biology students.	1	2	3	4	
3. These questions would be appropriate for A.P biology students.	1	2	3	4	
4. I found this test easy to complete.	1	2	3	4	
5. I use similar questions with my regular biology students.	1	2	3	4	N/A
6. I use similar questions with my Honors biology students.	1	2	3	4	N/A
7. I use similar questions with my A.P biology students.	1	2	3	4	N/A
8. Completing this test took a lot of effort/thought.	1	2	3	4	
9. Many biology teachers would have difficulty answering these questions.	1	2	3	4	
10. Biology majors (undergraduate) would have difficulty answering these questions.	1	2	3	4	

Comments (optional)

APPENDIX E:

BACKGROUND SURVEY

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Background Survey

Directions

Please answer the following questions as completely and honestly as you can.

Time

Please take no more than 10 minutes to complete this survey.

TEACHING EXPERIENCE

1. Counting this year as one full year, how many total years teaching experience do you have? _____

2. Counting this year as one full year, how many years have you taught high school biology? _____

3. Counting this year as one full year, how many years have you taught English language learners? _____

4. Have you also taught middle school life science?

   □ No

   □ Yes, how many years? ______

5. Besides biology, what other topics of science have you taught? Check all that apply and write the number of years you have taught that topic.

   □ Chemistry, ________ years

   □ Physics, ________ years

   □ Earth science, ________ years

   □ Environmental science, ________ years

   □ Other science topic(s), please specify and indicate the number of years:____

6. What level(s) of biology have you taught? Check all that apply and write how many years you have taught that level of biology.

   □ Honors high school biology, ________ years

   □ Regular high school biology, ________ years

   □ AP biology, ________ years

   □ Remedial level, ________ years

   □ English language learners/sheltered classes, ________ years

7. What level of biology would you consider yourself to be the most proficient at teaching? Please check only one.

   □ Honors high school biology

   □ Regular high school biology

   □ AP biology

   □ Remedial level

TEACHER CERTIFICATION

8. Which of the following best describes your teaching credential status?

   □ I am currently credentialed by California to teach biology

   □ I am in the process of becoming credentialed by California to teach biology

   □ I am not credentialed by California to teach biology and am not seeking to become credentialed at this time

   □ Other, please specify ____

9. As part of your teacher certification program, have you or will you take a course on the assessment of students? **Note: A course on assessment would include topics such as the different methods of testing students, how to make tests/quizzes, validity and reliability of tests/quizzes, etc.

   □ Yes, the entire course was on assessment

   □ Yes, part of a course was on assessment

   □ No

   □ I don’t know

COLLEGE DEGREES

10. What was your undergraduate college major?

□ Biology (this includes sub-specialties such as microbiology, molecular biology, ecology, etc.)

□ Other science degree (chemistry, physics, etc.)

□ Science education degree (from the education department and specializing in science)

□ Other education department degree, please specify ________________________________

□ Other degree, please specify ____

11. What college did you attend for your undergraduate degree? ___________________________

12. What was your approximate college GPA? ________

13. Have you completed a Master's degree?

□ No □ Yes

If yes, please check the specialty below:

□ Biology (this includes sub-specialties such as microbiology, molecular biology, ecology, etc.)

□ Other science Master's degree (chemistry, physics, etc.)

□ Science education Master's degree (from the education department and specializing in science)

□ Other education department Master's degree, please specify __________________________

□ Other Master's degree, please specify ____

14. What college did you attend for your Master's degree, if applicable? _____________________

15. Have you completed a Ph.D. or Ed.D?

□ No □ Yes

If yes, please check the specialty below:

□ Biology (this includes sub-specialties such as microbiology, molecular biology, ecology, etc.)

□ Other science department Ph.D. (chemistry, physics, etc.)

□ Science education Ph.D. or Ed.D (from the education department and specializing in science)

□ Other education department Ph.D. or Ed.D, please specify __________________________

□ Other Ph.D. or Ed.D please specify ____

16. What college did you attend for your Ph.D. or Ed.D., if applicable? _____________________

CLASSROOM ACTIVITIES

How often do you participate in the following activities?

