A COMPARATIVE STUDY OF ASSESSMENT FORMATS IN AN UNDERGRADUATE
ANATOMY AND PHYSIOLOGY COURSE

A Thesis

Presented to the faculty of the Department of Biological Sciences
California State University, Sacramento

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Biological Sciences
(Molecular and Cellular Biology)

by
Adriel Beredo Cruz

FALL
2015
A COMPARATIVE STUDY OF ASSESSMENT FORMATS IN AN UNDERGRADUATE ANATOMY AND PHYSIOLOGY COURSE

A Thesis

by

Adriel Beredo Cruz

Approved by:

__________________________________, Committee Chair
Dr. Kelly McDonald

__________________________________, Second Reader
Dr. Rosalee Sprowls

__________________________________
Date
Student: Adriel Beredo Cruz

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

________________________, Graduate Coordinator
Dr. Jamie Kneitel

Department of Biological Sciences
Abstract

of

A COMPARATIVE STUDY OF ASSESSMENT FORMATS IN AN UNDERGRADUATE ANATOMY AND PHYSIOLOGY COURSE

by

Adriel Beredo Cruz

To meet the needs of America’s growing science industry, national reports have called for reform in science education, including the way instructors assess their students. Research performed in the 1920s and 1930s encouraged the use of the then-new “choice-type” assessments (e.g. multiple-choice), showing them to be equally reliable as open-ended formats in assessing student learning. With the emergence of newer perspectives in cognition in the 1980s, researchers also began to question not just the reliability of an assessment, but an assessment’s ability to accurately evaluate multiple aspects of aptitude, including critical thinking. This prompted the development of more vigorously validated multiple-choice assessments known as concept inventories. Despite vigorous validation, concept inventories may be more limited in their ability to assess measures, such as critical thinking, than other formats, such as the oral exam.

This study compared three assessments formats – multiple-choice, written, and oral – in an undergraduate anatomy and physiology course. The study (1) revisited the
question of reliability by quantitatively comparing student scores across the three formats, (2) determined if any demographic variables were associated with potential differences in student performance, (3) identified student difficulties in the written and oral formats, and (4) examined student attitudes towards the written and oral formats through the use of a Likert-style survey. Question sets on the topics of hemodynamics and hemoglobin dissociation were validated and developed to be administered in either a multiple-choice, written, or oral format. Quantitative analyses – Kruskal-Wallis, Mann-Whitney U, correlations, and mean absolute differences – showed differential scoring between the multiple-choice and either of the open-ended formats, despite identical questions. Most quantitative analyses showed equal performance between the written and oral formats, although mean absolute differences showed slight, yet significant, differences. Two-way ANOVA analysis incorporating student demographic data detected no significant associations with most demographic variables. The only variable that showed a significant association was self-reported confidence level in the students that answered the hemoglobin questions. Qualitative analysis of the written and oral exams revealed student difficulties and misconceptions for both the hemodynamics and hemoglobin topics, with some of these difficulties revealed exclusively through the oral exam. Difficulties included the inability to define basic terms and the inability to appropriately apply mathematical models in explaining physiological concepts. Likert-style surveys revealed a diversity of attitudes among students. Most students felt uncomfortable, unable to organize their thoughts, and rushed on the oral exam.
However, other students preferred the oral exam, indicating that the act of speaking helped them better organize and elaborate on their thoughts. This study demonstrates that not all assessment formats are equal and highlights the importance of different formats in accurately assessing a diverse student population. The study also offers data on misconceptions that could lead to the development of a validated concept inventory in anatomy and physiology.

_______________________, Committee Chair
Dr. Kelly McDonald

_______________________
Date
ACKNOWLEDGEMENTS

I would like to thank Dr. Kelly McDonald for all she has done for me, both in supervising this thesis as well as all the guidance in pursuing my career as a community college professor. I would also like to thank the other members of my advisory committee: Dr. Christina Strandgaard for allowing me the use of her class and students for this project as well as her constant mentorship, and Dr. Rosalee Sprowls for all of her guidance in this project. I would also like to thank Dr. Marcello Caso and Dr. Rose Giordano for their help in validating the various assessments used in this project. Thanks to Gillian Andaya for helping with data input and analysis. Many thanks to the many students who participated in this project. Last, but definitely not least, heartfelt thanks to my family and friends for their constant moral support throughout the many years.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Components of Assessment</td>
<td>1</td>
</tr>
<tr>
<td>Cognitive Theories Applied to Assessment</td>
<td>2</td>
</tr>
<tr>
<td>Early Studies Comparing Assessment</td>
<td>3</td>
</tr>
<tr>
<td>Concept Inventories and Misconceptions</td>
<td>5</td>
</tr>
<tr>
<td>Limitation of Concept Inventories</td>
<td>7</td>
</tr>
<tr>
<td>Assessment Formats Revisited</td>
<td>9</td>
</tr>
<tr>
<td>Purpose of Thesis</td>
<td>10</td>
</tr>
<tr>
<td>PILOT STUDY METHODOLOGY</td>
<td>11</td>
</tr>
<tr>
<td>Study Population</td>
<td>11</td>
</tr>
<tr>
<td>Assessment Instrument Development</td>
<td>12</td>
</tr>
<tr>
<td>Assessment/Survey Administration</td>
<td>13</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>14</td>
</tr>
<tr>
<td>PILOT STUDY - RESULTS</td>
<td>16</td>
</tr>
<tr>
<td>One-way ANOVA Comparisons</td>
<td>16</td>
</tr>
<tr>
<td>Two-way ANOVA Comparisons</td>
<td>17</td>
</tr>
<tr>
<td>Correlation Analysis</td>
<td>19</td>
</tr>
<tr>
<td>Regression Analysis</td>
<td>20</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Two-way ANOVA with Gender</td>
<td>17</td>
</tr>
<tr>
<td>2. Two-way ANOVA between students who had taken an oral exam before vs. not</td>
<td>18</td>
</tr>
<tr>
<td>3. Two-way ANOVA with assessment preference</td>
<td>18</td>
</tr>
<tr>
<td>4. Summary of Demographic Data</td>
<td>44</td>
</tr>
<tr>
<td>5. Summary of Two-way ANOVA Analyses between format and demographic variables</td>
<td>45</td>
</tr>
<tr>
<td>6. Frequency of Misconceptions from the Diabetes Question Set</td>
<td>46</td>
</tr>
<tr>
<td>7. Summary of Demographic Data</td>
<td>67</td>
</tr>
<tr>
<td>8. Summary of Two-way ANOVA Analysis between assessment format and the demographic data</td>
<td>68</td>
</tr>
<tr>
<td>9. Frequency table of student difficulties detected by the written and oral exams</td>
<td>69</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figures</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Student scores on the three different assessment formats</td>
</tr>
<tr>
<td>2.</td>
<td>Average Scores</td>
</tr>
<tr>
<td>3.</td>
<td>Correlation Analysis</td>
</tr>
<tr>
<td>4.</td>
<td>Regression Analysis</td>
</tr>
<tr>
<td>5.</td>
<td>Average absolute differences</td>
</tr>
<tr>
<td>6.</td>
<td>Individual student scores for the hemodynamics questions</td>
</tr>
<tr>
<td>7.</td>
<td>Mean scores for the hemodynamics questions</td>
</tr>
<tr>
<td>8.</td>
<td>Correlation analysis for the hemodynamics questions</td>
</tr>
<tr>
<td>9.</td>
<td>Mean absolute differences for the hemodynamics questions</td>
</tr>
<tr>
<td>10.</td>
<td>Student responses on the Likert scale survey for the hemodynamics questions</td>
</tr>
<tr>
<td>11.</td>
<td>Student responses to the statement “Oral exams should be incorporated into the curriculum for science courses” for the hemodynamics questions</td>
</tr>
<tr>
<td>12.</td>
<td>Individual student scores for the hemoglobin questions</td>
</tr>
<tr>
<td>13.</td>
<td>Mean scores of the hemoglobin questions</td>
</tr>
<tr>
<td>14.</td>
<td>Correlation analysis for the hemoglobin questions</td>
</tr>
<tr>
<td>15.</td>
<td>Mean absolute differences for the hemoglobin questions</td>
</tr>
<tr>
<td>16.</td>
<td>Student responses on the Likert scale survey for the hemoglobin questions</td>
</tr>
<tr>
<td>17.</td>
<td>Student responses to the statement “Oral exams should be incorporated into the curriculum for science courses” for the hemoglobin questions</td>
</tr>
</tbody>
</table>
LITERATURE REVIEW

In a 2012 report to the president, the President’s Council of Advisor’s on Science and Technology (PCAST) reported a growth in North America’s science industry (Holdren and Lander, 2012). To meet this growth, the Council suggested that the nation will need to graduate approximately one million additional Science, Technology Engineering, and math (STEM) college graduates in the next decade. Along with increasing the number of graduates, the report also gave recommendations on how higher education can improve its training of STEM students. These recommendations included applying empirically validated teaching practices, which should have the added benefit of increasing diversity of the STEM student population.

When empirically testing teaching practices, researchers can look at the three aspects of education: curriculum, instruction, and assessment (Pellegrino, et al., 2001). Curriculum refers to the set of knowledge and skills that students are expected to achieve at the end of their education. Instruction refers to the methods used to impart these concepts and promote the development of competencies within students. Assessment refers to the methods used to determine how much of the curriculum a student has mastered after instruction.

Components of Assessment

There are three components to assessment that must be considered, as described in the assessment triangle (Bauerle, et al., 2001). The three corners of this triangle are observation, interpretation, and cognition. Observation refers to the “raw data” taken
from an assessment, such as a score on a multiple-choice exam or the essay written by a student. Interpretation refers to the inferences an instructor makes from these observations. Cognition is the underlying theory of how students develop and master material.

Cognition, one could argue, is the cornerstone of this triangle, and even the foundation of education overall. In terms of instruction, understanding how students learn informs the development of better instructional tools to address diverse learning preferences.

**Cognitive Theories Applied to Assessment**

For assessment, theories on cognition provides insight into what tasks are suitable for making observations. Furthermore, instructors must keep in mind that assessments can only estimate a student’s true ability. Therefore, cognitive theories help instructors understand and appreciate the limitations of whatever tasks are utilized in an assessment.

It is the constant evolution of cognitive theories that drives science education reform in assessment. For years, assessment tools have been developed using older theories of cognition, such as those grounded in the differential and behaviorist perspectives (Bauerle et al., 2001). The differential perspective assumes that individuals differ in ability (Binet and Simon, 1980). Therefore, an adequate assessment is one that can simply differentiate individuals from one another. The behaviorist perspective focuses on what is called “stimulus-response” associations (Hull, 1943; Skinner, 1938). Rewards, such as good grades, motivate students to acquire skills. Furthermore, this
perspective focuses on how basic skills need to be acquired before learning more advanced, complex skills. Thus an appropriate assessment according to this perspective is one that demonstrates those acquired skills.

The differential and behaviorist perspectives focus on the end goal of education – a student’s set of skills and abilities after instruction. Although these aspects of student learning are important, newer perspectives delve deeper into cognition. The cognitive perspective focuses on the mental processes used to develop a set of skills (Bransford et al., 1999). Therefore, an appropriate assessment aligned with this framework would be one that probes into how a student arrives at an answer. Lastly, the situational perspective focuses on how students apply their knowledge and skillset in real-world scenarios (Wertsch, 1998). An appropriate assessment for this perspective would be one that puts a student in an actual scenario, or at the very least simulates one.

**Early Studies Comparing Assessment Formats**

One other characteristic outside of the assessment triangle that must be considered is an assessment’s reliability. In other words, if two graders are asked to grade the same student’s responses, they should award the student the same grade. Studies performed as early as the 1920s addressed the issue of reliability of both free-response and choice-type questions (Hogan, 1981). In these studies, free-response questions ranged from full essays to short answer, whereas choice-type questions included multiple-choice, true/false, and even arguably fill in the blank. Three primary methods were used to compare the two assessment formats to each other were classified into three categories. Direct Correlation studies used statistical correlations to compare
student scores from two different assessment formats (Paterson, 1926; Corey, 1930; Hurd 1932). Criterion Correlation studies correlated the two formats of interest to an external criterion, such as a final grade, making the inference that the format with a stronger correlation was more reliable (Peters & Martz, 1931; Gates, 1921). Treatment Effect studies, a special kind of Criterion Correlation, incorporated some kind of treatment, such as the use of a new instructional tool. Whichever assessment format correlated better to changes in the external criterion due to the treatment were deduced to be the more reliable format (Crawford, 1925; Shulson & Crawford, 1928). The overall finding from these studies was that there was high correlation between free-response and choice-type questions. Therefore, it could be concluded that choice-type questions could replace free-response questions.

Limitations in these early studies underscore the need for more research comparing assessment formats. First, the students used in these studies were generally from psychology and physics classes. Therefore, the same inferences cannot be applied to biology students tested on concepts related to the biological sciences. Secondly, these studies presented no information on whether the tests being compared were validated for similar content, difficulty, and clarity. Furthermore, these studies were based on early cognitive theories as opposed to newer ones. One could make the argument that if a multiple-choice assessment correlates well with a free-response assessment designed to evaluate higher-order thinking, then the multiple-choice assessment would evaluate higher order thinking as well; yet more research is needed to make this assumption.
Concept Inventories and Misconceptions

To address the concerns of validity, more modern reform in assessment was pioneered with the development of the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992). This multiple-choice test was used to assess the aptitudes of physics students for basic concepts in mechanics. However, unlike common multiple-choice questions, each question was subjected to multiple levels of validation. “Experts” within the field of physics, typically other professors at other institutions, were asked to provide input on each question’s relevance to the course, the clarity of the question, and scientific accuracy. An open-ended interview was then administered to gain student input on relevance and clarity, as well as to identify common student misconceptions to use as “distractor” answer choices when forming the multiple-choice version. It was assumed that by using these student-derived distractors, open-ended formats were no longer needed, since common misconceptions could be identified by students picking that distractor choice.

