


March 2019

## Investigating Student Conceptual Understanding of Structure and Function by Using Formative Assessment and Automated Scoring Models

Kelli Patrice Carter

University of South Florida, [kellcarter@mail.usf.edu](mailto:kellcarter@mail.usf.edu)

Follow this and additional works at: <https://scholarcommons.usf.edu/etd>

 Part of the [Biology Commons](#), [Physiology Commons](#), and the [Science and Mathematics Education Commons](#)

---

### Scholar Commons Citation

Carter, Kelli Patrice, "Investigating Student Conceptual Understanding of Structure and Function by Using Formative Assessment and Automated Scoring Models" (2019). *Graduate Theses and Dissertations*.  
<https://scholarcommons.usf.edu/etd/7761>

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact [scholarcommons@usf.edu](mailto:scholarcommons@usf.edu).

Investigating Student Conceptual Understanding of Structure and Function  
by Using Formative Assessment and Automated Scoring Models

by

Kelli Patrice Carter

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
with a concentration in Biology Education  
Department of Integrative Biology  
College of Arts and Sciences  
University of South Florida

Major Professor: Luanna Prevost, Ph.D.  
Stephen Deban, Ph.D.  
Philip Motta, Ph.D.  
Jeffrey Raker, Ph.D.

Date of Approval:  
March 20, 2019

Keywords: core concepts, lexical analysis, machine scoring, anatomy & physiology

Copyright © 2019, Kelli P. Carter

## DEDICATION

I dedicate this work to my Dad, Mike Carter. Without his constant love, support and guidance I would not be the woman I am today. Thank you for never giving up on me.

I would also like to dedicate this dissertation to my family and friends for their unwavering support and encouragement throughout this journey. I could not have done this without you!

## ACKNOWLEDGMENTS

I first would like to thank my advisor, Dr. Luanna Prevost, for the support and mentorship throughout this journey. This project would not have been possible without her advice and guidance. I would also like to thank my dissertation committee: Dr. Stephen Deban, Dr. Philip Motta, and Dr. Jeffrey Raker. They pushed me to think outside of the box and provided encouragement throughout this process.

Thank you to my undergraduate researchers: Kirsti Martinez, Bryan Macneill, Sachiko Mahabeer, and Nyasha Madzingaidzi, for the hours of assistance and for allowing me to be your mentor.

I would also like to thank the faculty, staff and graduate students in the Integrative Biology department for their assistance and support: Dr. Ben Predmore, Dr. John Lawrence, Dr. Jody Harwood, Christine Brubaker, Karena Nguyen, Charly Stinson, and Bryan Delius. Also thank you to Dr. Daniel Lende for chairing my committee and to Dr. Rick Pollenz for helping me to embrace my nonlinear journey. Thank you to my writing partner, Rebecca Campbell-Montalvo, for the countless hours spent in our online writing sessions. Finally, thank you to Darrin King for help with editing this manuscript.

This research was supported in part by the Porter Family Foundation, the National Science Foundation, the American Physiological Society, the Human Anatomy and Physiology Society, the Pasco Hernando State College Foundation and the University of South Florida.

## TABLE OF CONTENTS

List of Tables .....	v
List of Figures .....	viii
Abstract .....	x
Chapter One: Introduction .....	1
Formative Assessment .....	2
Constructed Response .....	3
Written Assessment .....	4
Conceptual Understanding .....	5
Research Goals and Hypotheses .....	6
References .....	7
Chapter Two: Instrument Development.....	9
Abstract .....	9
Introduction.....	10
Automated Scoring .....	11
Lexical Analysis.....	13
Machine Scoring.....	14
Research Objectives and Questions .....	16
Methods.....	16
Question Development and Administration.....	20
Question Development and Administration.....	16
Coding for Structure Relates Function by Using Logistic Regression .....	18
Human Scoring of Student Responses .....	18
Lexical Analysis.....	20
Classification of Student Responses by Using Logistic Regression.....	21
Coding for Scientific and Non-scientific Ideas Using Machine Scoring.....	24
Human Scoring of Student Responses .....	24
Machine Scoring .....	26
Student Interviews .....	29
Results.....	30
Human Scoring of Define and Give Example Questions .....	30
Lexical Analysis.....	31
Model performance: Logistic regression .....	32
Model performance: Accuracy of Human-Computer Agreement .....	35
Human Scoring of Remaining Questions .....	36
Machine Scoring.....	36

Model Performance: Confusion Matrix .....	38
Model Performance: Cohen's Kappa.....	39
Model Performance: Precision and Recall.....	40
Student Interviews .....	40
"Two layers of skin" question.....	40
Discussion.....	43
Research Question 1 .....	44
Model Performance.....	44
Limitations of Model Performance.....	46
Research Question 2 .....	50
Conceptual Understanding of the Structure-Function Relationship .....	50
Students Lack a Conceptual Framework for Structure and Function .....	54
Levels of Organization.....	55
Implications for Teaching.....	57
Conclusion .....	57
References.....	58
Chapter Three: Comparison of Two-year and Four-year Students.....	62
Abstract.....	62
Introduction.....	63
Research Question .....	65
Research Hypothesis.....	65
Methods.....	65
Question Development and Administration.....	65
Human Scoring .....	67
Statistical Analyses .....	68
Results.....	68
Topic 1: Integumentary System/Skin Layers.....	69
SRF Concept 1: Sensation.....	69
SRF Concept 2: Protection.....	69
SRF Concept 3: Regulation .....	69
Topic 2: Muscular System/Skeletal Muscle Contraction.....	70
SRF Concept 4: ATP Necessary for Contraction to End.....	71
SRF Concept 5: Myosin Binds to Actin .....	71
SRF Concept 6: Muscle Contracts due to Calcium .....	71
SRF Concept 7: Sarcomere Contractile Unit.....	72
SRF Concept 8: ATP no Longer Available .....	72
SRF Concept 9: Muscle Shortening.....	72
Topic 3: Digestive System/Small Intestine.....	73
SRF Concept 10: Absorption.....	73
SRF Concept 11: Digestion .....	74
SRF Concept 12: Secretion.....	74
SRF Concept 13: Protection.....	74
Topic 4: Cardiovascular System/Blood vessels.....	75
SRF Concept 14: Blood Pressure Regulation.....	75