	Never	A few times a year	A few times a semester	A few times a month	About once a week	Daily or almost daily
17. Creating tests/quizzes	1	2	3	4	5	6
18. Administering district-created tests	1	2	3	4	5	6
19. Attending school/district-sponsored professional development	1	2	3	4	5	6
20. Attending non-school/district-sponsored professional development	1	2	3	4	5	6
21. Providing written feedback to students on assignments	1	2	3	4	5	6
22. Assigning full lab reports	1	2	3	4	5	6
23. Assigning long writing assignments (e.g. essays) other than lab reports	1	2	3	4	5	6
24. Conducting formative assessments of students	1	2	3	4	5	6
25. Analyzing student work for misconceptions	1	2	3	4	5	6
26. Using direct instruction	1	2	3	4	5	6
27. Conducting lab activities	1	2	3	4	5	6
28. Leading group discussions	1	2	3	4	5	6

Rate your skill at the following classroom activities:

	Highly unskilled	Unskilled	Skilled	Highly skilled
29. Creating tests/quizzes	1	2	3	4
30. Giving students feedback that will improve their understanding of the topic	1	2	3	4
31. Creating formative assessments	1	2	3	4
32. Analyzing student work for misconceptions	1	2	3	4
33. Diagnosing the level of student understanding of a topic	1	2	3	4

Continue to last page

BACKGROUND INFORMATION

34. Ethnicity (choose only one):

□ Biracial / multiethnic
□ African-American
□ Asian or Pacific Islander
□ Hispanic / Latino/a
□ Native-American
□ White, non-Hispanic
□ Other ______________________

35. Gender: □ Male □ Female

THANK YOU FOR COMPLETING THIS SURVEY!

APPENDIX F:

PROFESSIONAL DEVELOPMENT SURVEY

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Professional Development Survey

For each question, circle the number that shows how you feel.

This professional development was:	Strongl y agree	Agree	Disagree	Strongly disagree
1. …interesting	1	2	3	4
2. …organized	1	2	3	4
3. …helpful	1	2	3	4
4. …boring	1	2	3	4
5. …informative	1	2	3	4

Please rate how much you agree/disagree with each statement below.

	Strongly agree	Agree	Disagree	Strongly disagree
1. I learned about natural selection from this training.	1	2	3	4
2. This training wasn’t worth the amount of time it took.	1	2	3	4
3. I would recommend this training to a colleague.	1	2	3	4
4. I will use the examples from this training in my classroom.	1	2	3	4
5. I enjoyed participating in this training.	1	2	3	4