Use of the Force Concept Inventory has inspired some instructors to re-evaluate their courses. Many professors, confident in their students’ proficiency, administered this test expecting amazing results. Once administered however, some professors found that their students were not as proficient as anticipated (Mazur, 2009). Thus, the Force Concept Inventory allowed physics instructors to get a better view of how prepared their students were and how effective their curriculum was at preparing students. With this information, professors were then able to adjust or even overhaul their instruction to better address any new deficiencies revealed by the Force Concept Inventory.
Following in the footsteps of the Force Concept Inventory, researchers in other science disciplines, including the biological sciences, began to develop validated assessments for their respective fields. Similar multiple-choice tests began to emerge, including the Genetics Concept Assessment (Smith, Wood, & Knight, 2008), the Biology Concept Inventory (Klymkowsky, Underwood, & Garvas, 2010), and the Introductory Molecular and Cell Biology Assessment (Shi et al., 2010). In developing these multiple-choice tests, researchers first interviewed students by simply asking them open-ended questions. Through these interviews, researchers were able to identify commonly held misconceptions on specific biological topics. For instance, interviews for the development of the Introductory Molecular and Cell Biology Assessment revealed that students held the misconception that individual chromosomes contained genetic material from both parents (Shi et al., 2010). The Biology Concept Inventory interviews revealed that students believed that biology was governed by directed, as opposed to random, processes (Garvin-Doxas & Klymkowsky, 2008).

One surprising finding during these interviews was that students receiving exemplary grades in their respective courses still held misconceptions. Even more surprising was the fact that some students with below average grades in their courses did not possess misconceptions and could answer the questions correctly (Smith et al., 2008). This phenomenon could only be observed through the use of oral interviews, but Smith and colleagues did not provide further explanation into this phenomenon.

Although no formal concept inventory has been published for the field of physiology, researchers have discovered commonly held misconceptions in various
subfields of physiology. Researchers at Liverpool John Moores University found that students in fields of exercise physiology believe that lactate is produced in muscle tissue only during fully anaerobic conditions (Morton et al., 2008). Students at University of Kentucky were found to believe that as solutes were cleared through the kidneys, the remaining solutes would not diffuse through the remainder of the plasma (Richardson & Speck, 2004). The International Intermedical School Physiology Quiz, an annual competitive quiz between medical school students across various countries, revealed that even medical students held the basic misconceptions that the partial pressure of oxygen in air is greater in the lungs than the outside atmosphere (Cheng and Durairajanayagam, 2012) A researcher at Rush Medical College identified a variety of student misconceptions, including that the nervous system constantly holds heartbeat strength, regardless of exercise (Michael, 1998), and that when carbon monoxide displaces oxygen from hemoglobin, this causes the partial pressure of oxygen in the blood to decrease as well (Michael et al., 1999).

**Limitation of Concept Inventories**

Although the development of concept inventories is gaining widespread popularity within the science education community, critics have noted some inherent flaws. The first of these flaws is the fact that these assessment tools are closed-ended multiple-choice tests. The multiple-choice format is limited in its ability to measure a student’s true understanding of the material. For example, multiple-choice questions, by nature, are unable to assess a student’s ability to synthesize new ideas from a foundational understanding of biology (Crowe et al., 2008). Students can also rely on
intuition in choosing the correct answer (Smith & Tanner, 2010). Other factors that affect a student’s understanding are also not (or are poorly) assessed with these tools. Researchers at James Madison University administered the Natural World-9 instrument, a content-non-specific tool designed to assess a student’s abilities in scientific reasoning. The researchers found that the lack of basic scientific reasoning skills may be the reason why students are not performing at expected levels (Hurney et al., 2011). Another example of how these concept inventories may not be adequate can be found in the research performed by a team at Loyola Marymount University. These researchers administered both the Force Concept Inventory as well as the Lawson test, another tool designed to assess a student’s abilities in scientific reasoning. They found that students’ scores on the Lawson test showed a strong, positive correlation with their learning gains as assessed by the Force Concept Inventory (Coletta & Philips, 2005). In other words, students with better scientific reasoning skills were found to have learned more throughout the school term, an aspect of scientific ability that cannot be assessed by a simple multiple-choice test such as a concept inventory.

Other studies have attempted to utilize different testing formats in order to better assess students in varied aspects of scientific reasoning. An example of an expanded multiple-choice test is the “two-tier” test, in which one multiple-choice question is followed by another multiple-choice question that asks the student to select from a list of statements that provide a reason for the student’s answer on the previous question (Tamir, 1971). Another form of assessment is called the Problem-sorting Task, in which students are prompted to sort multiple-choice questions by similar content. Researchers
at the University of Pittsburgh administered this in a physics class and found that novice students sorted questions by superficial features, such as problems involving blocks on inclined planes, whereas expert students sorted questions by underlying concepts, such as Newtonian mechanics (Chi et al., 1981).

**Assessment Formats Revisited**

A recent study performed by Nehm and Schonfeld (Nehm & Schonfeld, 2008) made a comparison between assessment formats in ecology and evolution. Using questions from validated tests, the authors evaluated a multiple-choice exam, a written exam, and an oral exam for their abilities to assess student’s understanding in certain areas of ecology and to uncover student misconceptions. The authors found that oral exams did not elicit any novel misconceptions when compared to the other two formats, and therefore concluded that written essay exams were equivalent to oral exams in that aspect. When comparing scores, they used what he termed the Natural Selection Performance Quotient (NSPQ). This value determines student competency by measuring both understanding of key concepts and a lack of misconceptions. They performed correlations similar to the previous Criterion Correlation studies and found that the NSPQ scores correlated with the percentage scores on both the written essay exam and oral exam. They also found that the essay exam and oral exam scores correlated with each other, supporting the claim that written essay exams are an equivalent assessment format to the oral exam. Based on these data, the authors recommend the use of written essays, since written essay exams can be administered and graded more easily and efficiently than oral exams.
Purpose of Thesis

The use of oral exams may still be crucial to adequately assessing a science student’s critical thinking skills. This format allows for the use of open-ended questions that can fully evaluate the student’s thinking and reasoning processes, as well as their content knowledge (Wright et al., 1998). Although an essay-style exam also uses open-ended questions, the oral exam allows for follow-up questions, providing the interviewer with the opportunity to conduct a more thorough assessment. Considering the complex nature of anatomy and physiology concepts, the many misconceptions prevalent in anatomy and physiology students and the lack of a validated concept inventory to assess students, the oral exam may be the best way to truly understand what anatomy and physiology students know. Furthermore, the oral format of assessment may appeal to certain students. This study explores these questions and aims to reveal the potential importance of this particular assessment format.
PILOT STUDY METHODOLOGY

Study Population and Context for Participation

A pilot study was conducted to develop the methodology for the final study. Methods were developed to compare student responses across three assessment formats – multiple-choice, written, and oral. Methods were also developed to explore demographic variables of the student population and student attitudes towards the written and oral formats. A total of 18 students participated in this pilot study during the Spring 2012 semester. These students were enrolled in BIO 26: Anatomy and Physiology II, at California State University, Sacramento. Students participated as part of a required course assignment. The multiple-choice questions that students answered were incorporated into an actual midterm and were thus graded for accuracy. However, the open-ended questions were answered in a separate meeting. Students received participation points for these questions; their grade was not affected by the quality of the answers given during the open ended assessments. Students arranged an appointment to participate in the study, and if a student was unable to make an appointment or did not feel comfortable participating in the study, an alternative assignment was provided. This was solely for the purpose of fairness for all the students in the class.

The pilot study was approved as exempt by the IRB under protocol number 11-12-099.
Assessment Instrument Development

A copy of an actual multiple-choice midterm was obtained from the lead instructor. A subset of four questions pertaining to the hemoglobin dissociation curve was used as a content basis for creating the questions for the written and oral assessments for this study. These open-ended assessment formats showed students an empty dissociation curve and asked students to describe the curve. Question included labeling the axes, describing the relationship between variables, and describing what would happen to the curve if an external factor, such as carbon dioxide, entered the system (Appendix H and I). The written assessment and the oral assessment were identical except for the fact that the answers were provided by students in different formats, i.e. written out versus spoken aloud.

Along with creating the written and oral assessments, a demographics survey was developed (Appendix F). The survey asked students for information, such as the perceived learning style, study habits, the major and career aspirations. This survey was developed for the purposes of identifying possible factors that may influence scores on either the written or oral assessments.

Finally, a reflective essay (Appendix J) that asked students to reflect on how they felt taking each assessment format was created. Students were asked which assessment format they preferred as well as to elaborate on the reasons for their preference. Students were also asked whether they felt oral assessments should be used as a common assessment format in science curricula.
Assessment/Survey Administration

After taking the multiple-choice midterm from which the questions were obtained for content basis, students made an appointment with the researcher to take the written and oral exams, the demographic survey, and the reflective survey. Students made these appointments anywhere from 1 day to 4 weeks after taking the respective multiple-choice midterm.

A coin was flipped to decide which assessment format would be administered first. This strategy was used to eliminate the possible bias that a student will always perform better on the second test administered, since both assessments evaluated the student on identical material. Each format took a maximum of 10 minutes for a total of 20 minutes for the entire session. A transcriber was used to record the oral assessment. At a later date, the primary researcher listened to each recording and typed out a transcript in order for further quantitative and qualitative analysis.

Follow-up questions were asked after the oral assessment was administered, regardless of which format was administered first. These questions were unique to each student as they attempted to further understand the student’s thought processes behind his or her personal answers to the initial questions. Because these questions were unique to each student, they were not factored into the score the student received on the oral assessment. The answers to the follow-up questions were used to obtain qualitative data on student conceptions and misconceptions. After taking both assessment formants, the student was asked to respond to both the demographic survey and the reflective survey.
Data Analysis

Scantron answer sheets from the multiple-choice midterm were obtained for each participant. The four multiple-choice questions used were then scored on a right/wrong basis for a total of four possible points. The open-ended formats were scored on a 10-point scale according to a rubric (Appendix K) Due to differing total points, percentages were calculated in order to make comparisons between the multiple-choice and open-ended formats. Due to uncontrollable variables, such as differing academic ability and time between the authentic midterm and the open-ended formats, each individual student served as an individual block. A one-way ANOVA was performed between the three assessment formats. Because individual students were treated as individual blocks, there was no replication; therefore, post-hoc tests could not be performed. To compensate, all possible paired t-tests were performed with an adjusted alpha level of 0.017 to account for stacked Type I errors.

Following the methods of previous researchers (Paterson, 1926; Corey, 1930; Hurd 1932), correlations were conducted between assessment formats. Regression analysis was also performed. Specifically, the y-intercept was tested against zero. The aforementioned statistical methods analyzed the students as an aggregate group. To analyze students against themselves, absolute differences were calculated. Absolute differences are defined as:
Abs(percentage on format A – percentage on format B)

<table>
<thead>
<tr>
<th>Ex.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage on written assessment = 45%</td>
</tr>
<tr>
<td>Percentage on oral assessment = 25%</td>
</tr>
<tr>
<td>Absolute difference = Abs(45-25) = 20</td>
</tr>
</tbody>
</table>

Once absolute differences were calculated for all students and for all three format combinations, the averages of these absolute differences were calculated. Then one-sample t-tests were performed against the null hypothesis that the average absolute difference is 0, i.e. there is no difference between test formats.

Using the information from the demographics survey, students were grouped according to gender, assessment format preference, and whether or not they had taken an oral assessment before. This allowed for the performance of two-way ANOVA’s to identify potential factors that may underlie any scoring differences in between formats.

Statistics were calculated using IBM SPSS (version 19) and Microsoft Excel 2011.

Responses to the reflective survey were first analyzed qualitatively. Common themes in the responses were identified and categories were created. Responses were the re-read and the number of responses that contained a common theme were counted in order to provide quantitative data.
PILOT STUDY – RESULTS

One-way ANOVA Comparisons

A scatterplot of student scores gives an initial indication of differences in student scoring between formats. (Figure 1). A blocked ANOVA showed that at least one format is different from the other (Figure 2). Paired t-tests show that there is a significant difference between the multiple-choice exam and both of the free-response formats. However, there is no significant difference between written-essay exams versus the oral exams (Figure 2).

Figure 1: Student scores on the three different assessment formats
Figure 2: Average scores. Average score on the multiple-choice exam was significantly greater than either of the open-ended formats. The written and oral exams were not significantly different from each other.

**Two-way ANOVA Comparisons**

Two-way ANOVAs show that none of the aforementioned factors have statistically significant score differences (Tables 1-3).

**Table 1: Two-way ANOVA with Gender**

Tests of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>.911²</td>
<td>5</td>
<td>.182</td>
<td>3.035</td>
<td>.018</td>
</tr>
<tr>
<td>Intercept</td>
<td>6.647</td>
<td>1</td>
<td>6.647</td>
<td>110.707</td>
<td>.000</td>
</tr>
<tr>
<td>Format</td>
<td>.805</td>
<td>2</td>
<td>.403</td>
<td>6.705</td>
<td>.003</td>
</tr>
<tr>
<td>Preference</td>
<td>.037</td>
<td>1</td>
<td>.037</td>
<td>.611</td>
<td>.438</td>
</tr>
<tr>
<td>Format * Preference</td>
<td>.096</td>
<td>2</td>
<td>.048</td>
<td>.798</td>
<td>.456</td>
</tr>
<tr>
<td>Error</td>
<td>2.882</td>
<td>48</td>
<td>.060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11.238</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>3.793</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .240 (Adjusted R Squared = .161)
Table 2: Two-way ANOVA between students who had taken an oral exam before vs. not

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>.814*</td>
<td>5</td>
<td>.163</td>
<td>2.622</td>
<td>.036</td>
</tr>
<tr>
<td>Intercept</td>
<td>7.164</td>
<td>1</td>
<td>7.164</td>
<td>115.417</td>
<td>.000</td>
</tr>
<tr>
<td>format</td>
<td>.754</td>
<td>2</td>
<td>.377</td>
<td>6.070</td>
<td>.004</td>
</tr>
<tr>
<td>Oral before</td>
<td>.021</td>
<td>1</td>
<td>.021</td>
<td>.331</td>
<td>.568</td>
</tr>
<tr>
<td>format * Oral before</td>
<td>.009</td>
<td>2</td>
<td>.005</td>
<td>.073</td>
<td>.930</td>
</tr>
<tr>
<td>Error</td>
<td>2.979</td>
<td>48</td>
<td>.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11.238</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>3.793</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .215 (Adjusted R Squared = .133)

Table 3: Two-way ANOVA with assessment preference

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>.845*</td>
<td>5</td>
<td>.169</td>
<td>2.757</td>
<td>.029</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.958</td>
<td>1</td>
<td>4.958</td>
<td>80.760</td>
<td>.000</td>
</tr>
<tr>
<td>format</td>
<td>.424</td>
<td>2</td>
<td>.212</td>
<td>3.455</td>
<td>.040</td>
</tr>
<tr>
<td>Gender</td>
<td>.042</td>
<td>1</td>
<td>.042</td>
<td>.684</td>
<td>.412</td>
</tr>
<tr>
<td>format * Gender</td>
<td>.015</td>
<td>2</td>
<td>.007</td>
<td>.120</td>
<td>.887</td>
</tr>
<tr>
<td>Error</td>
<td>2.947</td>
<td>48</td>
<td>.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11.238</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>3.793</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .223 (Adjusted R Squared = .142)
Correlation Analysis

There is a significant correlation between the written essay exam and the oral exam (Figure 3). This suggests that the two formats are equally reliable and that one can be used in place of the other. However, we were unable to reject the null hypothesis for correlations between the multiple-choice exam and either free-response exam (Figure 3). Therefore, the reliability between these formats is unsatisfactory, indicating that the multiple-choice exam may not equally assess student aptitude as the open ended formats.