Discussion .....	76
College Readiness .....	86
Research Question .....	77
Research Hypothesis .....	77
College Readiness .....	77
Academic Integration .....	79
Conceptual Understanding in A&P II .....	80
Conceptual Understanding of Structure-Function .....	82
Implications for Teaching .....	83
Conclusion .....	84
References .....	85
Chapter Four: Comparison of Question Features .....	88
A Note to Reader .....	88
Abstract .....	88
Introduction .....	89
Investigating Student Responses to Varying Question Features .....	91
Cognitive Level .....	91
Guiding Context .....	92
Question Sequencing .....	93
Research Questions .....	95
Research Hypotheses .....	95
Methods .....	95
Question Development and Administration .....	95
Cognitive Level .....	96
Guiding Context .....	96
Question Sequencing .....	97
Human Scoring of Responses .....	98
Computer-Automated Scoring .....	98
Statistical Analyses .....	100
Cognitive Level .....	100
Guiding Context .....	101
Question Sequencing .....	101
Student Interviews .....	102
Results .....	103
Cognitive Level .....	103
Topic 1: Integumentary System/Skin Layers .....	103
Topic 2: Muscular System/Skeletal Muscle Contraction .....	104
Guiding Context .....	105
Question 1: Muscular System/Skeletal Muscle Contraction/Rigor Mortis .....	107
Question 2: Digestive System/Small Intestine/Celiac Disease .....	108
Question 3: Cardiovascular System/Blood Vessels/Arteries and Arterioles .....	108
Question 4: Cardiovascular System/Blood Vessels/Arteriosclerosis .....	109
Student Interviews Related to Guiding Context .....	109

Question Sequencing .....	111
Human Scoring .....	111
Response Length.....	113
Lexical Analysis.....	113
Student Interviews .....	117
Discussion.....	118
Overall Research Question .....	119
Overall Research Hypothesis.....	119
Cognitive Level.....	119
Research Question .....	119
Research Hypothesis.....	119
Guiding Context.....	122
Research Question .....	122
Research Hypothesis.....	122
Question Sequencing .....	123
Research Question .....	123
Research Hypothesis.....	123
Implications for Teaching.....	125
Conclusion .....	126
References.....	126
Appendices.....	130
Appendix A: Tables .....	131
Table A.1. Description of conceptual rubric for each short answer question prompt.....	131
Table A.2. Metrics of model performance for each conceptual model .....	135
Table A3. Chi-square analysis of structure-function concepts from Institutional comparison.....	141
Table A4. McNemar analysis of cognitive level of structure-function concepts.....	143
Table A5. Chi-square analysis of guiding context of structure-function concepts.....	144
Table A6. Chi-square analysis of guiding context of structure-function concepts by course .....	145
Appendix B: Interview protocol for Anatomy and Physiology assessment .....	147
Appendix C: IRB Approval Letters .....	150
USF .....	150
HCC .....	152
PHSC.....	154
Appendix D: Publication consent from APS .....	155
APS Publication.....	156



## LIST OF TABLES

Table 2.1:	Short answer structure-function questions .....	17
Table 2.2:	Human scoring of student responses using 3 bin rubric .....	20
Table 2.3:	Example student responses from “Define” and “Give Example” questions, and categorization of student responses in SPSS Modeler .....	22
Table 2.4:	Hierarchical structure and function lexical categories from SPSS Modeler.....	23
Table 2.5:	Confusion matrix of student responses to “Define principle” and use of “structure” .....	25
Table 2.6:	Description of conceptual rubric for each short answer question prompt .....	25
Table 2.7:	Human scoring of student responses to “two layers of skin” question using conceptual rubric .....	27
Table 2.8:	Machine scoring algorithms in the ensemble.....	28
Table 2.9:	Logistic regression model for the question “Define the principle: form reflects function” .....	34
Table 2.10:	Logistic regression model for the question “Give an example of the principle form reflects function from the human body” .....	34
Table 2.11:	Accuracy and goodness of fit of logistic regression models for “Define the principle” and “Give an example of the principle” .....	36
Table 2.12:	Confusion matrix of student responses to “two layers of skin” and <i>function protection</i> category .....	38
Table 2.13:	Examples of types of disagreements between human-scored and machine-scored explanations .....	49
Table 3.1:	Short-answer structure-function questions administered at one 4-year institution and two 2-year institutions.....	66
Table 3.2:	Number of responses collected for structure-function questions administered at one 4-year institution and two 2-year institutions .....	67

Table 4.1:	Short answer questions administered to students in General Physiology and Human Anatomy and Physiology with question prompts at the understand and apply cognitive levels .....	97
Table 4.2:	Short answer questions administered to students in General Physiology and Human Anatomy and Physiology .....	99
Table 4.3:	Description of question format DX and XD. Each question format was administered to half of a General Physiology class .....	100
Table 4.4:	Number of responses collected for short answer structure-function questions at the understand and apply cognitive levels from students in General Physiology and Human Anatomy and Physiology .....	103
Table 4.5:	Number of responses collected for short answer structure-function questions with either prompt or no prompt to the structure-function relationship.....	106
Table 4.6:	Format DX structure lexical categories with frequency by question prompt and Fisher's Exact Test results comparing structure lexical categories by question prompt .....	115
Table 4.7:	Format DX function lexical categories with frequency by question prompt and Fisher's Exact Test results comparing function lexical categories by question prompt .....	116
Table 4.8:	Format XD structure lexical categories with frequency by question prompt and Fisher's Exact Test results comparing structure lexical categories by question prompt .....	116
Table 4.9:	Format XD function lexical categories with frequency by question prompt and Fisher's Exact Test results comparing function lexical categories by question prompt .....	117
Table A.1:	Description of conceptual rubric for each short answer question prompt .....	138
Table A.2:	Metrics of model performance for each conceptual model .....	142
Table A3:	Chi-square analysis of structure-function concepts from institutional comparison.....	147
Table A4:	McNemar analysis of cognitive level of structure-function concepts .....	149
Table A5:	Chi-square analysis of guiding context of structure-function concepts.....	150

Table A6: Chi-square analysis of guiding context of structure-function concepts by course .....	151
---	-----

## LIST OF FIGURES

Figure 2.1:	Human scoring of student responses to “Define” and “Give Example” questions .....	31
Figure 2.2:	Categories from lexical analysis of student responses to “Define the principle” .....	32
Figure 2.3:	Categories from lexical analysis of student responses to “Give an example of the principle” .....	32
Figure 2.4:	Frequency of occurrence of concepts scored as “1” (present) between human scored and machine scored explanations of the structure-function relationship in the “two layers of skin” question .....	37
Figure 2.5:	Frequency of occurrence of students linking structure and function in their responses for the eight constructed response questions.....	38
Figure 2.6:	“Two layers of skin” categories with training kappa and testing kappa values as a measure of model performance .....	39
Figure 3.1:	Percentage of student responses from 2-year and 4-year institutions that included integument structure-function concepts .....	70
Figure 3.2:	Percentage of students’ responses from 2-year and 4-year institutions that included muscle contraction structure-function concepts.....	73
Figure 3.3:	Percentage of students’ responses from 2-year and 4-year institutions that included small intestine structure-function concepts .....	75
Figure 3.4:	Percentage of students’ responses from 2-year and 4-year institutions that included blood pressure regulation structure-function concepts .....	76
Figure 4.1:	Percentage of student responses for “Two layers of skin” (Understand) and “Third degree burn” (Apply) questions that included integument structure-function concepts.....	104
Figure 4.2:	Percentage of student responses for “Contractile proteins” (Understand) and “Rigor mortis” (Apply) questions that included skeletal muscle contraction structure-function concepts .....	105