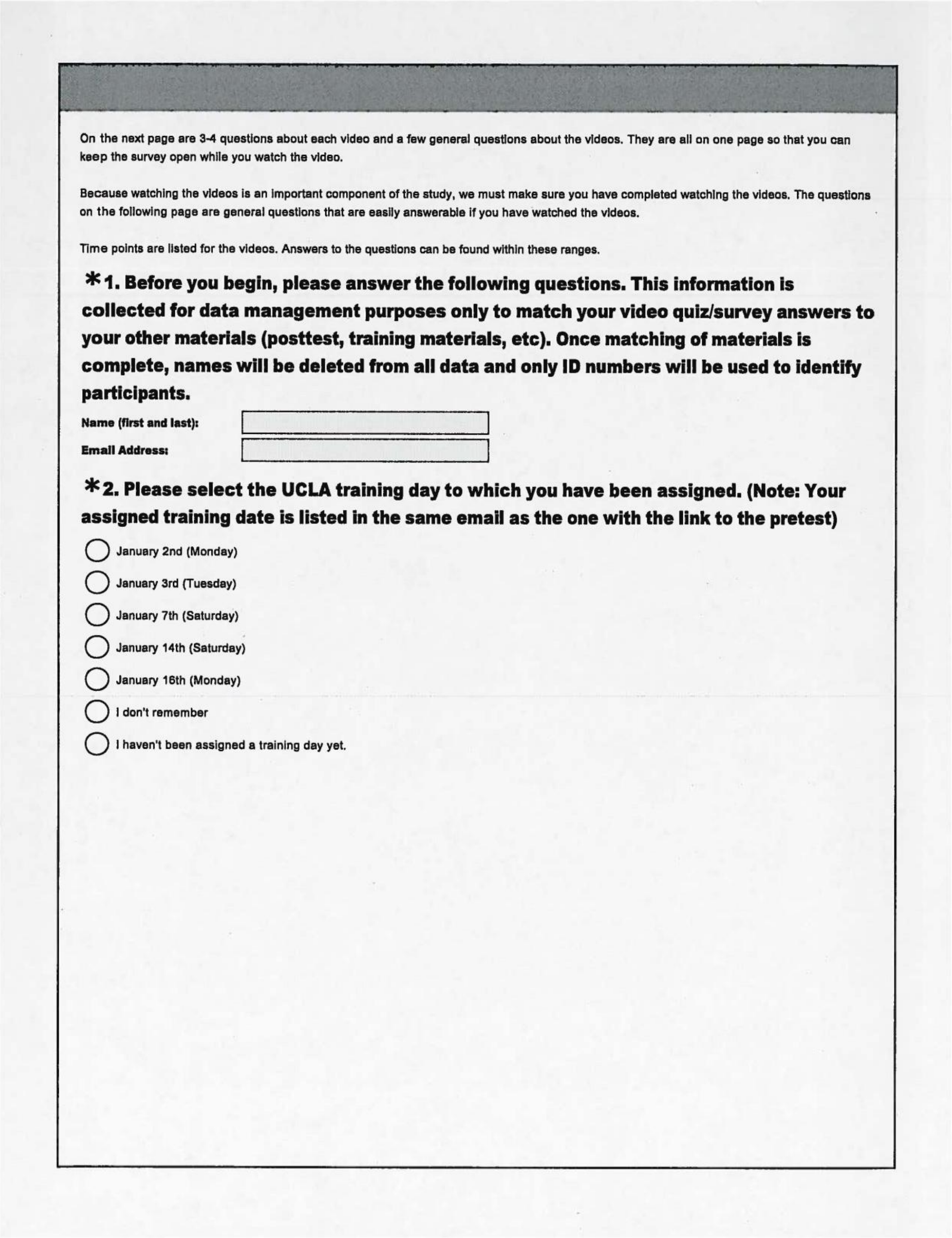
What I liked best about today’s training…

What I liked least about today's training…

Other comments (optional)…

APPENDIX G:

VIDEO QUIZ AND SURVEY



On the next page are 3-4 questions about each video and a few general questions about the videos. They are all on one page so that you can keep the survey open while you watch the video.

Because watching the videos is an important component of the study, we must make sure you have completed watching the videos. The questions on the following page are general questions that are easily answerable if you have watched the videos.

Time points are listed for the videos. Answers to the questions can be found within these ranges.

***1. Before you begin, please answer the following questions. This information is collected for data management purposes only to match your video quiz/survey answers to your other materials (posttest, training materials, etc). Once matching of materials is complete, names will be deleted from all data and only ID numbers will be used to identify participants.**

Name (first and last):

Email Address:

***2. Please select the UCLA training day to which you have been assigned. (Note: Your assigned training date is listed in the same email as the one with the link to the pretest)**

○ January 2nd (Monday)

○ January 3rd (Tuesday)