Figure 3: Correlation analyses between the written and multiple-choice format, between the oral and multiple-choice format, and between the oral and written format.
Regression Analysis

Regression analyses show that only the constants from the regressions between Written/Multiple-choice and Oral/Multiple-choice are statistically significant (Figure 4). However, the constant from the Written/Oral regression is marginally significant (Figure 4). Statistically, this provides a model to which one can predict a student’s score on one format by knowing the score on another format. However, because this model includes statistically significant constants, a base percentage must be added to the score. This base percentage indicates that percentage discrepancy between one format to another.
Figure 4: Regression between the written and oral exams ($R^2=0.55$, $A=0.11$, $P=0.056$), between the multiple-choice and oral exams ($R^2=0.12$, $A=0.40$, $P=0.004$), and between the multiple-choice and written exams ($R^2=0.08$, $A=0.408$, $P=0.008$).
Absolute Differences

Because multiple t-tests were performed, the alpha level was distributed between the three tests, thus lowering it to alpha=0.017. At this new alpha, we were able to reject the null hypothesis of an expected average absolute difference of zero, i.e. no difference between test formats within a student. Only for the comparisons between the multiple-choice exam with either of the open-ended formats. (Figure 5). A sign test was performed for the comparison between the written exam and the oral exam due to non-normality. The sign test was not able to reject the null hypothesis at $\alpha=0.017$, although it does at $\alpha=0.05$.

![Absolute Differences](image)

Figure 5: Average absolute differences. Only the Written-MC and Oral-MC absolute differences were statistically significantly different from zero. The written-oral absolute difference was not statistically significantly different.
Student Attitudes – Data from the Reflective Essays

Of the 18 students included in this pilot study, 13 preferred taking the written assessment. Of these 13 students, when asked if oral exams should be used as a normal form assessment 8 answered “no” while 5 answered “yes”. Students who thought that oral assessments should not be used thought so because they themselves were not comfortable with the oral format.

“No, because I think that we have all been trained to do written tests, so I think it would be a hard transition for some students.”

However, there were students that, although they themselves preferred the written essay assessment, were empathetic to the possibility that other students may benefit from the oral format. Some of these students even acknowledged the fact that the oral format may have some inherent advantages in terms of assessing student understanding.

“If more students excel from [oral exams], then I think it would be beneficial to add to the [science] courses.”

“The university should have some oral assessments because it shows the depth of understanding a person actually has of the material, as opposed to guessing multiple-choice questions.”
The reflective essays also provided some insight as to why these 13 students preferred the written assessment. Five of these students mentioned that they were more nervous or felt more pressure while taking the oral exam.

“Oral exams make me feel somewhat uncomfortable because you have to be clear and once you say something, you can’t take it back. When a person is nervous, most of [the] time wrong answer is blurted out.”

Four students mentioned that, in written exams, students have time to think about the answer as opposed to having to answer right away.

“When doing the oral assessment, I felt like I had to give responses right away, which lead to me saying whatever came to mind, whether it made sense or not.”

In terms of performance, of these 13 students that preferred the written format, 7 actually performed better on the written exam, 2 students performed better on the oral exam, and 4 students performed equally on both.

Of the 18 students that participated in the study, the remaining 5 students preferred the oral assessment. Of these 5 students, when asked if oral exams should be used as a normal form of assessment, 2 answered yes, 2 answered no, and 1 student forgot to answer the question.

“Although I felt very hesitant at first, hearing myself think through my responses seemed to help jog my memory for more information”
“I believe oral exams would be beneficial in any scenario where critical thinking may be applied.”

“However, in a course such as chemistry, balancing out equations may prove a little difficult in that manner.”

The first response indicates that students may have different testing preferences. Although some students in the group that preferred the written exam did not like the fact that they felt pressured to respond as quickly as possible, some students in the other group found that speaking out loud was beneficial. The second response further supports the use of oral assessments as a better format when the goal is to assess a student’s critical thinking skills. The third response shows that the student recognizes that depending on the context of the content being assessed, different formats may be more appropriate. In terms of performance, out of these 5 students, 3 actually performed better on the oral exam, 1 performed better on the written, and 1 performed equally well on both formats.
OBJECTIVES AND HYPOTHESES

The first objective of the final study was to quantitatively compare student scores across the multiple-choice, written and oral assessment formats. Results from the pilot study showed statistically significant differences between student scores on the three formats. However, the differences in the number and type of questions as well as the small sample size weakened the statistical comparisons. The assessments were revised to be more comparable to each other, more students were recruited in the final study, and the following hypothesis was tested.

- Hypothesis 1: Student scores on the three assessment formats would be statistically different from each other.

The second objective of this study was to determine if any demographic variables were associated with differences in student performance across the assessment formats. There are potentially a variety of reasons that a student may perform better on one format than another. For instance, students that speak English as their second language (ESL learners) may perform better on the written exam. However, students who have more experiences in speaking in other classes or jobs may perform better on the oral exam. This and other factors, including but not limited to gender and self-confidence, were tested for any kind of association with potential differential performance.

A third objective of this study was to identify student difficulties and misconceptions through qualitative analysis of the written and oral exams. There is
evidence from the literature that oral exams are better able to assess a student’s 
scientific understanding, or lack thereof. Oral exams were also a key step in identifying 
misconceptions during the construction of concept inventories in other fields of biology. 
For this objective, the following hypothesis was tested.

- Hypothesis 2: The oral exam would be better at identifying student 
  misconceptions than the written exam.

The last objective of this study was to examine student attitudes towards the written 
and oral exams. The pilot study gave initial insight into the attitudes of students towards 
the various assessment formats. This study was expanded with two additional cohorts 
and a revised assessment plan in the primary study. Identifying student attitudes will 
help guide approaches to both curriculum development and implementation.
FINA STUDY – METHODS

Study Population

A total of 67 participants were enrolled in this study from BIO 26: Anatomy and Physiology II during the Spring 2013 semester. Students were given points following the protocol from the Pilot Study.

The final study was approved as exempt under IRB protocol number 12-13-124.

Assessment Instruments and Validation

One of the main concerns raised from the pilot study design was that, although the multiple-choice questions and the open-ended questions covered similar material, they were still different questions. Differences in both the difficulty level, as defined by Bloom’s Taxonomy (Bloom, 1956), and the specific content of the questions reduced the relevance of any statistical comparison made during the pilot study. To resolve this concern, the same set of questions were implemented in all three assessment formats occurring after the pilot. For the multiple-choice assessment, students were given answer choices for each question and asked to choose the correct answer. For the two open-ended assessment formats, students were given the identical question without the answer choices and asked to produce an answer on their own.

Two sets of questions were used for the Final Study. One question set assessed student understanding of the hemoglobin dissociation curve, while the other question set assessed student understanding of hemodynamics. Each question set was incorporated into different in-class multiple-choice exams. All students took both multiple-choice assessments, yet were only asked to make an appointment to take the corresponding oral
and written exams for one of the two question sets. This appointment was set for
day to 4 weeks after the corresponding multiple-choice exam. A
total of 33 students made appointments to re-answer the hemoglobin question set, and a
total of 34 students made appointments to re-answer the hemodynamics question set.

In order to assess the robustness and clarity of the questions that students
answered as a part of this study, the question set underwent both expert validations
(Appendices D and E) and student validation (Appendices B and C). Validation
followed a simplified protocol described by Smith et al. (2008). For the expert
validation, faculty members from the Department of Biological Sciences that teach the
various Anatomy and Physiology courses were given the multiple-choice questions as
well as written and oral assessment questions. These experts were asked to validate each
question using the following criteria:

- Is this question appropriate for students taking BIO 26?
- Is the content of the question scientifically accurate?
- Is the question written clearly and precisely?

Following faculty validation, the questions underwent student validation by a
student cohort separate from the ones that answered the final questions. Students in this
cohort answered the same questions and were asked to circle any words that were
unclear. Students were asked to comment on the clarity and preciseness of each
question. Comments from the faculty and student validation were considered and the
questions were revised appropriately before being administered in the final study.
In order to quantify the attitudinal data, the reflective essay developed for the pilot study was replaced with a reflective survey that incorporated a Likert-like scale (Appendix G). Based on the pilot data, the following statements were made:

1. I was able to organize my thoughts/answers on the **written essay/oral exam**.
2. I felt comfortable taking the **written essay/oral exam**.
3. I felt pressured/nervous while taking the **written essay/oral exam**.
4. I felt rushed taking the **written essay/oral exam**.
5. The **written essay/oral exam** was able to assess critical thinking.
6. **Oral exams** should be incorporated into the curriculum for science courses.

Students were then asked to rate each statement with an integer from 1 to 5, with 1 representing that the student “highly disagreed” with the statement and 5 representing that the student “highly agreed” with the statement. In addition, students were allowed to elaborate on their responses in order to get a more in-depth understanding of student attitudes towards the different assessment formats.

**Assessment/Survey Administration**

Another flaw identified in the pilot study was the inherent presence of two independent variables during the oral exam: 1) how the question was delivered to the student, and 2) how the student delivered a response. In the pilot study, students received the question by either reading it (written) or listening to it read aloud (oral). Afterwards, students answered the question by either writing down a response or speaking the response aloud. This made it impossible to determine which factor was truly affecting student performance. To control for this in the final study, all students
taking the written essay and oral assessment received the question by reading it. However, students responded by writing it (written assessment) or speaking out loud (oral assessment).

As in the pilot study, students were asked follow-up questions based on their unique responses to the standardized portion of the oral assessment. Again, since these follow-up questions and responses were not standardized and varied depending on the student, these responses did not affect the score on the oral assessment. Responses to the follow-up questions were analyzed qualitatively (see sub-section of Qualitative Data Analysis).

After all tests were administered, the student was given the Likert-like reflective survey and the demographic survey. All other aspects, such as appointment scheduling and randomization of order of open-ended assessments, were the same as in the Pilot Study.

**Quantitative Data Analysis**

During the pilot study, only the primary researcher graded the written essays and oral assessments. With the absence of additional graders, any biases inherent to the primary researcher grader were left unchecked. Therefore, in the final study, the course coordinator for BIO 26 was introduced as a second grader in order to reduce any grading bias. Two random subsets with 10 student responses each, one from the hemoglobin question set and one from the hemodynamics question set, were graded by both graders and any discrepancies in grading were discussed and resolved. The
remainder of the responses from both question sets were graded by the primary researcher only.

During the pilot study, not all of the demographic data collected from the surveys were analyzed. However, to perform a more thorough statistical analysis, all demographic data were included in the final study. Students were first grouped based on their responses to the questions on the survey. These groups were determined based on potential factors that may result in students performing differentially on different assessment formats.

The groupings are as follows.

- **Confidence** – Students gave self-reported confidence levels in their ability to learn Biology. The following groups were created.
  - Students who self-reported neutral, insecure, or very insecure.
  - Students who self-reported somewhat confident and very confident

- **English** – Students were asked how long it had been since they had taken an English class. The following groups were created.
  - Students who had taken an English course the previous semester prior to taking BIO 26.
  - Students who had taken an English course two semesters prior to taking BIO 26.
  - Students who had taken an English course longer than two semesters prior to taking BIO 26

- **Gender** – Student genders.
- First A&P – Students were asked whether or not they had taken any Anatomy of Physiology courses prior to taking BIO 26 or its pre-requisite, BIO 25.
- Oral Before – Students were asked whether or not they had taken an oral exam prior to participating to this study.
- Study Skills – Students were asked if they felt they had good study skills. Students gave one of the following answers.
  o Yes
  o No
  o I don’t know.
- Pre – health – Students were categorized by whether or not they were pursuing any careers in the health/medical field.

After students were grouped, a series of two-way ANOVA’s were performed in order to determine if students in different groups performed differentially across the different assessment formats. All other statistical analyses from the pilot study were performed using the same approaches. These included regression analyses and absolute difference calculations.

To analyze the Likert-like data from the updated reflective survey, student responses on similar questions were paired.
Student numerical responses on the written essay statements were then arithmetically subtracted from the numerical responses from the oral assessment.

Example:

I was able to organize my thoughts/answers on the **written essay**.
(Student’s response: 4)

I was able to organize my thoughts/answers on the **oral assessment**.
(Student’s response 2)

Positive differences indicate student preferences for the written essay for the particular statement whereas negative differences indicate student preference for the oral assessment. Histograms were plotted for the frequencies of these differences. The only statement where this did not apply is the statement pertaining to whether a student believed oral assessments should be used in a science course curriculum. A separate histogram displays the responses to this statement.

**Qualitative Data Analysis**

Qualitative analysis, especially in exploratory studies such as this one, often utilize grounded theory to establish a framework of study. Grounded theory is named for the mindset that any hypothesis or theories inferred from the data should be grounded in the data itself. Unlike other research methods that create hypotheses before collecting data and then use the data to either confirm or reject those hypotheses, grounded theory is achieved by looking at raw data without any preconceptions or expectations. Inferences are then made organically as the data is analyzed.
In order to create a grounded theory, answers to the follow-up questions in the final study underwent a rigorous coding process. First, student responses underwent simultaneous use of First Cycle Coding Methods (Saldana). The primary method employed was Initial Coding, also known as Open Coding. In initial coding, the researcher assesses the students’ responses and looks for similarities between the responses in order to make preliminary categories. Some coding sub-methods also used included in vivo coding and descriptive coding. In vivo coding uses verbatim phrases from the actual transcripts to try and create categories while descriptive coding uses one- to two-word phrases to summarize a short passage within the transcript.