Figure 4.3:	Percentage of student responses from prompt SRF and no SRF prompt that included muscle contraction structure-function concepts.....	107
Figure 4.4:	Percentage of student responses from prompt SRF and no SRF prompt that included small intestine structure-function concepts.....	108
Figure 4.5:	Percentage of student responses from Arteries/arterioles and Arteriosclerosis questions with prompt SRF and no SRF prompt that included blood pressure regulation structure-function concepts .....	109
Figure 4.6:	Human scoring of student responses to format DX.....	112
Figure 4.7:	Human scoring of student responses to format XD.....	112
Figure 4.8:	Lexical categories contained in student responses to “Define the principle form reflects function”.....	114
Figure 4.9:	Lexical categories contained in student responses to “Give an example of the principle form reflects function” .....	114

## ABSTRACT

There has been a call from the national community of biologists and biology educators to increase biological literacy of undergraduate students, including understanding and application of core concepts. The structure and function relationship is a core concept identified by the wider biology community and by physiology faculty. Understanding of the core concept structure and function across multiple levels of organization may promote biological literacy. My research focused on the development of formative written assessment tools to provide insight into student understanding of structure and function in anatomy and physiology.

In chapter two I developed automated scoring tools to facilitate the evaluation of written formative assessment based on structure and function. Formative written assessments allow students to demonstrate their thinking by encouraging students to use their diverse ideas to construct their responses. However, formative written assessments are not often used in the undergraduate biology classroom due to barriers, such as time spent grading and the intricacy of interpreting student responses. Automated scoring, such as lexical analysis and machine scoring, can examine student thinking in formative written responses. The core concept structure-function provides a foundation upon which many topics in anatomy and physiology can be built across all levels of organization. My research focused on the development of formative written assessment tools and automated scoring models to provide insight into student understanding of structure and function. My research objective was to examine student understanding of a core concept in anatomy and physiology by using automated scoring. Ten short answer questions were administered to students in a junior-level General Physiology course and a sophomore level

Human Anatomy and Physiology course at a large Southeastern public university, and to students in Human Anatomy and Physiology courses at two Southeastern two-year colleges. Seventeen students were interviewed to determine if their responses to the short answer questions accurately reflected their thinking. Lexical analysis and machine scoring were used to build predictive models that can analyze student thinking about the structure-function relationship in anatomy and physiology with high agreement to human scoring. Less than half of the student responses in this study demonstrated conceptual understanding of the structure-function relationship. Automated scoring can successfully evaluate a large number of student responses in Human Anatomy and Physiology and General Physiology courses.

In chapter three I compared conceptual understanding of structure and function in 2-yr and 4-yr student responses. Anatomy and physiology is taught at a variety of institutions, including 2-year community colleges and 4-year research universities. Regardless of the type of institution offering anatomy and physiology, conceptual understanding of the structure-function relationship is necessary to understand physiological processes. The focus of my research was to compare conceptual understanding of 2-year versus 4-year anatomy and physiology students by using written formative assessment. I hypothesize that differences in students' academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance. Based on prior research, I predict that there will be a difference in conceptual understanding of the core concept structure and function between two-year and four-year students in anatomy and physiology, and that the students at the two-year institution will not perform as well as the students at the four-year institution, as measured by performance on the constructed response questions. Responses to eight short answer essay questions were collected from students at both types of institutions from students in human anatomy and physiology over

six semesters. My results demonstrated that there is a difference in conceptual understanding of the structure-function relationship between 2-year and 4-year students in anatomy and physiology with more 4-year students mentioning SRF concepts in their responses compared to the 2-year students. A potential reason for this difference may be college readiness. There was no difference in performance between institution types on structure-function concepts examined in the A&P II course. My results suggested that students may benefit from a focus on core concepts within the content of anatomy and physiology courses. This focus should occur in both the first and second semesters of anatomy and physiology. Instructors can use written formative assessment to allow students to demonstrate their conceptual understanding within the organ systems.

In chapter four I investigated how question features affect student responses to anatomy and physiology formative assessment questions. Short answer essay questions contain features which are elements of the question which aid students in connecting the question to their existing knowledge. Varying the features of a question may be used to provide insight into the different stages of students' emerging biological expertise and differentiate novice students who have memorized an explanation from those who exhibit understanding. I am interested in examining the cognitive level of questions, the use of guiding context/references in question prompts, and the order of questions, and how these features elicit student explanations of the core concept structure-function in anatomy and physiology. I hypothesized that varying the features of short answer questions may affect student explanations. Short answer questions based on the core concept 'structure-function' were administered to 767 students in a junior level General Physiology course and to 573 students in a sophomore level Human Anatomy and Physiology course at a large southeastern public university. Student responses were first human scored and



then scored by using lexical analysis and machine scoring. Students were interviewed to examine their familiarity with levels of organization and to confirm their interpretation of the questions. Students demonstrated more conceptual understanding of four of the structure-function concepts when answering the understand questions and more conceptual understanding of two structure-function concepts when answering the apply questions. The question prompts provided a different context which may have influenced student explanations. There was no difference in conceptual understanding of the structure-function relationship with and without the use of a guiding context in the wording of the question prompt. For question sequence, students performed better on the last questions in the sequence, regardless of whether the last question was easier or more difficult. Instructors should provide students with questions in varying contexts and cognitive levels will allow students to demonstrate their heterogeneous ideas about a concept.

## CHAPTER ONE INTRODUCTION

There has been a call from the national community of biologists and biology educators to increase biological literacy of undergraduate students (AAAS, 2011). Biological literacy includes understanding and application of core concepts, such as those identified by the Vision and Change Report: 1. evolution, 2. structure and function, 3. information flow, exchange and storage, 4. pathways and transformations of energy and matter, and 5. systems (AAAS, 2011). Structure and function is a core concept identified by the wider biology community, including physiology education researchers and physiology faculty (AAAS, 2011; Michael & McFarland, 2011). In a study by Michael & McFarland (2011), eighty-one college faculty from diverse institutions identified fifteen physiology core concepts. My research will focus on one of these core concepts, structure and function, which provides a foundation upon which many topics in anatomy and physiology may be built across all levels of organization. According to Michael and McFarland (2011), the relationship between structure and function is described as

“The function of a cell, tissue, or organ is determined by its form. Structure and function (from the molecular level to the organ system level) are intrinsically related to each other.” (p. 338)

The Human Anatomy and Physiology Society provides learning goals for students in Anatomy and Physiology (HAPS, n.d.), which include an understanding of structure and function defined as the ability to

“Use anatomical knowledge to predict physiological consequences and use knowledge of function to predict the features of anatomical structures.”

Understanding of this core concept serves as a foundation in the learning process and provides coherence of the other core concepts (Michael & McFarland, 2011). Understanding of the core concept of structure and function across multiple levels of organization may promote biological literacy. Within this framework, I will assess student understanding of the core concept of structure and function from the molecular to organ system levels of organization by using formative assessment.