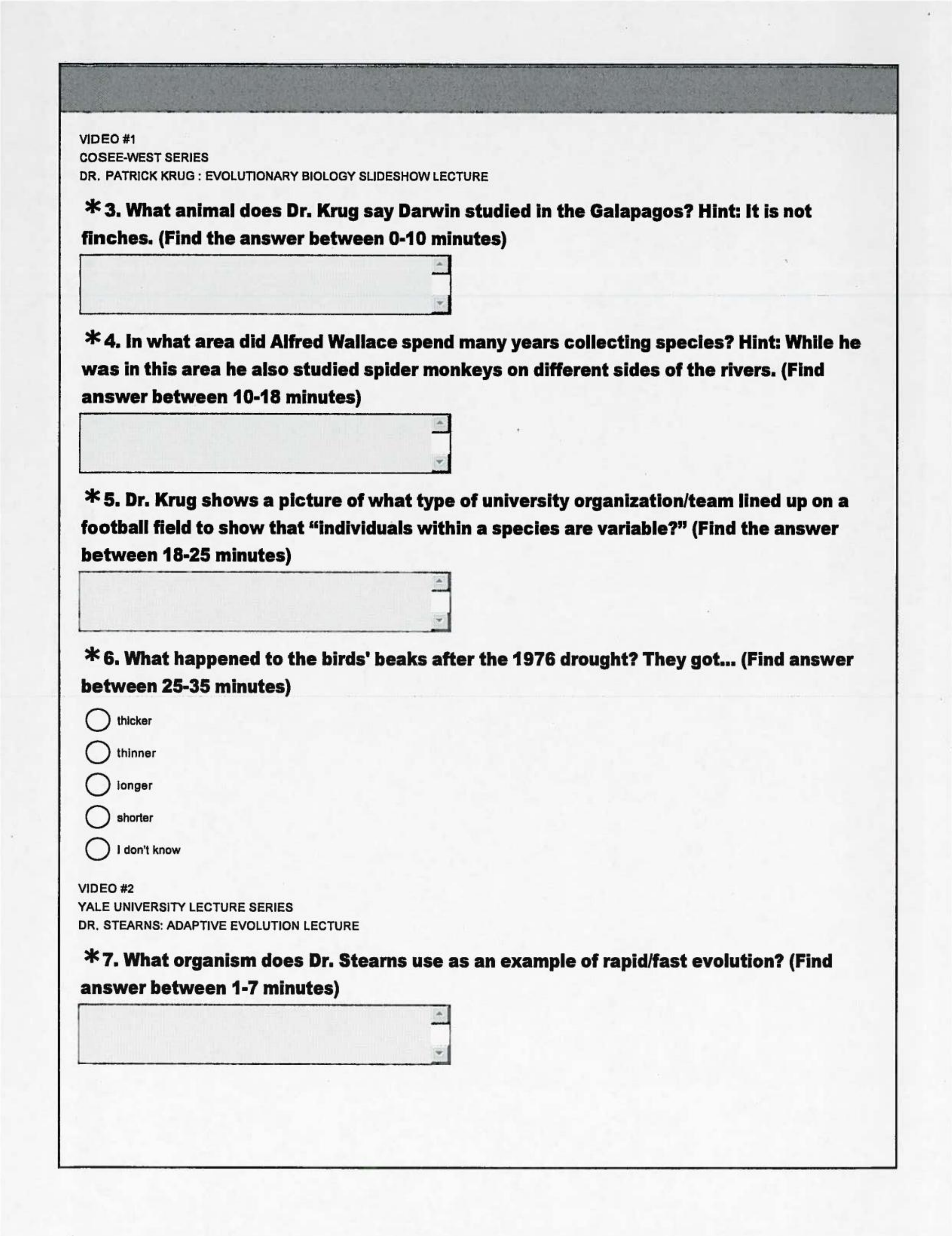
○ January 7th (Saturday)

○ January 14th (Saturday)

○ January 16th (Monday)

○ I don't remember

○ I haven't been assigned a training day yet.