After the First Cycle Coding Methods were employed, a Second Cycle Method known as Focused Coding was used. In Focused Coding, the codes created during the First Cycle Methods were then grouped together into categories. Afterwards, the number of codes grouped into each category were counted to see how prominent each category was within the set of transcripts.
RESULTS – HEMODYNAMICS QUESTIONS

Individual Scores Across Three Assessment Formats

Of the 85 students enrolled in BIO 26 during the Spring 2013 semester, 34 students answered identical questions, posed in three different formats (multiple-choice, written and oral), concerning how diabetes affects hemodynamics. The data demonstrated that, despite identical content, students scored differentially when the question formats differed (Figure 6). Of the 34 students, only 2 students received the same score on all three formats, 21 students received a different score on one of the formats, and 11 students received different scores on each of the three formats.

Figure 6: Individual student scores for the hemodynamics questions. Individual student scores on the multiple-choice, written, and oral assessment formats.
Comparison of Mean Scores Across Three Assessment Formats

The mean student scores on the various assessment formats were 3.94 ± 0.40 for the multiple-choice, 2.79 ± 0.51 for the written and 2.71 ± 0.46 for the oral format (Figure 7). Kruskal-Wallis analysis indicated that there was a statistically significant difference between the mean student scores ($\chi^2 = 16.561$, $P<0.001$). Post-hoc Mann-Whitney U tests showed that this significance resulted from differences between the mean scores on the multiple-choice format and the written format ($p<0.001$) and the multiple-choice and oral formats ($p<0.001$). However, post-hoc tests detected no statistical difference ($p=0.816$) between the mean scores on the written and the oral exam formats (Figure 7).
Figure 7: Mean scores for the hemodynamics questions. Mean scores for the multiple-choice format (MC Score), written format (Written Score), and oral format (Oral Score). Asterisks show statistically significant mean scores based on Kruskal-Wallis and post-hoc Mann-Whitney U tests.
Correlation Analysis of Individual Scores

Correlation analysis was performed between the multiple-choice and written formats, the multiple-choice and oral formats, and the written and oral formats (Figure 8) to determine whether these exam formats assessed students equally. Correlation analysis showed poor and statistically insignificant correlations between scores on the multiple-choice and the written formats (R = 0.223; P = 0.202, Figure 3), as well as scores on the multiple-choice and the oral formats (R=0.110; p=0.543, Figure 4), suggesting that these pairs do not assess students equally. However, there was a moderate and statistically significant correlation between scores on the written and oral formats (R=0.688; p<0.001, Figure 5).
Figure 8: Correlation analysis for the hemodynamics questions. Correlation between multiple-choice and written scores ($R=0.223; P=0.202$), between multiple-choice and oral scores ($R=0.110; P=0.543$), and between the written and oral scores ($R=0.688; P<0.001$).

**Analysis of Mean Absolute Differences**

Although analyses showed no statistical difference between the mean scores on the written and oral exams (Figure 11), differences may have gone undetected if the student population included both individuals who scored higher on the written format.
with those who scored higher on the oral format. Furthermore, although the correlation between the written and oral formats was statistically significant, the correlation was only moderate. The scatterplot between student scores on the written and oral formats showed students distant from the line of best fit (Figure 14). To determine whether individual differences were missed in the previous analyses, absolute differences were calculated by subtracting a student’s score on one format from that student’s score on another format and taking the absolute value of that difference. After doing this for each student and each format pairing, means were calculated for the absolute differences within a format pairing. The mean absolute differences were 1.71 ± 0.40 for the multiple-choice and oral formats, 1.62 ± 0.41 for the multiple-choice and written formats, and 0.79 ± 0.27 for the written and oral formats. One-sample t-tests were then performed on the mean absolute differences against a null hypothesis of zero to determine whether differential scoring was observed for individual students. According to these tests, all differences between every possible format combination are statistically significant from zero (Figure 9). These data supported the mean comparison (Figure 7) and correlation analyses (Figure 8) between the multiple-choice format to either the written or oral format. However, these data contradicted mean comparison (Figure 7) and correlation analyses (Figure 8) between the written and oral format, suggesting that there may be some differences when comparing individual students to themselves (Figure 6).
Figure 9: Mean absolute differences for the hemodynamics questions. Mean absolute differences between the multiple-choice and oral formats (MO), the multiple-choice and written formats (MW), and the written and oral formats (WO). All means were statistically significant from the null hypothesis of zero (p<0.001).
Analysis of Demographic Information

A survey was used to collect student demographic data as well as responses to questions about confidence and study skills. All 34 students completed the survey and the data were counted and summarized (Table 4). Subsequent two-way ANOVA analyses were performed to determine whether or not various demographic factors might have influenced the tendency of students to score differentially across the different exam formats. As seen in Table 5, there appeared to be no association between the demographic variables considered in this study and student performance on a particular question format. It should be noted that only the categories of “Taken an Oral Exam Before,” “Self-reported Study Skills,” and “Pre-health” met the assumption of equal variances as tested by Leven’s Test of Homogeneity.
Table 4: Summary of Demographic Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5</td>
<td>14.7</td>
</tr>
<tr>
<td>Female</td>
<td>29</td>
<td>85.3</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>Last English Class Taken</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1 year ago</td>
<td>13</td>
<td>38.2</td>
</tr>
<tr>
<td>1 year ago</td>
<td>9</td>
<td>26.5</td>
</tr>
<tr>
<td>≤2 years ago</td>
<td>12</td>
<td>35.3</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>English as Primary Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20</td>
<td>58.8</td>
</tr>
<tr>
<td>No/Prefer not to Answer</td>
<td>14</td>
<td>41.2</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>Taken an oral exam before BIO 26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>13</td>
<td>38.2</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>61.8</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>&quot;How Confident are you in your ability to learn Biology?&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Confident/Confident</td>
<td>25</td>
<td>73.5</td>
</tr>
<tr>
<td>Neutral/Insecure/Very Insecure</td>
<td>9</td>
<td>26.5</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>&quot;Do you feel that you have good study skills?&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>18</td>
<td>52.9</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>26.5</td>
</tr>
<tr>
<td>I don't know</td>
<td>7</td>
<td>20.6</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>First Anatomy and Physiology Course</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>26</td>
<td>76.5</td>
</tr>
<tr>
<td>No</td>
<td>8</td>
<td>23.5</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>Pre-health Student</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25</td>
<td>73.5</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>26.5</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Table 5: Summary of two-way ANOVA Analyses between assessment format and demographic variables

<table>
<thead>
<tr>
<th></th>
<th>d.f</th>
<th>F</th>
<th>P</th>
<th>Levene's Test of Homogeneity (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>2</td>
<td>0.19</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td>Last English course taken</td>
<td>4</td>
<td>0.71</td>
<td>0.58</td>
<td>0.012</td>
</tr>
<tr>
<td>English as primary language</td>
<td>2</td>
<td>0.02</td>
<td>0.98</td>
<td>0.022</td>
</tr>
<tr>
<td>Taken an oral exam before</td>
<td>2</td>
<td>0.91</td>
<td>0.40</td>
<td>0.119*</td>
</tr>
<tr>
<td>Self-reported confidence level</td>
<td>2</td>
<td>1.25</td>
<td>0.29</td>
<td>0.031</td>
</tr>
<tr>
<td>Self-reported study skills</td>
<td>4</td>
<td>1.09</td>
<td>0.36</td>
<td>0.638*</td>
</tr>
<tr>
<td>First A&amp;P Course</td>
<td>2</td>
<td>0.24</td>
<td>0.78</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pre-health</td>
<td>2</td>
<td>0.09</td>
<td>0.91</td>
<td>0.251*</td>
</tr>
</tbody>
</table>

d.f. – degrees of freedom  
F – F-value from Two-way ANOVA analysis  
P – P-value of the corresponding F-value

**Student Difficulties Detected by the Written Essay and Oral Exams**

The written and oral exams, both scripted and non-scripted, were qualitatively analyzed using an open coding strategy. This method allowed for the identification of misconceptions and other student difficulties from assessment formats in which students were not prompted with answer choices. Analysis of student responses revealed frequent problems with three course-related concepts (Physiological Concept). From these concepts, misconceptions were identified and categorized and their frequencies determined for both the written and oral assessments (Table 6). Both the written and oral exams detected misconceptions related to all three concepts. In addition, an “Other”
category was formed to group responses that demonstrated conceptual difficulties that were unrelated to the three Physiological Concepts. Both the written and oral exams equally detected student difficulties for “Resistance’s Effect on Blood Pressure,” “Blood Vessel Diameter,” and “Miscellaneous Definition Misconceptions.” However, qualitative analysis of the oral exams revealed greater insight as to why students may hold these misconceptions. Furthermore, the oral exam was better at detecting misconceptions for the “Definition of Resistance.”

Table 6: Frequency of Misconceptions from the Diabetes Question Set

<table>
<thead>
<tr>
<th>Physiological Concept</th>
<th>Misconception</th>
<th>Exam Format</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Written</td>
<td>Oral</td>
</tr>
<tr>
<td><strong>Definition of Resistance</strong></td>
<td>Pressure</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Force</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Blockage</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>“Insulin Resistance”</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td><strong>The Effect of Resistance on Blood Pressure</strong></td>
<td>Inverse Relationship</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Pressure Causes Resistance</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>CO = SV x HR</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td><strong>Blood Vessel Diameter</strong></td>
<td>Vasoconstriction</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Vasodilation</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sugar Causes Clot</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td><strong>Other Misconceptions</strong></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Definition of Resistance

Students were asked how increased levels of blood sugar would affect resistance. Resistance is defined as anything that slows down blood flow in the body. The unscripted oral exam revealed that students held misconceptions in even the basic definition of this term. Two students defined resistance using the word “force.” The responses suggested that students thought that resistance was the actual force that affected blood flow.

*Examiner: “Can you define resistance?”*

*Student 18: “The amount of force needed to push the blood through the arteries.”*

Similarly, two students defined resistance using the word “pressure.”

*Student 12: “Resistance is... I guess the amount of pressure exerted on the vessels by fluid.”*

Two students defined resistance as a blockage in the blood vessel. Although a blockage would increase resistance, it is not the only factor that affects resistance.

*Student 1: “Resistance... [is] what happens when the blood vessels, there’s, there’s something about causes it to … To either be blocked.*

The oral exam also detected three miscellaneous misconceptions regarding the definition of resistance.
Student 11: “Well, I was, I guess I was thinking of like, if there is higher blood sugar, blood is not going to be flowing through as quickly. It’s not going to have as much resistance.”

Examiner: “What is resistance?”

Student 31: “The way the body reacts to abnormal levels.”

Examiner: “Abnormal levels of?

Student 31: “of (long pause) homeostasis.”

Student 32: “Cardiac resistance is (long pause) is when your heart is working harder to pump out blood.”

An interesting finding by both the oral and written exams was that some students thought that resistance referred to insulin resistance. This particular confusion may be unique to this question set, since it asked questions regarding resistance in the context of diabetes.

Student 28: “…High levels of blood sugar will cause the factor to decrease. Therefore, overall resistance… I was assuming that the overall resistance to insulin is increasing

The Effect of Resistance on Blood Pressure

Later in the question set, students were asked how the resistance would affect mean arterial blood pressure. Blood pressure is the physical pressure exerted by blood,
similar to the pressure exerted by any fluid. In the context of the question set, blood pressure is directly related to resistance in the mathematical equation $\text{CO} = \text{MAP} \times \text{Resistance}$, where CO refers to cardiac output, the amount of blood ejected by the heart in one minute, and MAP refers to the mean arterial pressure, or blood pressure. Simply put, as resistance increases, blood pressure must increase in order to maintain the same cardiac output, and it is understood that changes in resistance are the cause and subsequent changes in blood pressure are the effect.

Responses from both the written and oral exams indicated that some students held the misconception that an increase in resistance somehow decreased mean arterial blood pressure. However, the unscripted oral exams provided insight into the reasons for the faulty logic.

*Examiner:* “How did you get from increased resistance to lower blood pressure?”

*Student 1:* … because if resistance increased, the blood pressure would decrease because the... blood flow wouldn’t flow as quickly as normal.”

The unscripted oral exam detected two students that thought that blood pressure was the cause and resistance was the effect.

*Student 4:* “So, blood pressure, high blood pressure would be the cause and resistance would be the effect…. If you have high blood pressure,
then you have vasoconstriction, which would then increase resistance.”

One underlying student difficulty is the use of the alternate equation that describes cardiac output, \( CO = SV \times HR \), where the two factors that affect cardiac output are stroke volume (SV), the amount of blood ejected in one heartbeat, and heart rate (HR); how fast the heart beats. Although this is a correct description of cardiac output, there is no direct relationship between resistance to either stroke volume or heart rate. Despite this difficulty, students were still able to come to a conclusion about how resistance would affect blood pressure, including that blood pressure would increase. Only the written and oral exams were able to detect when students used the incorrect equation, as opposed to the correct equation relating cardiac output to resistance.

Student 21: “It’s that equation. It was blood pressure equals stroke volume times cardiac output. And since you mentioned before that, since stroke volume increased due to the increase of resistance, the cardiac output should decrease in order to counteract it.”

Miscellaneous misconceptions in this category were detected by both the written and oral exams, although the oral exams were better at detecting additional student misunderstandings. Misconceptions ranged from incorrect definitions of blood pressure to nonsensical logic on how sugar would affect blood physiology.

Student 3: “Blood is more viscous, there’s more resistance.”

Examiner: “…how does that necessarily affect blood pressure?”
Student 3: “...the more it is hitting against the wall of the blood vessel.”

Student 10: “…Because the heart rate is basically affected a lot by the resistance. So, if the resistance is higher, the harder it is going to pump harder. So the heart rate is going to increase.”

Student 16: “Because if it is more dense, then you have more pressure in it. It’s exerting more pressure on the walls.”