### **Formative Assessment**

Formative assessment occurs during the process of learning and provides feedback to both the instructor and students (Bell & Cowie, 2001). Feedback to instructors and students occurs during learning, not afterwards. Instructors need to know students' existing conceptions and misconceptions in order to help them overcome misconceptions. Students must be given feedback about their existing conceptions, and how to modify their thinking, in order for learning to occur. Students can use this feedback to determine what information they need to study further and what adjustments in their thinking need to be made.

The purposes of formative assessment are to facilitate student learning and inform pedagogy (Bell & Cowie, 2001). Learning is an adaptive process in which students' mental schema are reconstructed based on formative feedback (Chi et al., 1981; Driver, 1989). The development of this mental schema is necessary for a student to apply concepts in appropriate contexts. As students develop subject matter expertise, there are corresponding changes in how they represent problems cognitively (Chi et al., 1981). Students' existing mental schema and their cognitive processes both must be taken into account to inform pedagogy. Formative assessment allows teachers to discover the effectiveness of learning activities within the classroom. Examples of formative assessment include clicker questions, case studies, group

worksheets, or think-pair-share activities. As a teacher gathers this formative assessment feedback and information about student thinking and learning, the teacher may make pedagogical adjustments to further support learning. Therefore, formative assessment is at the intersection of teaching and learning.

### **Constructed Response**

Constructed response (CR) questions are open-ended questions, such as short answer essay questions, drawing, or an oral examination, whereby students use their own knowledge to construct their responses rather than choose from a list of options such as responding to multiple choice questions (Martinez, 1991). CR questions provide advantages not afforded by multiple choice questions. In addition to formative assessment providing insight into student understanding, CR questions also identify student misconceptions (Birenbaum & Tatsuoka, 1987). CR questions reveal student thinking by allowing students to use their various knowledge elements to construct their written responses (Nehm & Haertig, 2012). Multiple choice questions have less diagnostic value, whereas CR questions have more diagnostic value and may be used to assess student conceptions and misconceptions (Birenbaum & Tatsuoka, 1987; Martinez, 1999). Multiple choice questions allow for guessing and response elimination strategies, while CR questions require a written response that does not permit guessing (Kuechler & Simkin, 2010; Martinez, 1991). By using CR questions, a student must generate knowledge and organize information both cognitively and within his or her written response; thus, CR questions may help to enhance critical thinking skills. This higher order thinking is similar to real-world tasks and is necessary for the development of biological literacy (Nehm & Haertig, 2012). However, there are drawbacks to CR questions. Resource constraints of time, money, and expertise limit the use of CR questions; formative feedback to students about their learning may be delayed due to these

constraints (Ha et al, 2011). There is also difficulty with the reliability and consistency of human grading of CR questions (Ha et al, 2011). Such drawbacks to the use of CR questions to provide insight into student thinking may be ameliorated by using automated scoring (Martinez, 1999). Automated scoring helps to alleviate resource constraints and inconsistency of human grading.

### **Written Assessment**

Written assessment is one form of constructed response. Writing is a tool for enhancing student higher-order cognitive skills (Marzano, 1993). When students are encouraged to write, they use metacognition to reflect, construct, and explore their ideas (Glynn & Muth, 1994; Keys, 1999). Student learning is not necessarily linear but involves the exploration of ideas. For example, journal writing allows a student to explore his or her thinking without fear of being graded. Informal journal writing encourages students to reflect on their understandings and misunderstandings as they communicate their ideas (Newell, 2006). As students write, they can organize distinct facts into a more coherent form. This process does not mean that the act of writing will inevitably lead to a better understanding of information. However, if a student has prior knowledge of a concept, the act of writing and organizing their ideas will aid in the learning process (Newell, 2006). If a student is developing conceptual understanding, informal writing may help him or her to clarify ideas (Glynn & Muth, 1994). Informal, unstructured writing further allows students to explore their thinking and construct knowledge (Glynn & Muth, 1994; Keys, 1999; Newell, 2006). Additionally, having students informally write will enhance their writing skills. Such writing proficiency is important for computer assisted scoring as students with poor writing skills have a difficult time accurately expressing their ideas, which may prevent computer assisted scoring (lexical analysis) from recognizing their knowledge (Nehm & Haertig, 2012).

## Conceptual Understanding

Conceptual understanding refers to the ability to apply knowledge in a variety of contexts. Formative assessment is crucial for conceptual understanding to occur (Bell, 1995; Bell & Cowie, 2001). Conceptual understanding necessitates students' awareness of their existing conceptions and misconceptions, and how to make modifications to their thinking, through formative feedback. Often, students resort to rote memorization of facts rather than conceptual understanding (Michael, 2007). However, most students in anatomy and physiology courses are destined for healthcare professions where rote memorization is insufficient.

This dissertation will focus on measuring students' conceptual understanding of the structure-function relationship in anatomy and physiology. The content of anatomy and physiology courses includes many complex processes. The amount of detail in an anatomy and physiology textbook can be overwhelming to students, and students often resort to memorizing facts and narratives (Michael, 2007). Students need tools to make sense of the complex processes in anatomy and physiology and find patterns in the details of the information. Conceptual understanding of the structure-function relationship can help students recognize the relationship during learning and apply it to new information being learned. However, conceptual understanding of structure-function is not prevalent, even though the core concept is an organizing principle for biology, especially for anatomy and physiology. Recognizing the structure-function relationship can help students to organize the details and make sense of anatomical and physiological processes. This conceptual approach to teaching and learning about core concepts in anatomy and physiology, although not new, is not well established in anatomy and physiology education. The use of formative written assessments is one approach that can help give instructors insight into student conceptual understanding.

## Research Goals and Hypotheses

My research focuses on the development of formative written assessment tools to provide insight into student understanding of structure and function. In chapter 2, I have developed formative written assessments and automated scoring models that reveal student understanding about the relationship between structure and function in anatomy and physiology. My research goal is to examine how automated scoring may be used to examine student understanding of core concepts in anatomy and physiology and to build predictive models that mimic human scoring of structure-function formative assessment questions. In chapter 3, I compare conceptual understanding of students in anatomy and physiology at two-year institutions and a four-year institution. I hypothesize that differences in students' academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance. Based on prior research, I predict that there will be a difference in conceptual understanding of the core concept structure and function between two-year and four-year students in anatomy and physiology, and that the students at the two-year institution will not perform as well as the students at the four-year institution, as measured by performance on the constructed response questions. In chapter 4, I examine the effect of question features, such as cognitive level, guiding context, and question sequence, on prompting student understanding of structure and function in anatomy and physiology. I hypothesize that varying the features of short answer questions (cognitive level, guiding context, and question order) may affect student writing about core concepts in anatomy and physiology. Based on prior research, I predict a difference in student responses and conceptual understanding of the core concept structure and function based on the cognitive difficulty, guiding context, and question sequencing of the short answer questions.