VIDEO #1
COSEE-WEST SERIES
DR. PATRICK KRUG : EVOLUTIONARY BIOLOGY SLIDESHOW LECTURE

***3. What animal does Dr. Krug say Darwin studied in the Galapagos? Hint: It is not finches. (Find the answer between 0-10 minutes)**

***4. In what area did Alfred Wallace spend many years collecting species? Hint: While he was in this area he also studied spider monkeys on different sides of the rivers. (Find answer between 10-18 minutes)**

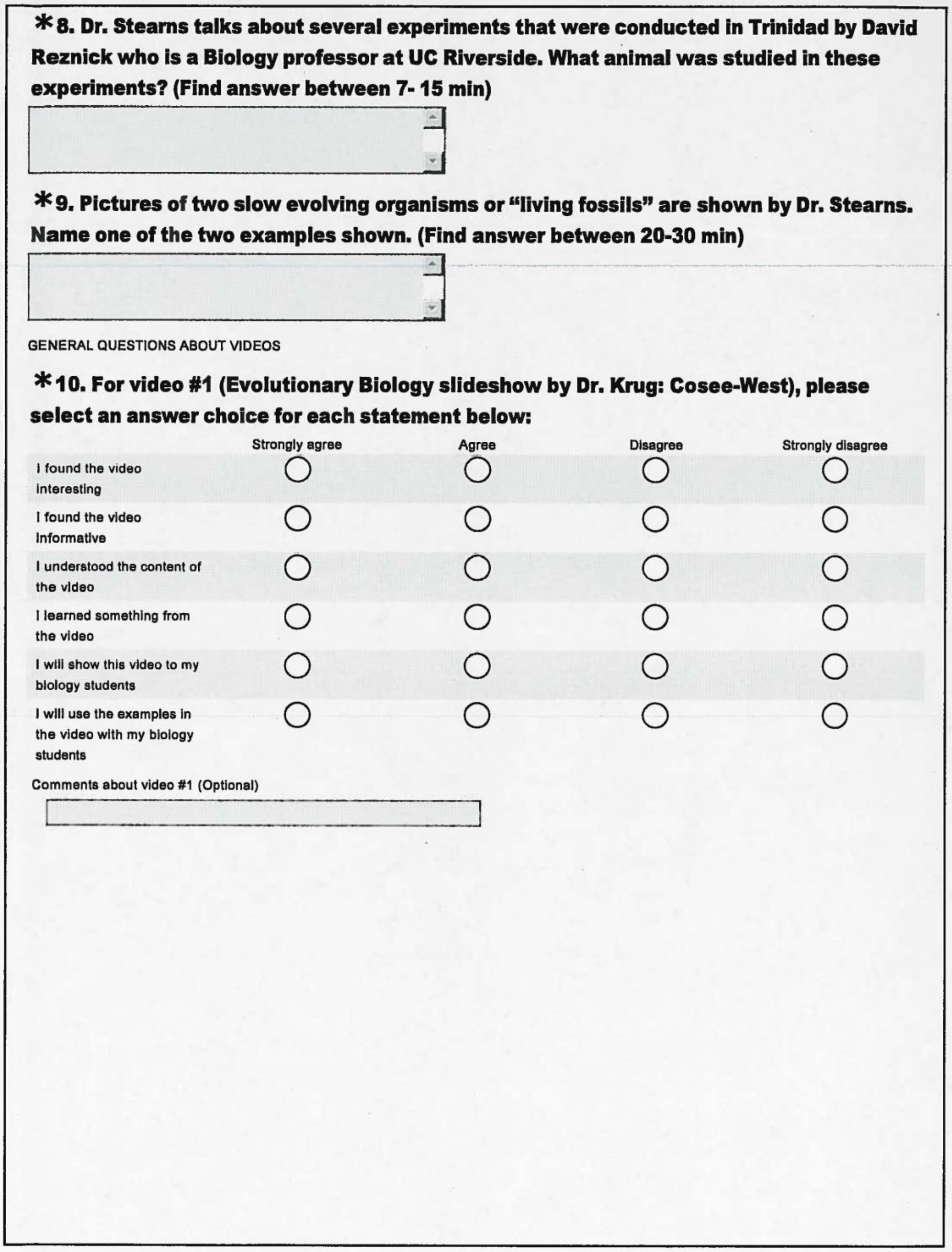
***5. Dr. Krug shows a picture of what type of university organization/team lined up on a football field to show that "individuals within a species are variable?" (Find the answer between 18-25 minutes)**