Examiner: “So you said that blood would travel faster because of the high levels of blood sugar. How did you come to that conclusion?”
Student 33: “I think that the person has high blood pressure, then I think the blood will travel faster.”

Examiner: “Can you explain how exactly sugar levels affect volume?”
Student 2: “I just said, you know more blood platelets. So it would be more erythrocytes in the blood... So your blood sugar increases, I just thought that maybe you have more blood cells in your bloodstream... more blood cells are taking in the sugar...Because there is more volume going int, so more pressure would be going in. Resistance would decrease.”
**Blood Vessel Diameter**

The high levels of blood sugar presented in the question set have no direct effect on blood vessel diameter, either increasing it (vasodilation) or decreasing it (vasoconstriction). Both the written and oral exams revealed student misunderstandings related to this concept. However, students responded to the written exam with a simple answer of either “vasodilation” or “vasoconstriction.” Responses to the oral exam were more elaborate. Two students thought that changes in blood pressure cause changes in blood vessel diameter.

*Student 4: “So, blood pressure, high blood pressure would be the cause and resistance would be the effect…. If you have high blood pressure, then you have vasoconstriction, which would then increase resistance”*

One student said that the high viscosity is what causes vasodilation.

*Student 18: “I assumed if there is more resistance, because the blood viscosity, because the blood is thicker, I assume that the artery had to open more. And that decreases the pressure.”*

Both the written and oral exams detected the misconception that sugar would cause a clot and then further defined this as vasoconstriction.

*Student 22: “So, when we refer to question one, it said high levels of blood sugar will affect which factor of resistance, it made me, the sugar... the closest thing I could come to was plaque... adhering to the
One student thought that high levels of blood would significantly increase blood volume, defined as the total amount of blood within the body, and then affect blood vessel diameter.

*Examiner:* “So how does blood sugar levels, or an increase in blood sugar levels...”

*Student 26:* “It increases the blood viscosity.”

*Examiner:* “It increases the blood viscosity.” So, what is viscosity?”

*Student 26:* “It is the thickness of blood.”

*Examiner:* “So why would that lead, eventually, to vasodilation?”

*Student 26:* “By increasing the blood volume.”

Furthermore, the unscripted oral exam detected one misconception in which blood vessel diameter would be affected, but admitted to not knowing why.

**Other Misconceptions**

The unscripted oral exam was also able to detect three misconceptions related to basic definitions that were unrelated to the three primary categories.

*Student 5:* “So the basically the sympathetic will cause vasoconstriction, so basically it will decrease the blood airways... Blood volume will probably decrease, because, basically, when you are vasoconstricting it, you’re basically closing off any of the blood
vessels.... So basically, when you have too much sugar in your blood, the sympathetic nervous system will come into play, so that’s where I got the vasoconstriction.”

Student 11: “Viscosity, as far as blood vessels go, is like, the speed...”

Examiner: “...What is contractility?”

Student 14: “The ability of blood vessels to contract.”

Attitudinal Data

Responses from a Likert survey regarding student attitudes towards the written exam were arithmetically subtracted from responses to identical statements about the oral exams. Positive differences indicate that students felt more strongly about the written exam for that particular statement, whereas negative differences indicate that students felt more strongly about the oral exam (Figure 10). Students were also given the opportunity to elaborate on their responses, which provided further insight into their attitudes about the different testing formats.
Figure 10: Student responses on the Likert scale survey for the hemodynamics questions. Students were asked about (a) how well they were able to organize their thoughts, (b) how comfortable they felt, (c) how pressured they felt, (d) how rushed they felt, and (e) how well the format could assess critical thinking. Responses were
arithmetically subtracted (written – oral). Positive differences (to the right of the dotted line) represent students who felt more strongly about the written exam. Negative differences (to the left of the dotted line) represent students who felt more strongly about the oral exams. Differences of zero (on the dotted line) represent students who did not feel differently about the two formats.

When asked about their ability to organize their thoughts during the written and oral formats, only five students felt that they performed this task more successfully on the oral exam, and four felt equally organized on both the written and oral exams. Not surprisingly, the majority (25) of students felt that they were better able to organize their thoughts on the written exam (Figure 10a). One student said that it was the actual process of writing that helped him/her think of the answer. On the other hand, another student felt that the process of speaking aloud helped him/her to provide a more elaborate answer.

In a related question about feelings of comfort with the two assessment formats, only four students felt more comfortable during the oral exam, whereas eighteen (18) students felt more comfortable during the written exam. Twelve students felt equally comfortable on both the written and oral exams (Figure 10b). One student felt uncomfortable with being questioned on the spot during the unscripted part of the oral exam. Another student felt awkward on the oral exam because he/she knew that the answer he/she was providing was incorrect.

In another question about the pressure that students felt on the two test formats, the majority (23) indicated feeling more pressure on the oral exam. Ten students felt an
equal amount of pressure on both the written and oral exam. Only one student felt more pressure on the written exam (Figure 10c). One student explained that he/she felt nervous because this was the first oral exam he/she had ever taken.

A related question dealt with time management on the written and oral exams. Seventeen (17) students indicated that they felt more rushed on the oral exam, while only one felt more rushed on the written. Fifteen (15) students felt equally rushed on both the written and oral exams (Figure 10d). One student felt that, even though he/she knew that there was the same amount of time given for both the written and oral formats, he/she felt as if there was less time on the oral. Another student felt rushed because he/she felt that silence during the conversation made the oral exam awkward.

When asked about the ability of the written and oral formats to assess critical thinking, responses were divided fairly equally. Twelve students believed that the written exam was better for assessing critical thinking, while ten students felt that this skill was required more for the oral exam. Twelve students felt that the written and oral exams assessed critical thinking equally (Figure 10e). One student felt that no elaboration was optional on the written exam. However, the same student felt that unscripted oral exam, he/she felt the exam truly tested her critical thinking.

Furthermore, student elaborations on the Likert-like survey also showed that these attitudinal categories were interrelated. For example, some students felt they were not able to organize their thoughts on the oral exam because of anxiety. Other students felt rushed because of the pressure they felt. Many students attributed their discomfort during the oral exam to feelings of being rushed, being pressured, or their inability to
organize their thoughts. For some students, the feelings of pressure and being rushed along with the inability to organize their thoughts were what led them to feel that the oral exam could not assess critical thinking.

*Student 12*: I felt I had more time to think [on the written essay]. [Oral exams] somewhat [assess critical thinking], but it is harder to critical think under pressure.

When asked if oral exams should be incorporated into course curricula more frequently, ten (10) students agreed or strongly agreed, nine (9) students felt neutral and fourteen (14) students disagreed or strongly disagreed. One did not answer and was designated a score of zero (Figure 11). Students who disagreed tended to do so because they felt uncomfortable on the oral exam. Students who felt neutral or agreed looked past their own experiences and reflected on the actual worth of the oral exam. One student who felt neutral talked about how the oral exam would help pre-nursing students develop the ability to think on their feet. Another student who felt neutral admitted to hating the oral exam, but also being able to learn from it. One student who agreed that oral exams should be used also commented on the oral exam’s ability to help students remember concepts.
Figure 11: Student responses to the statement “Oral exams should be incorporated into the curriculum for science courses” for the hemodynamics questions
RESULTS – HEMOGLOBIN QUESTIONS

Individual Scores Across Three Assessment Formats

Of the 85 students enrolled in BIO 26 during the Spring 2013 semester, 33 students answered identical questions, posed in three different formats (multiple-choice, written and oral), concerning the hemoglobin dissociation curve. One question on the multiple-choice format did not have a correct answer. Therefore, that question was removed from all three formats for the quantitative analysis. The data demonstrated that, despite identical content, students scored differentially when the question formats differed (Figure 12). Of the 33 students, only 4 students received the exact score on all three formats, 23 students received a different score on one of the formats, and 6 students received different scores on each of the three formats.

Figure 12: Individual student scores for the hemoglobin questions. Individual student scores on the multiple-choice, written, and oral assessment formats.
Comparison of Mean Scores Across Three Assessment Formats

The mean student scores on the various assessment formats were 3.09 ± 0.30 for the multiple-choice, 1.36 ± 0.39 for the written and 1.48 ± 0.42 for the oral format (Figure 13). Kruskal-Wallis analysis indicate that there is a statistical difference between the mean student scores ($\chi^2 = 36.318$, P<0.001). Post-hoc Mann-Whitney U tests showed that this significance resulted from differences between the mean scores on the multiple-choice format and the written format (p<0.001) and the multiple-choice and oral formats (p<0.001). However, post-hoc tests detected no statistical difference (p=0.638) between the mean scores on the written and the oral exam format (Figure 13).

Figure 13 - Mean scores for the multiple-choice format (MC Score), written format (Written Score), and oral format (Oral Score). Asterisks show statistically significant mean scores based on Kruskal-Wallis and post-hoc Mann-Whitney U tests.
Correlation Analysis of Individual Scores

Correlation analysis was performed between the multiple-choice and written format, the multiple-choice and oral format, and the written format and oral format (Figure 14) to determine whether these exam formats assessed students equally. Correlation analysis showed poor and statistically insignificant correlations between scores on the multiple-choice and the written formats ($R = 0.230; P = 0.197$, Figure 14), as well as scores on the multiple-choice and the oral formats ($R=0.145; p=0.426$, Figure 14), suggesting that these pairs do not assess students equally. However, there was a moderate and statistically significant correlation between scores on the written and oral formats ($R=0.840; p<0.001$, Figure 14).
Figure 14: Correlation analysis for the hemoglobin questions. Correlation between multiple-choice and written scores (R=0.230; P=0.197), between multiple-choice and oral scores (R=0.145; P=0.426), and between the written and oral scores (R=0.840; P<0.001).
Analysis of Mean Absolute Differences

Although analyses showed no statistical difference between the mean scores on the written and oral exams (Figure 13), differences may have gone undetected if the student population included both individuals who scored higher on the written format with those who scored higher on the oral format. Furthermore, although the correlation between the written and oral formats was statistically significant, the correlation was only moderate. The scatterplot between student scores on the written and oral formats showed students distant from the line of best fit (Figure 14). To determine whether individual differences were missed in the previous analyses, absolute differences were calculated by subtracting a student’s score on one format from that student’s score on another format and taking the absolute value of that difference. After doing this for each student and each format pairing, means were calculated for the absolute differences within a format pairing. The mean absolute differences were 1.73 ± 0.42 for the multiple-choice and oral formats, 1.79 ± 0.41 for the multiple-choice and written formats, and 0.30 ± 0.21 for the written and oral formats. One-sample t-tests were performed on the mean absolute differences against a null hypothesis of zero to determine whether differential scoring was observed for individual students. According to these tests, all differences between every possible format combination were statistically significant from zero (Figure 15). These data support the findings from the mean comparison (Figure 13) and correlation analyses (Figure 14) between the multiple-choice format and either the written or oral format. However, these data contradicted mean comparison (Figure 13) and correlation analyses (Figure 14) between
the written and oral format, suggesting that there may be some differences when comparing individual students to themselves (Figure 12).

Figure 15: Mean absolute differences for the hemoglobin questions. Mean Absolute Differences between the multiple-choice and oral formats (MO), the multiple-choice and written formats (MW), and the written and oral formats (WO). All means were statistically significant from the null hypothesis of zero (p<0.001)
Analysis of Demographic Information

A survey was used to collect student demographic data as well as responses to questions about confidence and study skills. All 33 students completed the survey and the data were counted and summarized (Table 7). Subsequent two-way ANOVA analyses were performed to determine whether or not various demographic factors may have influenced the tendency of students to score differentially across the different exam formats. As seen in Table 8, the only variable that showed a statistically significant association with assessment format was self-reported confidence. There appears to be no association between any other demographic variables considered in this study and student performance on a particular assessment format. It should be noted that the categories “Gender” and “English as a Primary Language” did not meet the assumption of equal variances as tested by Levene’s Test of Homogeneity.
Table 7: Summary of Demographic Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>10</td>
<td>30.3</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>63.6</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Last English Class Taken</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1 year ago</td>
<td>10</td>
<td>30.3</td>
</tr>
<tr>
<td>1 year ago</td>
<td>9</td>
<td>27.3</td>
</tr>
<tr>
<td>≤2 years ago</td>
<td>14</td>
<td>42.4</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>English as Primary Language</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>21</td>
<td>63.6</td>
</tr>
<tr>
<td>No/Prefer not to Answer</td>
<td>14</td>
<td>42.4</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Taken an oral exam before BIO 26</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>13</td>
<td>39.4</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>63.6</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>103.0</td>
</tr>
<tr>
<td><strong>&quot;How Confident are you in your ability to learn Biology?&quot;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Confident/Confident</td>
<td>25</td>
<td>75.8</td>
</tr>
<tr>
<td>Neutral/Insecure/Very Insecure</td>
<td>8</td>
<td>24.2</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>&quot;Do you feel that you have good study skills?&quot;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>9</td>
<td>27.3</td>
</tr>
<tr>
<td>No</td>
<td>15</td>
<td>45.5</td>
</tr>
<tr>
<td>I don't know/Prefer Not To Answer/Never Studied for Previous Classes</td>
<td>9</td>
<td>27.3</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>First Anatomy and Physiology Course</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>19</td>
<td>57.6</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>42.4</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Pre-health Student</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>27</td>
<td>81.8</td>
</tr>
<tr>
<td>No</td>
<td>16</td>
<td>48.5</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Table 8: Summary of two-way ANOVA Analyses between assessment format and the demographic data

<table>
<thead>
<tr>
<th></th>
<th>d f</th>
<th>F</th>
<th>P</th>
<th>Levene's Test of Homogeneity (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>2</td>
<td>2.724</td>
<td>0.071</td>
<td>0.008</td>
</tr>
<tr>
<td>Last English course taken</td>
<td>4</td>
<td>0.29</td>
<td>0.884</td>
<td><strong>0.809</strong>*</td>
</tr>
<tr>
<td>English as primary language</td>
<td>2</td>
<td>0.049</td>
<td>0.952</td>
<td>0.026</td>
</tr>
<tr>
<td>Taken an oral exam before</td>
<td>2</td>
<td>0.13</td>
<td>0.878</td>
<td><strong>0.168</strong>*</td>
</tr>
<tr>
<td>Self-reported confidence level</td>
<td>2</td>
<td>3.317</td>
<td><strong>0.041</strong>*</td>
<td>0.415*</td>
</tr>
<tr>
<td>Self-reported study skills</td>
<td>4</td>
<td>0.397</td>
<td>0.81</td>
<td><strong>0.155</strong>*</td>
</tr>
<tr>
<td>First A&amp;P Course</td>
<td>2</td>
<td>0.414</td>
<td>0.662</td>
<td><strong>0.296</strong>*</td>
</tr>
<tr>
<td>Pre-health</td>
<td>2</td>
<td>0.064</td>
<td>0.938</td>
<td><strong>0.227</strong>*</td>
</tr>
</tbody>
</table>

d.f. – degrees of freedom
F – F-value from Two-way ANOVA analysis
P – P-value of the corresponding F-value.