## References

- American Association for the Advancement of Science (AAAS) (2011). Vision and change in undergraduate biology education: A call to action.
- Bell, B. (1995). Interviewing: A technique for assessing science knowledge. In S. Glynn & R. Duit (Eds.), *Learning Science in Schools: research reforming practice*. New Jersey: Lawrence Erlbaum Associates.
- Bell, B. & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85, 536-553.
- Birenbaum, M. & Tatsuoka, K. K. (1987). Open-ended versus multiple choice response formats - it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4), 385-395.
- Chi, M.T., Feltovich, P.J., & Glasner, R. (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152
- Driver, R. (1989). Students' conceptions and the learning of science. *International Journal of Science Education*, 11(5), 481-490.
- Glynn, S.M. & Muth, K.D. (1994). Reading and writing to learn science: Achieving scientific literacy. *Journal of Research in Science Teaching*, 31(9), 1057-1073.
- Ha, M. Nehm, R. Urban-Lurain, M & Merrill, J.E. (2011). Applying computerized scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE-Life Sciences Education*, 10, 379-393.
- Human Anatomy and Physiology Society (HAPS), (n.d.). Learning goals for students in Anatomy and Physiology. Retrieved from <http://www.hapsweb.org/?page=LearningGoals>
- Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education*, 83(2), 115-130.
- Kuechler, W.L. & Simkin, M.G. (2010). Why is performance on multiple choice tests and constructed response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, 8, 55-73.
- Martinez, M. (1991). A comparison of multiple-choice and constructed response figural items. *Journal of Educational Measurement*, 28(2), 131-145.
- Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.



Marzano, R. J. (1993). How classroom teachers approach the teaching of thinking. *Theory Into Practice*, 32(3), 154-160.

Michael, J. (2007). What makes physiology hard for students to learn? Results of a faculty survey. *Advances in Physiology Education*, 31, 34-40.

Michael, J. & McFarland, J. (2011). The core principles ("big ideas") of Physiology: results of faculty surveys. *Advances in Physiology Education*, 35, 336-341.

Nehm, R.H. & Haertig, H. (2012) Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, 21, 56-73.

Newell, G. (2006). Writing to learn: How alternative theories of school writing account for student performance. In *Handbook of Writing Research*. Eds. pp. 235-246

## CHAPTER 2 INSTRUMENT DEVELOPMENT

### **Abstract**

Formative written assessments allow students to demonstrate their thinking by encouraging students to use their diverse ideas to construct their responses. However, formative written assessments are not often used in the undergraduate biology classroom due to barriers, such as time spent grading and the intricacy of interpreting student responses. Automated scoring, such as lexical analysis and machine scoring, can examine student thinking in formative written responses. The core concept structure-function provides a foundation upon which many topics in anatomy and physiology can be built across all levels of organization. My research focused on the development of formative written assessment tools and automated scoring models to provide insight into student understanding of structure and function. My research objective was to to examine student understanding of core concepts in anatomy and physiology by using automated scoring. Ten short answer questions were administered to students in a junior-level General Physiology course and a sophomore level Human Anatomy and Physiology course at a large Southeastern public university, and to students in Human Anatomy and Physiology courses at two Southeastern two-year colleges. Seventeen students were interviewed to determine if their responses to the short answer questions accurately reflected their thinking. Lexical analysis and machine scoring were used to build predictive models that can analyze student thinking about the structure-function relationship in anatomy and physiology with high agreement to human scoring. This study showed that less than half of the student responses in this study demonstrated

conceptual understanding of the structure-function relationship. Automated scoring can successfully evaluate a large number of student responses in Human Anatomy and Physiology and General Physiology courses.

## **Introduction**

Formative assessment occurs during learning (Bell & Cowie, 2001). Written assessment includes open-ended questions wherein students use their own knowledge to construct their response rather than by choosing a multiple-choice option (Martinez, 1991). Formative written assessments facilitate student learning and can enhance pedagogy by providing feedback to both instructors and students (Bell & Cowie, 2001). Formative written assessments allow students to demonstrate their thinking by encouraging students to use their diverse ideas to construct their responses. However, formative written assessments are not often used in the undergraduate biology classroom.

Barriers to the use of written assessments include time spent grading and training other graders, inconsistency in grading due to fatigue, and the intricacy of interpreting student responses (Nehm & Haertig, 2012). There is also difficulty with the reliability and consistency of human grading of written assessments (Ha et al., 2011). Subjectivity is a problem with human grading, which leads to issues with reliability. Multiple graders are often necessary in large enrollment courses, and training is time-consuming. In addition, multiple graders may not be comparatively consistent in their grading. Formative feedback provided to students about their learning may thus be delayed due to such constraints (Ha et al., 2011). However, the effectiveness of written assessments to provide insight into student thinking may be augmented by using automated scoring (Martinez, 1999).

## Automated Scoring

Automated scoring, such as lexical analysis and machine scoring, resolves some of the concerns that limit the use of written formative assessment. Automated scoring is a branch of computer science in which computers learn through experience (Abu-Mostafa, 2012). In this way, computers can utilize patterns to build predictive models that can then be used to evaluate future cases. The computer matches characteristics of responses with the scores assigned by human scorers, then uses the characteristics to predict scores on new responses; this algorithmic process is similar to how Netflix recommends content to its viewers based on prior user-selection or how Amazon makes predictions about one's future purchases (Abu-Mostafa, 2012). Thus, the development of automated grading tools will assist with reliability and consistency of grading, alleviate grading fatigue, and diminish scorer training time. Automated scoring may facilitate the analysis of written responses from large enrollment classes and provide formative feedback to instructors and students in a timely manner as it is capable of scoring a large number of student written responses in a short amount of time.

There are numerous examples of automated scoring tools being used in education. In each of the following examples, patterns are used to build predictive models, future essays are compared to those models, and scores are assigned. For example, Project essay grader (PEG) evaluates formative and summative essays for writing quality by analyzing essays for fluency, word choice, and grammar (Page, 1994). Intelligent essay assessor (IEA) uses latent semantic analysis to evaluate text for specific words and phrases (Foltz, et al., 1999). IEA is a Pearson tool for WriteToLearn, which is a web-based tool for improving writing skills and reading comprehension (Foltz et al., 2013). IEA assesses formative and summative essays for quality as well as spelling and grammar. E-rater is the tool used by the Educational Testing Service to

evaluate formative and summative essays by using natural language processing (Burstein, Kukich, Wolff, Lu & Chodorow, 1998). E-rater scores essays based on the presence or absence of “features” of quality writing, such as lexical complexity, grammar, mechanics, and organization (Attali & Burstein, 2006). E-rater is used in high-stakes assessments and standardized tests including the GRE and GMAT.