***6. What happened to the birds' beaks after the 1976 drought? They got... (Find answer between 25-35 minutes)**

○ thicker
○ thinner
○ longer
○ shorter
○ I don't know

VIDEO #2
YALE UNIVERSITY LECTURE SERIES
DR. STEARNS: ADAPTIVE EVOLUTION LECTURE

***7. What organism does Dr. Stearns use as an example of rapid/fast evolution? (Find answer between 1-7 minutes)**



***8. Dr. Stearns talks about several experiments that were conducted in Trinidad by David Reznick who is a Biology professor at UC Riverside. What animal was studied in these experiments? (Find answer between 7- 15 min)**

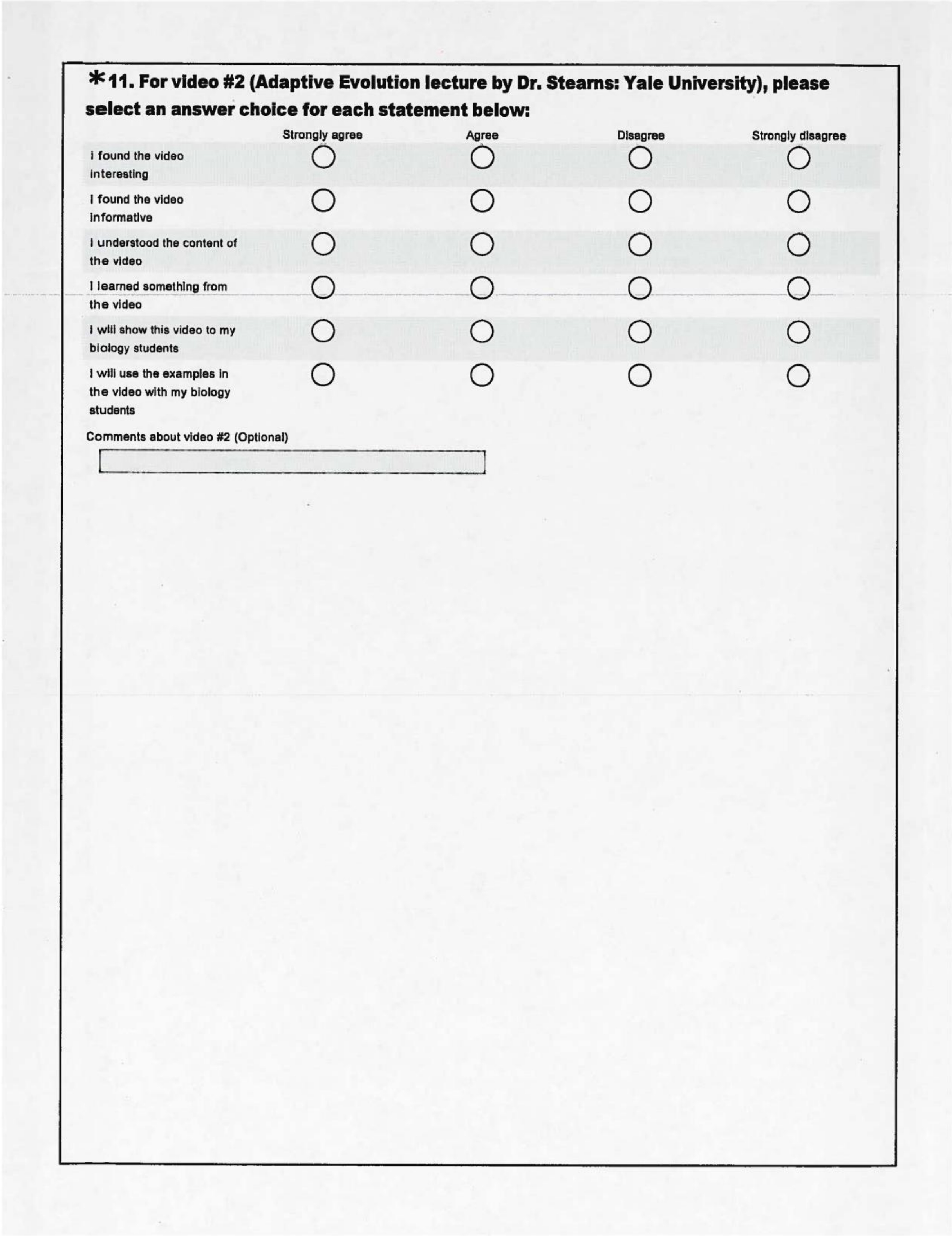
***9. Pictures of two slow evolving organisms or "living fossils" are shown by Dr. Stearns. Name one of the two examples shown. (Find answer between 20-30 min)**