**Student Difficulties Detected by the Written Essay and Oral Exams**

The written and oral exams, both scripted and non-scripted, were qualitatively analyzed using an open coding strategy. This method allowed for the identification of misconceptions and other student difficulties from assessment formats in which students were not prompted with answer choices. Analysis of student responses revealed frequent problems with four course-related concepts (“Physiological Concept”). From these concepts, misconceptions were identified and categorized and their frequencies determined for both the written and oral assessments (Table 3). Both the written and oral exams detected misconceptions related to all four concept categories. Qualitative analysis of the oral exams revealed more insight as to why students may hold these misconceptions. The most problematic concepts for students were “Shifts in the
Hemoglobin Dissociation Curve” and “Surfactant’s Role in Ventilation,” and problems with these concepts were almost equally detected by both the written and oral exams (Table 9). However, the oral exam was the only format able to detect additional student difficulties for “Graph Reading” and “Hemoglobin Saturation.”

Table 9: Frequency table of student difficulties detected by the written and oral exams

<table>
<thead>
<tr>
<th>Physiological Concept</th>
<th>Misconception</th>
<th>Exam Format</th>
<th>Written</th>
<th>Oral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Reading</td>
<td>Inverse Relationship</td>
<td></td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td></td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Hemoglobin Saturation</td>
<td>Misconception - Saturation causes Partial Pressure</td>
<td></td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Incomplete Definition</td>
<td></td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Shifts in the Hemoglobin Dissociation Curve</td>
<td>Backwards</td>
<td></td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Hemoglobin Levels Change</td>
<td></td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td></td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>Surfactant's Role in Ventilation</td>
<td>Lubrication</td>
<td></td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Related to gas exchange</td>
<td></td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Obstructed airways</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Miscellaneous</td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>23</td>
<td>21</td>
</tr>
</tbody>
</table>

Although the written and oral exams elicited similar frequencies of difficulties for the “Shifts in the Hemoglobin Dissociation Curve” and “Surfactant’s Role in Ventilation” categories, qualitative analysis of the oral exams allowed for the
determination of why students have these difficulties. Specific examples of the student misconceptions detected by the unscripted only oral exam are described below.

**Graph Reading – Basic Skills**

The hemoglobin question set tested student’s ability to interpret a specific graph known as the hemoglobin dissociation curve. Students were shown the following graph.

![Graph Image](image.png)

Students were first asked to label the axes of this graph, as this basic and foundational knowledge is crucial to more advanced interpretations of the graph. The written essay and scripted oral exam both revealed the problems that some students had labeling the axes. However, the unscripted oral interviews detected additional student difficulties in basic graph reading.

Even without knowing the axes labels, the student should be able to see that the graph shows a direct relationship between the independent and dependent variables; as one variable increases in value, the other increases as well. However, six of the unscripted oral interviews revealed an inability to make this interpretation, which would be considered a basic graph reading skill. Two of these six responses explicitly
indicated an inverse relationship between the variables. The following represents one example.

*Examiner:* “What is the relationship between CO2 [student response to x-axis] and oxygen saturation percentage [student response to y-axis]?”

*Student 15:* “So as CO2 increases, oxygen saturation would decrease. Right?”

**Hemoglobin Saturation**

In answering the axes-labeling questions, most students correctly identified the y-axis as percent saturation of hemoglobin, or at the very least, mention the term as something associated with the graph. Students are taught that saturation is the term used to define how many of the hemoglobin’s heme groups are bound to oxygen. A higher saturation means that more of the heme groups on hemoglobin are occupied by oxygen. Furthermore, students are taught that an increase in the partial pressure of oxygen will lead to higher percent saturation.

The written essay and scripted oral exam did not ask students to define percent saturation, making the assumption that if the student used the term, they understood its definition. However, the unscripted oral exams indicated that six students had misunderstood this concept. Two of the students held the misconception that percent saturation was what determined partial pressure of oxygen.
Student 3: “The less oxygen you dump, the less percent saturation, then the less partial pressure.”

Examiner: “…So saturation is the cause and partial pressure of oxygen is the effect?”

Student 3: “Yeah.”

Two students gave incomplete definitions of saturation, saying that oxygen was simply bound to blood.

Examiner: “What exactly is percent oxygen saturation?”

Student 14: “The amount of oxygen present in the blood.”

**Shifts in the Hemoglobin Dissociation Curve**

Students were then asked about “shifts” in the hemoglobin dissociation curve. Students are taught that the introduction of carbon dioxide to the globin portion of hemoglobin decreases hemoglobin’s affinity to oxygen. Thus, the heme portions of hemoglobin release any bound oxygen, a process known as oxygen unloading. Oxygen unloading results in a decreased hemoglobin percent saturation at every given partial pressure of oxygen. Graphically, this is demonstrated by re-drawing the dissociation curve to the right, or a “right shift.” Any changes that increase hemoglobin’s affinity to oxygen results in increased percent saturation, thus re-drawing the dissociation curve to the left, or a “left shift.” Examples of factors that would lead to a “left shift” include a decrease in carbon dioxide concentration, a decrease in temperature and an increase in pH.
The written essay and scripted oral exam were both able to elicit the student misconception that the introduction of carbon dioxide would cause a left shift in the curve, as opposed to a right shift. However, the interactive oral interviews gave some insight as to why this faulty logic may occur. Some students correctly stated that carbon dioxide would cause more oxygen unloading, but associated that with a left shift.

*Student 3*: “I don’t remember. Oh, well I remember. So if for left shifting we are unloading, more oxygen is being unloaded, so the percent saturation of oxygen is lower.”

Some students thought that shifts in the curve, either left or right, represented a change in the levels of hemoglobin.

*Examiner*: “Do you know what a left shift means?”

*Student 13*: “A lower amount of hemoglobin.”

Some students associated shifts with changes in ventilation, indicating that a person would not be able to hold their breath as long.

*Student 16*: “That, what I got this. I know this one. Your body produces too much CO2 so it needs to release it. So, you would have to take another breath as soon as you reach threshold.”

**Surfactant’s Role in Ventilation**

The last question in this question set pertained to a separate topic in respiratory physiology. The question tested the student’s knowledge of surfactant in the context of
a critical thinking question. Students are taught that water molecules present in the alveoli of lungs can exert surface tension. This surface tension can prevent the alveoli from expanding and potentially cause the alveoli to collapse. Students are then taught that a substance known as surfactant breaks up this surface tension in order to help the alveoli expand and function properly. For this particular question, the written exam, scripted oral exam, and unscripted oral exam were all able to detect various student difficulties, including misconceptions, about surfactant. A frequent misconception was that surfactant normally acts as a lubricant. In this case, the student possibly mistakes surfactant for the serous fluid in the pleural cavity.

*Student 17:* “So it is a lubrication. So essentially, without the lubrication, it will not be able to expand or contract as freely.”

Some students thought that decreased function of surfactant would somehow obstruct the air vessels of the conducting zone.

*Student 46:* “So I figured that the surface area was either increasing or decreasing. But it causes breathing problems. So I figured, that something must be constricting the airways.”

Two students gave responses that were partially correct. For example, the following student correctly says that decreased function of surfactant would lead to decreased surface. However, the student shows that an inability to define basic definitions, let alone be able to demonstrate an understanding that an increased surface tension leads to decreased surface area in the alveoli.
Examiner: “Deceased functional surfactant will cause breathing problems because...”

Student 9: “Because there is not enough surface area.”

Examiner: “... How did you get form surfactant to not enough surface area?”

Student 9: “Just basically, I don’t remember the word surfactant is, but from the prefix pat of the word, surfactant, I got surface, and I know surface are is a big thing. And so...”

Examiner: “And so why would not enough surface area lead to breathing problems?”

Student 9: “Not enough surface area on the lungs. I’m not really recalling why or how. But, I know something about surface area.”

The following example demonstrates that the student understands that a decrease in functional surfactant leads to decreased surface tension, but also demonstrates an inability to define what surface tension is.

Examiner: “What would surface tension be?”

Student 22: “How much pressure is onto it... how much pressure is on the cell.”

Examiner: “On the cell? So, what does surfactant do, then, to the surface tension?”
Student 22: “It decreases it.”

Examiner: “Okay. Why would you want to do that?”

Student 22: “Why would I want to decrease the surface tension?... I don’t know... Maybe it’s putting more pressure on the lungs.”

**Attitudinal Data**

Responses from a Likert survey regarding student attitudes towards the written exam were arithmetically subtracted from responses to identical statements about the oral exams. Positive differences indicated that students felt more strongly about the written exam for that particular statement, whereas negative differences indicated that students felt more strongly about the oral exam (Figure 16a-e). Students were also given the chance to further elaborate on their responses, which provided additional insight into their attitudes about the different testing formats.
Figure 16: Student responses on the Likert scale survey for the hemoglobin questions. Students were asked about (a) how well they were able to organize their thoughts, (b) how comfortable they felt, (c) how pressured they felt, (d) how rushed they felt, and (e)
how well the format could assess critical thinking. Responses were arithmetically subtracted (written – oral). Positive differences (to the right of the dotted line) represent students who felt more strongly about the written exam. Negative differences (to the left of the dotted line) represent students who felt more strongly about the oral exams. Differences of zero (on the dotted line) represent students who did not feel differently about the two formats.

When asked about their ability to organize their thoughts during the written and oral exams, only two students felt that they performed this task more successfully on the oral exam, and eight felt equally organized on both the written and oral exams. Not surprisingly, the majority (23) of students felt that they were better able to organize their thoughts on the written exam. (Figure 16a). The process of writing helped one student in organizing his/her thoughts. Similarly, another student did not like having to keep her thought organization all in his/her head during the oral exam.

In a related question about feelings of comfort with the two assessment formats, only three students felt more comfortable during the oral exam, whereas twenty students felt more comfortable during the written exam. Ten students felt equally comfortable on both the written and oral exams (Figure 16b). One student felt uncomfortable because he/she could not find the appropriate words to communicate his/her answers. Another student was afraid that the examiner would thinking negatively of him/her. In contrast, one student who had never taken an oral exam before was pleasantly surprised at how comfortable he/she felt during the oral exam.
In another question about the pressure that students felt on the two test formats, the majority (23) indicated feeling more pressure on the oral exam. Eight students felt an equal amount of pressure on both the written and oral exam. Only two students felt more pressure on the written exam (Figure 16c). One student commented that the written test felt less personal, thus easing the pressure off of the written exam. Another student felt that the quiet atmosphere of the written exam helped him/her concentrate. The unscripted portion of the oral exam placed pressure on another student because he/she could not predict what questions be asked.

A related question dealt with time management on the written and oral exams. Ten students indicated that they felt more rushed on the oral exam, while only four students felt more rushed on the written. Nineteen students felt equally rushed on both the written and oral exams (Figure 16d). One student felt rushed on the oral exam because he/she did not like thinking about the question for long periods of time. Another student said he/she felt rushed because he/she felt the need to keep talking and fill in moments of silence.

When asked about the ability of the written and oral formats to assess critical thinking, responses were divided fairly equally. Twelve students believed that the written exam was better for assessing critical thinking, while five students felt that this skill was more required more for the oral exam. Sixteen students felt that the written and oral exams assessed critical thinking equally. (Figure 16e). One student on the oral exam felt the need elaborate on his/her answer as opposed to just giving a simple answer.
Furthermore, student elaborations on the Likert-like survey also showed that these attitudinal categories were interrelated. For example, one student felt uncomfortable on the oral exam because of the silence. Because of the discomfort, the student felt that he/she felt rushed and unable to effectively organize his/her thoughts. Similarly, another student felt uncomfortable on the oral exam because he/she preferred to organize his/her thoughts on paper. When asked about the ability of both formats to assess critical thinking, many students referred back to feelings of discomfort and nervousness.

When asked if oral exams should be incorporated, eleven (11) students agreed or strongly agreed, seven students felt neutral, and thirteen student responses either disagreed or strongly disagreed. One student did not respond and was designated a score of zero (Figure 17). Students who disagreed talked about their own attitudes, saying that because they felt more pressured on the oral exam, it should not be used. Students who felt neutral or agreed looked past their own experiences and reflected on the actual worth of an oral exam. One student who felt neutral admitted that, although he/she might not perform as well on oral exams, that oral exams would help him/her get more out of the class. Similarly, another neutral student commented about how although students might feel more nervous on an oral exam, an oral exam would incentivize a deeper understanding of the material. Several other students who agreed to the use of oral exams mention how the oral exam forces students to truly understand the material. One student who agreed even mentioned how the oral exam not only tested for content
knowledge but student communication skills, commenting how those skills were necessary for students entering the medical field.

Figure 17 - Student responses to the statement “Oral exams should be incorporated into the curriculum for science courses.” For the hemoglobin questions.
DISCUSSION

In this study, data comparing the multiple-choice format to the written and oral formats showed that the multiple-choice format is not an appropriate substitute for either of the open-ended formats. Furthermore, quantitative analyses from two cohorts of students tested on two different physiological concepts showed consistent results. On the other hand, results from this study are inconsistent with direct correlations between free-response and choice type questions in previous studies (Paterson, 1926; Corey, 1930; Hurd, 1932). A key difference in this study was the use of more comparable exams. Unlike previous studies, this study kept the questions identical, changing only the presence or absence of answer choices. This allowed for more meaningful statistical results from the direct correlations. Furthermore, the direct correlations were supported with other statistical analyses, such as Kruskal-Wallis with post-hoc Mann-Whitney U tests and comparisons of absolute differences.