Within science education, automated scoring, including lexical analysis and machine scoring, is used to examine conceptual understanding. For instance, Prevost et al. (2016) used lexical analysis to identify terms that students used in their written responses about replication, transcription, and translation. They next used this information to build predictive scoring models of student explanations of the central dogma (Prevost et al., 2016). Similarly, Weston et al. (2015) used lexical analysis to investigate student understanding of and misconceptions about the process of photosynthesis. They found that the content of student responses varied by changing the question prompt order and the plant species (Weston et al., 2015). Haudek et al. (2012) used lexical analysis to reveal terms and phrases students used to explain the chemistry of acids and bases in the context of cellular biology. They then used the terms and phrases to predict human scoring of written responses (Haudek et al., 2012). Ha et al. (2011) used machine scoring to analyze key concepts of evolution, including variation, heredity and limited resources. For their study, responses were collected from majors and non-majors at two different institutions. They found that sample source did not affect the performance of the machine-scoring models. However, the frequency of occurrence of concepts was associated with model performance: e.g., “competition” occurred in 0.02% of the student responses, which corresponded with poor model performance (Ha et al., 2011). They also investigated sample size and model performance and

found that sample sizes of 500 compared to 1,000 did not have significant effect on model performance (Ha et al., 2011).

This chapter builds upon one study that used automated scoring/lexical analysis to examine student understanding of the structure-function relationship in physiology (Carter & Prevost, 2018).

### Lexical Analysis

Lexical analysis uses linguistic-based computer analysis to identify, extract and categorize text (Nehm & Haertig, 2012). Lexical analysis has been described as a “bag of words” model in which words are extracted from a text document. With a “bag of words” model, the order of the diction and grammar is ignored during extraction (Zhang et al., 2010). Lexical analysis has been used for marketing research to thematically evaluate open-ended survey responses (Espinoza et al., 2018). For example, if asked “How did you like your hotel stay?”, responses could include:

- a. I would definitely recommend this hotel. The location was great!
- b. Had I known the hotel would be so noisy, I would not have chosen it for my work trip since I needed quiet time to work.
- c. The rooms were decent, and the bed was comfortable.

Lexical analysis would explore these types of responses and provide a quantitative view of terms and phrases in the responses. Once terms and phrases are identified in the responses, categories could be formed, such as location (a), noise level (b), and room satisfaction (c) from the above examples. This automated analysis of open-ended survey responses could occur more quickly and be more consistent than human scoring (Espinoza et al., 2018).

Prior work in science education has shown lexical analysis, extraction, and categorization to reliably reveal student thinking (Haudek et al., 2012; Weston et al., 2015; Prevost et al., 2016; Carter & Prevost, 2018). Haudek et al. (2012) performed lexical analysis, including extraction

and categorization, on student responses to an acid-base chemistry constructed response question. This two-step, linguistic-based approach to lexical analysis using SPSS Modeler was used by Prevost et al. (2016) in which categories were created to categorize student responses to genetics questions. Additionally, Weston et al. (2015) analyzed student responses to photosynthesis questions by using lexical analysis with both extraction and categorization. Lexical analysis also was successfully used to analyze physiology students' written responses to structure-function questions in which the order of question prompts varied (Carter & Prevost, 2018).

### Machine Scoring

In supervised machine scoring, the computer “learns” the rules of scoring student responses from human scoring (Kotsiantis, 2007). The machine scoring uses a set of human-scored student responses to discover patterns, such as the presence or absence of physiological concepts relating to the structure-function relationship. Thus, the human scoring of student responses is taken into account by the machine scoring algorithms, which learn patterns from the human scoring to classify the written responses and mimic human scoring. The patterns detected in the student responses can then be applied to a new set of student responses. Both correct and incorrect ideas can thus be recognized and classified by the software.

As an example of machine scoring, a bank has a large customer base, and many of their customers have various loans. The bank knows the characteristics of people who make timely payments on their loans, such as debt to income ratio, credit score, and other credit obligations. Machine scoring algorithms detect the patterns of people who make their loan payments. When a customer applies for a loan, the machine scoring algorithms use that customer's characteristics to predict whether or not he or she will make the loan payments based on patterns that the

algorithms detected in data from previous customers. The bank uses that information to decide whether or not to offer the customer a loan. In the same vein, machine scoring may be used to build computer scoring models of students' written responses. These predictive scoring models thus would mimic human scoring.

Prior work in education has demonstrated machine scoring to reliably detect patterns of student thinking in evolution (Ha et al., 2011; Nehm et al., 2012). A high level of agreement ( $\kappa$ ) between human scoring and machine scoring was assessed in student responses for key concepts related to evolutionary change (Ha et al., 2011). Livne et al. (2007) used machine scoring of mathematics questions to evaluate student responses in the form of mathematical equations by using a holistic rubric; equations were assessed as correct, partially correct or incorrect. In broader strokes, machine scoring has also been used to predict student performance in distance learning classes (Kotsiantis et al., 2004; Kotsiantis, 2012). The predictive models considered student demographics, prior educational experience, and whether or not the student sought help from a tutor to predict academic achievement. This information was used to design tutoring interventions with the goal of students' academic success (Kotsiantis, 2012).

One of the main differences between lexical analysis and machine scoring is that human scoring of written responses to interpret lexical expressions is needed prior to machine scoring to train the computer for patterns to detect. Lexical analysis can be used initially to identify terms and phrases in student responses in an exploratory fashion. Therefore, machine scoring has the ability to function as a confirmatory analysis, which will measure the deviation from the human scoring (Haudek et al., 2011; Nehm et al., 2012). Neither lexical analysis nor machine scoring is capable of detecting meaning in students' written responses, yet both are quite sensitive to words



and patterns. Further work using lexical analysis and machine scoring is needed to determine the appropriateness of each method for science education research.

My investigation of lexical analysis and machine scoring is focused on a core concept in physiology education: the structure-function relationship (AAAS, 2011, Michael et al., 2009). Automated scoring, such as lexical analysis and machine scoring, can examine student thinking about the structure-function relationship in formative written responses. In this study, I use lexical analysis to evaluate two structure-function short answer questions, and I use machine scoring to assess eight short answer questions.

### Research Objectives and Questions

1. How can automated scoring methods, such as lexical analysis and machine scoring, be used to build predictive models that mimic human scoring of structure-function formative assessment questions?
2. What do the predictive models built from automated scoring demonstrate about student conceptual understanding of the structure-function relationship in physiology?

### Methods

#### Question Development and Administration

I developed ten short answer questions based on the core concept of “structure-function” (Table 2.1). The questions were developed with feedback from an anatomy and physiology instructor, two physiology instructors, and two science education researchers. I also interviewed students for their feedback and interpretation of the questions. The study protocol was approved under the Institutional Review Board (Pro00027955, HCC IRB #2017\_009), and students provided consent prior to participation.