GENERAL QUESTIONS ABOUT VIDEOS

***10. For video #1 (Evolutionary Biology slideshow by Dr. Krug: Cosee-West), please select an answer choice for each statement below:**

	Strongly agree	Agree	Disagree	Strongly disagree
I found the video interesting	○	○	○	○
I found the video informative	○	○	○	○
I understood the content of the video	○	○	○	○
I learned something from the video	○	○	○	○
I will show this video to my biology students	○	○	○	○
I will use the examples in the video with my biology students	○	○	○	○

Comments about video #1 (Optional)



***11. For video #2 (Adaptive Evolution lecture by Dr. Stearns: Yale University), please select an answer choice for each statement below:**

	Strongly agree	Agree	Disagree	Strongly disagree
I found the video interesting	○	○	○	○
I found the video informative	○	○	○	○
I understood the content of the video	○	○	○	○
I learned something from the video	○	○	○	○
I will show this video to my biology students	○	○	○	○
I will use the examples in the video with my biology students	○	○	○	○

Comments about video #2 (Optional)

APPENDIX H:

ITEM-LEVEL DESCRIPTIVE STATISTICS

Analyzing Student Responses. The experimental condition scored higher or equal to the control condition on the number of errors, understandings, and omissions identified in each student response. The experimental condition more accurately rated two of the four students.

Table 21

Description of the Scores and Item-Level Data for the ASR Measure

	Control (*n* = 20)		Experimental (*n* = 20)	
Task	*M*	*SD*	*M*	*SD*
Number of embedded components identified in: (maximum score 4 per student response)				
Task A	1.35	0.81	1.65	0.75
Task B	1.20	1.06	1.65	0.93
Task C	0.85	0.81	1.40	0.60
Task D	1.40	1.05	1.40	0.99
Participant ratings for: (maximum score 1 per student response)				
Task A	.63	.27	.50	.36
Task B	.50	.00	.58	.18
Task C	.15	.29	.30	.38
Task D	.48	.34	.30	.38

Examination of the identification of individual errors, understandings, and omissions shown in Table 22 indicates that in most cases, participants in the experimental condition identified each component at a higher frequency than the control condition, and that some components seem to be more difficult to identify than others regardless of condition assignment.

Table 22

Number of Components of Student Responses Identified by Participants

	Control (*n* = 20)		Experimental (*n* = 20)	
Individual coded components	*n*	%	*n*	%
Task A, presence of the following codes:				
misunderstands that change is not directional	0	0%	6	30%
has the misconception that change happens because of need	10	50%	15	75%
misunderstands survival of the fittest and links it to strong and weak	9	45%	5	25%
omitted variation in response	8	40%	7	35%
Task B, presence of the following codes:				
associates strong/weak incorrectly with sight	10	50%	14	70%
understands differential reproduction	2	10%	8	40%
understands variation (basic/some understanding)	2	10%	0	0%
understands inheritance (basic/some understanding)	2	10%	11	55%
Task C, presence of the following codes:				
has the misconception that change happens because of need/desire	9	45%	17	85%
thinks change happens to all organisms in a population at once rather than individuals	1	5%	2	10%
misunderstands role of differential reproduction in natural selection	7	35%	9	45%
Task D, presence of the following codes:				
omits differential reproduction	11	55%	11	55%
understands the role of reproduction in natural selection	7	35%	8	40%
omits inheritance	5	25%	7	35%
understands variation (basic/some understanding)	5	25%	4	20%