The two-way ANOVA analysis showed that most demographic factors did not account for the differences in students’ scores on the multiple-choice and open-ended assessments. However, one of the statistically insignificant variables has a positive implication. Major reports have focused on the achievement gap between genders in the science, technology, engineering, and math (STEM) fields, both at the professional and educational level (Hewlett, 2008; Simard, 2010). Previous studies looked specifically at the gender stereotype threat in science courses at the undergraduate level.
These studies found no difference between males and females in exam scores. Although not a focus of our study, the lack of a statistically significant difference between genders is consistent with these previous studies, and may suggest a closing of the achievement gap between genders in science courses.

Our results also found a lack of a statistical difference in the scores of students that had recently taken English courses and those that had not, as well as students that had taken oral exams in previous courses compared to those without such experiences. When combined with the statistically lower mean scores on the open-ended formats, the results question the ability of all students to utilize those skills in the context of a science course. Another related result is the statistically insignificant difference in performance between students that speak English as a second language and those whose native language is English. In fact, the only variable that showed a statistically significant result was the self-reported confidence level of students in the hemoglobin cohort, indicating a possible relationship between student confidence and differential scoring between the assessment formats for this physiological concept. However, the demographic analysis in this study is limited due to small sub-sample sizes. Future studies could replicate this study with a much larger sample size to further explore the potential influences that various demographic variables have on student performance and different assessment formats.

Between the written and oral formats, quantitative analysis showed little to no difference between student scoring, suggesting that both open-ended formats equally assess student learning. These results were consistent with the findings by Nehm and
Schonfeld in their comparison of assessment formats in an ecology and evolution course (2008). However, for the hemoglobin cohort, one question had to be removed from analysis, allowing us to explore how the number of questions could affect differential scoring. When the number of questions analyzed increased from four to five questions, the correlation coefficient ($R^2$) dropped from 0.706 to 0.473. Furthermore, the mean absolute difference increased from 0.303 to 0.794. Future studies could replicate this study with longer exams and examine how the quantitative analysis would change.

While no significant differences in student scores were found between the written and oral formats using quantitative methods of analysis, qualitative analysis did reveal some differences in student responses on the written and oral formats. Although both open-ended formats detected student difficulties equally, the oral format identified novel misunderstandings and misconceptions. Unlike the results of the quantitative analysis, these findings are contradictory to findings in Nehm and Schonfeld’s previous study (2008).

The qualitative analysis of this study also adds to the body of literature exploring misconceptions held by anatomy and physiology students. Michael’s study (1999) reported student misconceptions related to hemoglobin saturation causing partial pressure of oxygen in the blood, a misconception also found in this study. However, Michael’s study detected misconceptions solely through the use of a multiple-choice concept inventory. Furthermore, the researchers designed the inventory using only informal interactions with students and unpublished discussions with other faculty. The assessment in Michael’s study may have detected the student misconception of percent
saturation causing partial pressure because it was presented to the student as a possible answer choice. Our study elicited the same misconception through open-ended formats, thus students had to come up with explanations on their own. Data from the qualitative analysis in this study could provide a foundation for multiple-choice concept inventories in physiology, similar to those already designed for other biology courses (Garvin-Doxas, 2008; Smith, 2008).

Along with the misconception that percent saturation caused partial pressure, our study also detected other misconceptions and misunderstandings not found in the existing literature. Two student difficulties of importance related to 1) problems reading graphs in the hemoglobin cohort and 2) the application of the incorrect cardiac output equation in the hemodynamics cohort. This inability to use mathematical models in understanding physiological concepts showed a shortcoming in a competency outlined by the National Science Foundation in Visions and Change (Jungek, 1997; Brewer and Gross, 2003). Future studies could explore various interventions to help students develop mathematical and quantitative reasoning skills.

In 2006, the National Science Foundation began discussions on the future of biology education at the higher education level. By 2007, they were joined by the American Association for the Advancement of Science (AAAS), the Howard Hughes Medical Institute (HHMI), and the National Institutes of Health (NIH) in furthering these discussions with faculty and students from a variety of institutions ranging from the community colleges to research universities. In July 2009, the AAAS with the support of the NSF, held a conference to bring together the various conversations
throughout the years. This conference, “Visions and Change in Undergraduate Biology Education,” helped provide a framework and define expectations for the future of biology education.

The attitudinal data from the Likert-like surveys provided insight into student opinions about the written and oral formats. Overall, students felt more pressured and more rushed on the oral exam. Furthermore, students did not feel as comfortable, and some felt that they were unable to organize their thoughts as well on the oral exam. These attitudes may indicate an inability of students to effectively communicate their understanding of anatomy and physiology equally well on the oral and written formats.

Communication skills are another competency outlined in Visions and Change. This competency was included in the recommendations because scientists often work in teams that rely on effective communication between individuals. Furthermore, effective communication is particularly important for students in anatomy and physiology planning to pursue careers in healthcare fields. In terms of written communication, healthcare providers will need to learn to write out reports for their colleagues. For oral communication, providers will need to learn to talk not only with their colleagues but with their patients as well. Providing students with the opportunities to practice communicating the concepts of anatomy and physiology, in both written and oral formats, is crucial in developing necessary communication skills.

Although, overall, students held negative attitudes towards oral exams, the attitudinal data also showed that students felt that the oral exam could be used as a learning tool. Struvyen (2005) reviewed studies on how different exam formats affected
how students approached learning. Studies in this review showed that when students expect multiple-choice formats, they utilize surface approaches to learning; they learn only as much as needed to complete the expected tasks. However, when students expected essays, the studies show that students utilized deeper approaches of learning, striving to truly understand the material. Although our study did not look at whether expecting the oral exam changed student learning approaches, the attitudinal data showed that students felt they learned more when taking the oral exam. In light of these findings, future studies explore the use of oral exams as formative assessments and instructional tools for deeper learning.
APPENDIX A

California State University, Sacramento

Department of Biology

Information about Study & Consent Form

You are being asked to participate in a research project, which will be conducted by Adriel Cruz, a graduate student of the Department of Biology at California State University, Sacramento. The study will examine student performance on an oral examination compared to a written essay examination.

You will be asked to answer questions pertaining to material covered in the lecture portion of the Biology 26 – Anatomy and Physiology II course. These questions will be administered in both a verbal format as well as a written prompt. You will be asked to respond to the verbal format via spoken answers and to the written prompt via an essay. Each exam format will require up to 15 minutes of your time. After taking both exams, you may be asked further verbal questions in order to clarify your answers to your original responses.

During the verbal format, both you and the interviewer will be recorded. After participation in the study, the recording will be transcribed. Both the transcript of the oral interview as well as the written response will be kept private. To further maintain
your anonymity, your transcripts and essay responses will be tracked using random tracking numbers. Microcassettes, transcripts, and analyzed data will be stored in a locked file cabinet in a locked office.

After taking both assessment formats, you will be asked to fill out a questionnaire asking you to reflect on your experiences taking both the essay format as well as the verbal format.

To further analyze your relative performance, your responses to the multiple-choice exam administered during the courses will be obtained and compared. Your letter grade in the class will also be obtained for further comparison. This information will be kept confidential.

For participating in the oral and written essay assessment, you will be given points that will count towards your grade in the lab component of Biology 26. Your grades on other exams administered in either the lecture or lab component of Biology 26 will not be affected. If you do not wish to participate in this project, you will be given an alternative assignment that will count for the same number of points.

By participating in this research, you may gain additional insight into your learning style as well as your testing style. Findings from this research will add to the body of literature on assessment of science students.
All written responses, transcripts, and questionnaires will be destroyed upon completion of this project. Only aggregate or data made anonymous will be kept.

If at any time you do not wish to continue participation in this study, you are free to discontinue your involvement and you will be given an alternate assignment to complete for the same number of points.

If you have any questions about this research, you may contact Adriel Cruz, abc248@saclink.csus.edu.

Your participation in this research is entirely voluntary. **BY SIGNING BELOW, YOU CONFIRM THAT YOU ARE AT LEAST EIGHTEEN (18) YEARS OF AGE AND YOU ARE GIVING YOUR INFORMED CONSENT TO BE A PARTICIPANT IN THIS STUDY.**

___________________________________
Signature

___________________________________
Name (Print)

___________________________________
Date
APPENDIX B

Instructions:

- Please answer the questions to the best of your abilities.
- Please write to the side why you chose your answer choice. For example: “I chose answer A because it is was the longest answer choice”
- Please circle any words that may be confusing to you.

Type II Diabetes is a common, yet serious condition in the United States. In this condition, individuals are still able to produce insulin, but the body can no longer respond to it. Therefore, individuals with type II diabetes tend to have high levels of sugar in their blood.

1) High levels of blood sugar will effect which factor of resistance?
   a. Blood viscosity
   b. Blood vessel length
   c. Blood vessel radius
   d. Venous return

2) For the factor that you chose in question 1, high levels of blood sugar will cause the factor to
   a. Increase
   b. Decrease
   c. Stay the same

3) Therefore, overall resistance will
a. Increase
b. Decrease
c. Stay the same

4) In order to maintain the same cardiac output, what will happen to mean arterial blood pressure?
   a. Increase
   b. Decrease
   c. Nothing, mean arterial blood pressure will not be affected by blood sugar levels.

Individuals with type II diabetes also experience higher than normal levels of insulin. This may lead to blood vessel damage. This is particularly dangerous in blood vessels in the eye. Blood vessel damage in the eye may lead to a condition known as diabetic macula edema. In this condition, damaged blood vessels are known to be “leaky”.

5) Blood “leaking,” or more accurately, fluid moving out of the capillaries is known as what?
   a. Filtration
   b. Reabsorption
   c. Osmotic pressure
   d. Hydrostatic pressure
6) Normally, what physiological system plays a role in recycling fluid lost from the capillaries?
   a. Cardiovascular
   b. Digestive
   c. Lymphatic
   d. Urinary

One treatment for macular edema is the drug dexamethasone. Although the mechanism by which this drug works is complex, the ultimate goal of this drug is to reduce the amount of blood flowing through blood vessels.

7) Which force does dexamethasone affect?
   a. Interstitial fluid osmotic pressure
   b. Interstitial fluid hydrostatic pressure
   c. Capillary osmotic pressure
   d. Capillary hydrostatic pressure

8) For the force chosen in question 6, dexamethasone will
   a. Increase the force
   b. Decrease the force

9) How will dexamethasone affect capillary fluid movement?
   a. Increase filtration
   b. Decrease filtration
c. Increase reabsorption

d. Decrease reabsorption

Below is a graph that pertains to the heme group of hemoglobin. In a written essay, please address the following points to the best of your abilities.

10) What does the x-axis represent?

a. Total air pressure

b. Partial pressure of oxygen (O2)

c. Partial pressure of carbon dioxide (CO2)

d. Concentration of hemoglobin

11) What does the y-axis represent?

a. Hematocrit

b. Percent saturation of hemoglobin

c. Cardiac output

d. Blood pressure

12) How does an increase in carbon dioxide (CO2) change this curve?
a. A left shift, which leads to an increase in oxygen unloading
b. A right shift, which leads to an increase in oxygen unloading
c. A left shift, which leads to a decrease in oxygen unloading
d. A right shift, which leads to a decrease in oxygen unloading

13) Which of these will cause a left shift in the curve?
   a. An increase in BPG
   b. An increase oxygen (O2)
   c. A decrease in pH
   d. An increase in temperature

The Chernobyl incident is an infamous incident in which a nuclear power plant exploded. This caused the land around the plant to become extremely radioactive, causing health problems to the citizens living near the area. Scientists have found that one effect of the radioactive land is decreased function of surfactant.

14) Decreased function of surfactant will cause breathing problems because:
   a. There is an increased amount of friction in the respiratory passageways
   b. There is a decreased amount of friction in the respiratory passageways
   c. Alveoli may collapse due to an increase in surface tension
   d. Alveoli may collapse due to a decrease in surface tension
Student Evaluation of Concept Questionnaire

1. Did the wording of any of questions confuse you in any way? If so, please explain.

2. Were there any questions for which you felt that there was not an appropriate answer provided? If so, please explain.

3. Do you feel that you got any of the questions wrong as a result of the wording? In other words, did you understand the concept, but choose the wrong answer because of the way the question was written? If so, please elaborate; be sure to specify the answer you chose and why you did not choose the correct answer.
APPENDIX C

Instructions:

- Please answer the questions to the best of your abilities.
- Please circle any words that may be confusing to you.

Type II Diabetes is a common, yet serious condition in the United States. In this condition, individuals are still able to produce insulin, but the body can no longer respond to it. Therefore, individuals with type II diabetes tend to have high levels of sugar in their blood.

15) High levels of blood sugar will effect which factor of resistance?

16) For the factor that you chose in question 1, high levels of blood sugar will cause the factor to

17) Therefore, overall resistance will

18) In order to maintain the same cardiac output, what will happen to mean arterial blood pressure?

Individuals with type II diabetes also experience higher than normal levels of insulin. This may lead to blood vessel damage. This is particularly dangerous in blood vessels in the eye. Blood vessel damage in the eye may lead to a condition known as diabetic macula edema. In this condition, damaged blood vessels are known to be “leaky”.

19) Blood “leaking,” or more accurately, fluid moving out of the capillaries is known as what?
20) Normally, what physiological system plays a role in recycling fluid lost from the capillaries?

One treatment for macular edema is the drug dexamethasone. Although the mechanism by which this drug works is complex, the ultimate goal of this drug is to reduce the amount of blood flowing through blood vessels.

21) Which force does dexamethasone affect?

22) For the force chosen in question 6, dexamethasone will

23) How will dexamethasone affect capillary fluid movement?

Below is a graph that pertains to the heme group of hemoglobin. In a written essay, please address the following points to the best of your abilities.

24) What does the x-axis represent?

25) What does the y-axis represent?

26) How does an increase in carbon dioxide (CO2) change this curve?

27) What will cause a left shift in the curve?
The Chernobyl incident is an infamous incident in which a nuclear power plant exploded. This caused the land around the plant to become extremely radioactive, causing health problems to the citizens living near the area. Scientists have found that one effect of the radioactive land is decreased function of surfactant.