The short answer questions were administered to students in a junior-level General Physiology course and a sophomore level Human Anatomy and Physiology course at a large

**Table 2.1.** Short answer structure-function questions. GP=General Physiology, HAP= Human Anatomy & Physiology

Question prompt	Topic	N GP	N HAP
Define the principle: form reflects function	Concept definition	222	318
Give an example of the principle: form reflects function from the human body	Concept example	484	319
Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.	Integumentary system/Skin layers	0	597
Victims of third degree, or full thickness, burns have their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin.	Integumentary system/Skin layers	149	458
The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction.	Muscular system/Skeletal muscle contraction	262	462
A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.	Muscular system/Skeletal muscle contraction	463	509
Consider the mucosa of the small intestine. Based on form reflecting function, explain how this layer contributes to the functions of the digestive system.	Digestive system/Small intestine	0	314
Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system.	Digestive system/Small intestine	349	370
Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation.	Cardiovascular system/Blood vessels	334	376
Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.	Cardiovascular system/Blood vessels	465	311
Totals		2728	4034

Southeastern public university, and the short answer questions were also administered to students in Human Anatomy and Physiology courses at two Southeastern two-year colleges. The General Physiology course focuses on the structures and metabolic processes that vertebrate and invertebrate animals use to interact with their environment; the structure-function relationship is an underlying or implicit concept in the General Physiology course. The Human Anatomy and Physiology course is a sequential, two-term course (Human Anatomy & Physiology I and II) designed to introduce the form and function of the human body; the structure-function relationship is an explicit concept in the Human Anatomy and Physiology course.

The questions were administered throughout each semester over eight semesters as part of regular online homework via the course management system. Administration of each question occurred after the relevant topic was discussed in class. Students were asked to explain their answer to the best of their ability without the use of outside resources. I collected a total of 6,762 responses over the course of those eight semesters from 1,777 students at the 4-yr institution and from 437 students at the 2-yr institutions.

I analyzed the student responses to the short answer questions by using two approaches. The first approach was human scoring for the structure-function relationship followed by logistic regression. The second approach included human scoring for scientific and nonscientific ideas related to the structure-function relationship followed by machine scoring.

Coding for Structure Relates Function by Using Logistic Regression

#### Human Scoring of Student Responses

For the questions, “Define the principle: form reflects function” and “Give an example of the principle: form reflects function from the human body”, I used a 3 bin analytic rubric to code for the presence (1) or absence (0) of the concepts of *structure* and *function* and whether students

*relate structure and function* (Table 2.2). For example, responses that mentioned “teeth”, “vili (sp)”, “small intestine”, “female pelvis”, and “membrane” were coded 1 (present) for structure as shown in bold font in Table 2.2. Adjectives used to describe structures, such as “pointed”, “flat”, “large”, and “wide” were also coded 1 (present) for structure. Responses that mentioned “tearing”, “grinding”, “pumping of blood”, “heat loss”, and “transport” were coded 1 (present) for function as shown in Table 2.2. Responses that included a correct statement connecting structure and function were coded 1 (present) for the structure-relates-function concept. For example, in Table 2.2, “The female pelvis is large and wide for childbearing” was coded as a 1 for structure (female pelvis/large/wide), a 1 for function (childbearing), and a 1 for structure-relates-function because the student demonstrated the connection between the two. However, responses that mentioned a structure and a function but did not provide a correct statement linking the two were coded 0 for structure-relates-function: e.g., in Table 2.2, “Transport across the membrane” was coded as a 1 for structure (membrane) and a 1 for function (transport) but as a 0 for structure-relates-function because there was no connection between the two noted (Carter and Prevost, 2018).

We obtained inter-rater reliability by scoring a subset of responses (15%) with two coders. An inter-rater reliability of  $>0.70$  (kappa) was achieved (Landis & Koch, 1977), and then I coded the remaining responses. Analysis of the human scoring data consisted of determining the percent of student responses that mentioned structure, function or the concept of structure relating function.

**Table 2.2.** Human scoring of student responses using 3 bin rubric. Within student responses, structures are highlighted in bold and functions are underlined.

Student response	<b>Structure</b>	<u>Function</u>	Structure relates function
Some <b>teeth</b> are made for <u>tearing</u> ; therefore, they are <b>pointed</b> . Other <b>teeth</b> and meant for <u>grinding</u> , so they are <b>flat</b> .	1	1	1
<b>vili</b> ( <i>sic</i> ) of the <b>small intestine</b> .	1	0	0
<u>Pumping of blood</u> through body	0	1	0
The <b>female pelvis</b> is <b>large</b> and <b>wide</b> for <u>childbearing</u> .	1	1	1
<u>heat loss</u> and <u>preservation</u> to <u>maintain homeostasis</u> during <u>work</u> .	0	1	0
<u>Transport</u> across the <b>membrane</b>	1	1	0

### Lexical Analysis

The first step of lexical analysis is extraction, and the process involved the identification of text by building a custom lexical library using student responses. Student responses to “Define the principle: form reflects function” and “Give an example of the principle: form reflects function” were analyzed by using IBM SPSS Modeler Text Analysis version 16 (SPSS, 2013). The software has a standard library of common term, and a custom lexical library was built by using data from student responses. The custom library includes synonyms, abbreviations, variant spellings and misspellings as well as discipline-specific, technical terms used in physiology courses. For example, synonyms and misspellings of “absorption” included “absorptive”, “absortion”, and “absorbtion”, and all such terms were added to the custom lexical library.

Extraction occurred as the software identified key terms from the student responses by using the standard and custom lexical libraries. Examples of terms in the lexical libraries and

identified by the software are shown in Table 2.3, e.g., in the first response in Table 2.3, the software recognized the terms “femur”, “support” and the phrase “thick-walled”, while in the fifth response, only “process” was recognized.

The second step of lexical analysis is categorization, in which terms and phrases identified by the software are grouped into categories. (Throughout this chapter, categories will be represented in italics.) A category includes terms and phrases that represent a common meaning or a homogenous idea. For example, the category *structure/organ level* includes the terms “femur”, “small intestine”, “capillaries”, and “lungs.” The terms “thick walled”, “surface area”, and “biconcave shape” are used to describe structures and are considered *properties of structures* (Table 3). The student’s written responses were categorized into zero, one, or more categories following extraction: e.g., the student response of “The femur is a thick-walled long bone because of its support function for the trunk of the body” is categorized as *properties of structures*, *structure/organ level*, and *function/general* (Table 2.3). The category grain size was hierarchical based on biological levels of organization from molecular to organ system (Table 2.4) (Carter & Prevost, 2018).

#### Classification of Student Responses by Using Logistic Regression

Classification was performed once categorization was complete by using SPSS Modeler to build a predictive model of human scoring of student responses. Logistic regression uses a forward stepwise method to identify a subset of lexical categories that predict human scoring. Logistic regression through SPSS Statistics was used to determine the categories that predict the presence or absence of the human coding: structure, function, and structure-relates-function. A unique model was created for each category; thus, for each question, three logistic regression

models were built. Logistic regression was used because the dependent (response) variable, the prediction of human coding, is a binary variable: e.g., when building a model to predict student

**Table 2.3.** Example student responses from “Define” and “Give Example” questions, and categorization of student responses in SPSS Modeler.