Natural Selection Pretest. The percentages of participants answering correctly on each individual multiple-choice item are presented in Table 23. Data indicate that on four out of the five multiple-choice items, more participants in the experimental group answered correctly than in the control group.

Table 23

Description of the Individual Multiple-Choice Items on the Natural Selection Posttest

	Control (*n* = 20)		Experimental (*n* = 20)	
Multiple choice items	n	% correct	n	% correct
Inheritance of traits	12	60%	17	85%
Acquired traits item 1	16	80%	18	90%
Acquired traits item 2	15	75%	20	100%
Artificial selection item 1	16	80%	14	70%
Artificial selection item 2	10	50%	14	70%

APPENDIX I:

ITEM-LEVEL INFERENTIAL STATISTICS

Main Research Question 1: Are there differences between conditions on participants' overall analyses of student work? The number of errors, understandings, and omissions identified in individual student responses as shown in Table 24 indicate that participants in the experimental condition identified significantly more embedded components (M = 1.40, SD = 0.60) than those in the control condition (M = 0.85, SD = 0.81) for Task C, $t(38)$ = 2.44, p = .02. Cohen's d was estimated at 0.77. No differences were found between condition for Tasks A, B, and D. A moderate effect size was estimated for Tasks A and B. An effect size of zero was estimated for Task D.

Table 24

Differences on the Number of Embedded Components Identified on the ASR for Individual Students Between Conditions

	Control (*n* = 20)		Experimental (*n* = 20)					
Number of embedded components identified for:	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Task A (maximum score 4)	1.35	0.81	1.65	0.75	38	1.22	.23	0.38
Task B (maximum score 4)	1.20	1.06	1.65	0.93	38	1.43	.16	0.45
Task C (maximum score 3)	0.85	0.81	1.40	0.60	38	2.44	.02	0.77
Task D (maximum score 4)	1.40	1.05	1.40	0.99	38	0.00	1.00	0.00

Ratings of individual student understanding presented in Table 25 indicate moderate effect sizes were estimated for Tasks A through C. An effect size of zero was estimated for Task D.

Table 25

Differences on Ratings of Student Understanding on the ASR Between Conditions

Individual ASR Rating scores for:	Control (*n* = 20)		Experimental (*n* = 20)					
	M	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Task A	.63	.28	.50	.36	38	-1.29	.23	0.40
Task B	.50	.00	.58	.18	38	1.83	.08	0.63
Task C	.15	.29	.30	.38	38	1.42	.16	0.44
Task D	.48	.34	.48	.37	38	0.00	1.00	0.00

Note. Maximum 1 point.

Main Research Question 2: Are there differences between conditions in participants' identification of critical elements of student understanding and errors in written work related to natural selection? Examination of individual student responses as presented in Table 26 indicates that the participants in the experimental condition identified more critical elements than those in the control condition for Tasks A, B, and C at a statistically significant level. Cohen's *d* for Tasks A, B, and C were estimated between .80 and 1.10.

Table 26

Differences on the Number of Critical Elements Identified Between Conditions

Number of critical elements identified for:	Control (*n* = 20)		Experimental (*n* = 20)					
	M	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Task A	0.50	0.51	1.05	0.51	38	3.40	.002	1.08
Task B	0.60	0.60	1.10	0.64	38	2.55	.015	0.81
Task C	0.50	0.61	0.95	0.51	38	2.54	.015	0.80
Task D	0.90	0.72	0.85	0.88	38	-0.20	.844	-0.06

Note. Maximum 2 points.