28) Decreased function of surfactant will cause breathing problems because:

**Student Evaluation of Concept Questionnaire**

1. Did the wording of any of questions confuse you in any way? If so, please explain.
2. Do you feel that you got any of the questions wrong as a result of the wording? In other words, did you understand the concept, but answered incorrectly because of the way the question was written? If so, please elaborate; be sure to specify the answer you chose and why you did not choose the correct answer.
APPENDIX D

Instructions:

Please validate the following assessment. Indicate if any of the questions are:

- Inappropriate for students taking BIO 26 (please keep in mind that although these students are technically non-biology majors, they also are typically in Sac State’s pre-nursing program, a Health Science major, or a kinesiology major)
- Scientifically inaccurate
- Not written clearly and/or precisely

Also, please categorize each question into one of the categories described in Bloom’s taxonomy (please see attached article)

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

Type II Diabetes is a common, yet serious condition in the United States. In this condition, individuals are still able to produce insulin, but the body can no longer respond to it. Therefore, individuals with type II diabetes tend to have high levels of sugar in their blood.
29) High levels of blood sugar will affect which factor of resistance?

   a. Blood viscosity  
   b. Blood vessel length  
   c. Blood vessel radius  
   d. Venous return

   Bloom level ____________________

30) For the factor that you chose in question 1, high levels of blood sugar will cause the factor to

   a. Increase  
   b. Decrease  
   c. Stay the same

   Bloom level ____________________

31) Therefore, overall resistance will

   a. Increase  
   b. Decrease  
   c. Stay the same

   Bloom level ____________________
32) In order to maintain the same cardiac output, what will happen to mean arterial blood pressure?
   a. Increase
   b. Decrease
   c. Nothing, mean arterial blood pressure will not be affected by blood sugar levels.

Bloom level ____________________

Individuals with type II diabetes also experience higher than normal levels of insulin. This may lead to blood vessel damage. This is particularly dangerous in blood vessels in the eye. Blood vessel damage in the eye may lead to a condition known as diabetic macula edema. In this condition, damaged blood vessels are known to be “leaky”.

33) Blood “leaking,” or more accurately, fluid moving out of the capillaries is known as what?
   a. Filtration
   b. Reabsorption
   c. Osmotic pressure
   d. Hydrostatic pressure

Bloom level ____________________
34) Normally, what physiological system plays a role in recycling fluid lost from the capillaries?
   a. Cardiovascular
   b. Digestive
   c. Lymphatic
   d. Urinary

   Bloom level______________

One treatment for macular edema is the drug dexamethasone. Although the mechanism by which this drug works is complex, the ultimate goal of this drug is to reduce the amount of blood flowing through blood vessels.

35) Which force does dexamethasone affect?
   a. Interstitial fluid osmotic pressure
   b. Interstitial fluid hydrostatic pressure
   c. Capillary osmotic pressure
   d. Capillary hydrostatic pressure

   Bloom level______________
36) For the force chosen in question 6, dexamethasone will
   a. Increase the force
   b. Decrease the force

   Bloom level__________________

37) How will dexamethasone affect capillary fluid movement?
   a. Increase filtration
   b. Decrease filtration
   c. Increase reabsorption
   d. Decrease reabsorption

   Bloom level__________________

Below is a graph that pertains to the heme group of hemoglobin. In a written essay, please address the following points to the best of your abilities.

38) What does the x-axis represent?
   a. Total air pressure
   b. Partial pressure of oxygen (O2)
c. Partial pressure of carbon dioxide (CO2)

d. Concentration of hemoglobin

Bloom level ____________________

39) What does the y-axis represent?

a. Hematocrit

b. Percent saturation of hemoglobin

c. Cardiac output

d. Blood pressure

Bloom level ____________________

40) How does an increase in carbon dioxide (CO2) change this curve?

a. A left shift, which leads to an increase in oxygen unloading

b. A right shift, which leads to an increase in oxygen unloading

c. A left shift, which leads to a decrease in oxygen unloading

d. A right shift, which leads to a decrease in oxygen unloading

Bloom level ____________________

41) Which of these will cause a left shift in the curve?

a. An increase in BPG
b. An increase oxygen (O2)

c. A decrease in pH

d. An increase in temperature

Bloom level ____________________

The Chernobyl incident is an infamous incident in which a nuclear power plant exploded. This caused the land around the plant to become extremely radioactive, causing health problems to the citizens living near the area. Scientists have found that one effect of the radioactive land is decreased function of surfactant.

42) Decreased function of surfactant will cause breathing problems because:

   a. There is an increased amount of friction in the respiratory passageways
   b. There is a decreased amount of friction in the respiratory passageways
   c. Alveoli may collapse due to an increase in surface tension
   d. Alveoli may collapse due to a decrease in surface tension

Bloom level____________________
Instructions:

Please validate the following assessment. Indicate if any of the questions are:

- Inappropriate for students taking BIO 26 (please keep in mind that although these students are technically non-biology majors, they also are typically in Sac State’s pre-nursing program, a Health Science major, or a kinesiology major)
- Scientifically inaccurate
- Not written clearly and/or precisely

Also, please categorize each question into one of the categories described in Bloom’s taxonomy (please see attached article)

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

Type II Diabetes is a common, yet serious condition in the United States. In this condition, individuals are still able to produce insulin, but the body can no longer respond to it. Therefore, individuals with type II diabetes tend to have high levels of sugar in their blood.
43) High levels of blood sugar will affect which factor of resistance?

   Bloom level _______________

44) For the factor that you chose in question 1, high levels of blood sugar will cause the factor to

   Bloom level _______________

45) Therefore, overall resistance will

   Bloom level _______________

46) In order to maintain the same cardiac output, what will happen to mean arterial blood pressure?

   Bloom level _______________

Individuals with type II diabetes also experience higher than normal levels of insulin. This may lead to blood vessel damage. This is particularly dangerous in blood vessels in the eye. Blood vessel damage in the eye may lead to a condition known as diabetic macula edema. In this condition, damaged blood vessels are known to be “leaky”.

47) Blood “leaking,” or more accurately, fluid moving out of the capillaries is known as what?
48) Normally, what physiological system plays a role in recycling fluid lost from the capillaries?

49) Which force does dexamethasone affect?

50) For the force chosen in question 6, dexamethasone will

51) How will dexamethasone affect capillary fluid movement?
Below is a graph that pertains to the heme group of hemoglobin. In a written essay, please address the following points to the best of your abilities.

52) What does the x-axis represent?

53) What does the y-axis represent?

54) How does an increase in carbon dioxide (CO2) change this curve?
55) What will cause a left shift in the curve?

Bloom level ________________

The Chernobyl incident is an infamous incident in which a nuclear power plant exploded. This caused the land around the plant to become extremely radioactive, causing health problems to the citizens living near the area. Scientists have found that one effect of the radioactive land is decreased function of surfactant.

56) Decreased function of surfactant will cause breathing problems because:

Bloom level ________________
APPENDIX F

Oral Assessment vs. Written Exam survey

Instructions: This survey is part of a science education research project in Biology 26 – Introductory Anatomy and Physiology II that is being conducted by Adriel Cruz. I appreciate your participation. This survey consists of 25 questions and should take 15 minutes or less to complete. If you feel uncomfortable answering any individual question while taking this survey you may choose “prefer not to answer” or simply leave the question blank.

1. Name: (Your name helps us with tracking through the semester, but will be removed for data entry and analysis purposes)

2. What is your age?

3. What is your biological gender?

   A) Female           B) Male           C) Prefer not to answer

4. Is English your primary language at home?

   A) Yes           B) No           C) Prefer not to answer

5. Is this your first semester to attempt college-level work (not including AP classes)?

   A) Yes
B) No-it is my 2^{\text{nd}} \text{ semester}

C) No-it is my 3^{\text{rd}} \text{ semester}

D) No-it is greater than my 3^{\text{rd}} \text{ semester}

E) Prefer not to answer.

6. What type of degree are you currently pursuing?

A) Associates  B) Bachelors  C) Other  D) Prefer not to answer.

7. How many semester hours are you taking this semester?

A) < 6 hours  B) 6-12 hours  C) 12-16 hours  D) 16+ hours  E) Prefer not to answer.

8. Is this the first time you are taking an Anatomy and Physiology course (not including Biology 25 – Anatomy and Physiology I)?

A) Yes  B) No  C) Prefer not to answer

9. Are you repeating Biology 26 – Anatomy and Physiology II

A) Yes  B) No
10. If the answer to question 9 is ‘yes’, why are you repeating this course?

A) Withdrew from previous course
B) Failed previous course
C) Passed previous course but did not receive a high enough grade to move on within the major
D) Needed refresher course
E) Prefer not to answer

11. When is the last time (in years) you took a Chemistry class? (Do not count this semester).

A) < 1 year  B) 1 year  C) 2 years  D) 3 years  E) >3 years

12. Please provide the name of the Chemistry class and type of institution (high school, college) where you took this class.

13. When is the last time (in years) you took an English class? (Do not count this semester)

A) < 1 year  B) 1 year  C) 2 years  D) 3 years  E) >3 years

14. Please provide the name of the English class and type of institution (high school, college) where you took this class.
15. Have you ever taken a class in which oral assessments were used as a graded exam?

A) Yes           B) No

16. If you answered ‘yes’ to #16, please provide the name of the class in which you took an oral assessment.

17. If you answered ‘yes’ to #16, how long ago (in years) did you take that class?

   A) <1 year  B) 1 year  C) 2 years  D) 3 years  E) >3 years

18. How confident are you in your ability to learn Biology?

   A) Very confident
   B) Somewhat confident
   C) Neutral
   D) Insecure
   E) Very insecure

19. Please elaborate on your answer to #18.
20. Do you feel prepared by your previous courses to do well in this class?
   A) Very prepared
   B) Adequately prepared
   C) Neutral
   D) Somewhat unprepared
   E) Very unprepared

21. Please elaborate on your answer to #20.

22. Do you feel that you have good study skills?
   A) Yes
   B) No
   C) I don’t know
   D) I never had to study for previous classes
   E) Prefer not to answer

23. How many hours a week (on average) do you study for this Biology 26 class?
   A) 0-1    B) 2-3    C) 4-6    D) 7-8    E) more than 8
24. What is your current or intended major?

A) Biology

B) Chemistry

C) Another science-related field (please specify)________________

D) Other (please specify) __________________

E) Undecided

25. Are you a pre-medical, pre-pharmacy, pre-veterinary, pre-dental, pre-
nursing or allied health student?

A) Yes    B) No    C) Prefer not to answer
APPENDIX G

BIO 26 – Research Project: Reflective Survey

Please take a few moments to reflect on your participation in this study and write down your thoughts. Responses to this reflection will not affect your score in either the written essay or oral exam. They will also not affect your grades or performance in either the laboratory or lecture component of the BIO 26 course.

Please answer according to the following rubric.

1- Highly disagree
2- Disagree
3- Neutral
4- Agree
5- Highly agree
I was able to organize my thoughts/answers on the **written essay**. (Please elaborate if necessary)

1 2 3 4 5

I was able to organize my thoughts/answers on the **oral exam**. (Please elaborate if necessary)

1 2 3 4 5

I felt comfortable taking the **written essay**. (Please elaborate if necessary)

1 2 3 4 5

I felt comfortable taking the **oral exam**. (Please elaborate if necessary)

1 2 3 4 5

I felt pressured/nervous while taking the **written essay**. (please elaborate if necessary)

1 2 3 4 5
I felt pressured/nervous while taking the **oral exam**. (Please elaborate if necessary)

1 2 3 4 5

I felt rushed taking the **written essay**. (please elaborate if necessary)

1 2 3 4 5

I felt rushed taking the **oral exam**. (please elaborate if necessary)

1 2 3 4 5

The **written essay** was able to assess critical thinking. (Please elaborate)

1 2 3 4 5

The **oral exam** was able to assess critical thinking (Please elaborate)

1 2 3 4 5
Oral exams should be incorporated into the curriculum for science courses. (Please elaborate)

Which format (written essay/oral exam) did you prefer? Why?

Please guess your score on the written essay (score from 0-5)

Please guess your score on the oral exam (score from 0-5)
APPENDIX H

Below is a graph that pertains to the heme group of hemoglobin. In a written essay, please address the following points to the best of your abilities.

- What do the x-axis and y-axis represent?
- What does the overall curve represent, i.e. what does it mean?
- How does carbon dioxide (CO₂) change this curve?
  - What do the changes in the curve mean?
  - Why are these changes important? (think of oxygen unloading)
- Are there any other factors that might affect this curve?
APPENDIX I

Examiner: “You will now be tested on your knowledge of hemoglobin. The oral assessment will consist of two parts.

During the first part, I will read to you the question that you will need to answer. After reading the question, you will have ten (10) minutes to answer the question. You may ask me to repeat any part of the question at any time. If you finish answering the question before time is up, you may end the first part early by telling me that you are done. However, should you decide to end the first part early, you may not go back and continue to answer for the first part.

During the second part of the oral assessment, I will ask you follow-up questions based on the response given during the first part. This is to assess your ability to think scientifically.
Below is a graph that pertains to the heme group of hemoglobin. Please address the following points to the best of your abilities.

- What do the x-axis and y-axis represent?
- What does the overall curve represent, i.e. what does it mean?
- How does carbon dioxide (CO₂) change this curve?
  - What do the changes in the curve mean?
  - Why are these changes important? (think of oxygen unloading)
- Are there any other factors that might affect this curve?
APPENDIX J

BIO 26 – Research Project: Reflective Essay

Please take a few moments to reflect on your participation in this study and write down your thoughts. Responses to this reflection will not affect your score in either the written essay or oral exam. They will also not affect your grades or performance in either the laboratory or lecture component of the BIO 26 course. In your response, please think about these questions.

- What, in your opinion, are the pros and cons of the written exam.
- What, in your opinion, are the pros and cons of the oral exam.
- Which assessment format (written vs. oral) did you prefer and why?
- Assuming that the University has adequate resources to conduct oral assessments, should oral assessments be used as a normal form of assessment in science courses?

Thank you for your participation in this study.
APPENDIX K

Points for hemoglobin question:

1: x-axis

1: y-axis (+0.5 for hemoglobin saturation)

0.5: hemoglobin dissociation curve

0.5: correlation between pO2 and % saturation

1: CO2 causes a right shift

1: decreases Hb’s affinity to hemoglobin

1: increases oxygen unloading

1: needed for cells

1 for each: BPG, Temperature, pH
LITERATURE CITED