Response	Category					
	<i>Properties of structures</i>	<i>Structure/cellular level</i>	<i>Structure/organ level</i>	<i>Process</i>	<i>Function/cellular level</i>	<i>Function/general</i>
The femur is a thick-walled long bone because of its support function for the trunk of the body	x		x			x
Microvilli on the epithelial cells of the small intestine for faster absorption due to larger surface area.	x	x	x		x	
The loss of functions associated with a third degree burn would be the loss of sensation, loss of some movement, and loss of protection.						x
Red blood cells have a very distinctive biconcave shape that allows them to squeeze into the smallest capillaries in the human body.	x	x	x			x
A process by which a physiological reaction takes place in an organism.				x		
The alveoli of the lungs give a high surface area for gas exchange.	x	x	x		x	

**Table 2.4.** Hierarchical structure and function lexical categories from SPSS Modeler. Table from Carter and Prevost (2018).

Structure	Function	Other
<i>structure</i>	<i>function</i>	<i>dynamics</i>
<i>structure/biomolecules</i>	<i>function/cellular level</i>	<i>mechanism</i>
<i>structure/cell</i>	<i>function/organ level</i>	<i>organism</i>
<i>structure/cell components</i>	<i>function/organ system level</i>	<i>process</i>
<i>structure/tissue</i>	<i>function/organism level</i>	
<i>structure/tissue components</i>	<i>function/general</i>	
<i>structure/organ</i>	<i>function/disorder</i>	
<i>structure/organ components</i>		
<i>structure/organ system</i>		
<i>structure/part</i>		
<i>structure/complex structures</i>		
<i>properties of structures</i>		

understanding of the concept of structure, the response variable of structure has two values in the student response: the presence or absence of structure. The independent (predictor) variables for the logistic regression are the binary variables of the presence or absence of a student's response in a lexical analysis category. The logistic regression model thus predicts the likelihood that a response would be classified as either correct or incorrect.

The performances of the logistic regression models were evaluated based on three criteria: fitness, % agreement, and relevant biological interpretation. The fitness of the logistic regression model was evaluated by using a Pearson chi-squared test (Menard, 2002) ( $p < 0.05$ ). The % agreement of the logistic regression was determined by using a confusion matrix and human-computer agreement of 0.7 as the level of acceptable agreement where 0 is no agreement



and 1 is perfect agreement (Cohen, 1960). The confusion matrix contains information about the actual classification (from human coding) and predicted classifications from the logistic regression model. For example, in the short answer question, “Define the principle: form reflects function”, a confusion matrix shows the actual (human scored) and predicted (by logistic regression model) uses of “structure” (Table 2.5). The % agreement (accuracy) is a measure of how often the classification is correct. Precision is the proportion of positive human-scored and machine-scored responses, or true positives (TP), out of all of the positive results (TP + FP). Precision is based on when the model predicts a positive result and how often the positive prediction is correct. Recall is the proportion of true positives (TP) divided by the total number of actual positive responses (TP+FN) and is a measure of the actual positives being correctly identified. Negative predictive value is the proportion of true negatives (TN) divided by the true negatives and predicted negatives that are positives (TN+FN).

#### Coding for Scientific and Non-scientific Ideas Using Machine Scoring

##### Human Scoring of Student Responses

The human scoring rubric includes key concepts, terms, and phrases, and it was built by using a grounded theory approach (Glaser and Strauss, 1967). The remaining eight questions (Table 2.1) were scored by using a conceptual rubric designed for each question (Table 2.6 and Appendix, Table A.1). For example, terms such as protection, regulation, and sensation are important in understanding the structure-function relationship in human skin. The human scoring rubric can include both structures and functions related to protection, regulation, and sensation. For the “two layers of skin” question, six conceptual categories related to protection, regulation, and sensation were represented in student responses (Table 2.6).

**Table 2.5.** Confusion matrix of student responses to “Define principle” and use of “structure”.

		Predicted	
		Present (1)	Absent (0)
Actual	Present (1)	40 (TP)	15 (FN)
	Absent (0)	6 (FP)	193 (TN)
$\% \text{ Agreement (accuracy)} = (TP+TN)/\text{total}$ $= (40+193)/254$ $= 0.917$			
$\text{Precision} = TP/(TP + FP)$ $= 40/40+6$ $= 0.869$			
$\text{Recall} = TP/(TP + FN)$ $= 40/40+15$ $= 0.727$			
$\text{Negative predictive value} = TN/(TN+FN)$ $= 193/193+15$ $= 0.928$			

TP=true positive, FP=false positive, FN=false negative, TN=true negative

**Table 2.6.** Description of conceptual rubric for each short answer question prompt.

Question prompt	Conceptual rubric	Description
Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.	Structures protection	Pigments, cells, glands and tissues that provide protection.
	Function protection	Protective barrier
	Structures regulation	Cells, glands and tissues that regulate temperature, blood supply and cell division
	Function regulation	Homeostasis, thermoregulation, repair, and regeneration
	Structures sensation	Cells and tissues which provide sensation
	Function sensation	Sense of touch and sensory perception.

Four coders scored a subset of responses and achieved an inter-rater reliability of 0.7 or higher for each concept. I calculated the intraclass correlation (Cronbach's alpha), which is used to compare agreement among more than two raters. Cronbach alpha values of 0.7 and higher are considered acceptable levels of agreement for inter-rater reliability (Cronbach, 1984). Each rater was then assigned a subset of responses to code with at least two coders assigned to each response. After this round of independent coding, I resolved any disagreements.

Each student response was scored for the presence (1) or absence (0) of each concept. Each concept was represented in student responses as key terms or phrases. For example, in the first response in Table 2.7, the student mentions the function of protection as "protective barrier" and also specifically mentions the epidermis protecting the body (function protection =1) regulation of body temperature (function regulation =1) and insulates heat (function regulation =1). In the last response in Table 2.7, the student mentions "stratified squamous epithelial tissue" (structures protection =1), which provides protection (function protection =1), secretion of oils and sweat (function regulation =1) and maintain body temperature (function regulation =1).

### Machine Scoring

The final consensus scores from the human scoring were then divided into a training data set (70%) and a testing data set (30%) for machine scoring by using an ensemble method.

Ensemble methods combine machine scoring algorithms to obtain better predictive results than could be obtained from only one machine algorithm. There are eight algorithms used in the ensemble method in this study: support vector machine, supervised latent dirichlet allocation, logitboost, classification tree, bagging classification trees, random forest, penalized generalized linear model, and maximum entropy (Table 2.8). Each algorithm is used to predict the scoring

**Table 2.7.** Human scoring of student responses to “two layers of skin” question using conceptual rubric.

Student response	Conceptual rubric					
	structures protection	function protection	structures regulation	function regulation	structures sensation	function sensation
The integumentary system acts as the protective barrier to the body. It keeps bodily fluids inside, and helps regulate body temperature. The outer layer of the skin the epidermis, protects the body from disease and outside forces. The dermis, which is under the epidermis, insulates heat.	0	1	0	1	0	0
The epidermis contains many different types of cells and provides a barrier for the body; however, the dermis provides blood vessels, connective tissue, fluids, insulation and nerves that help the epidermis. I would have to say that the dermal layer contributes to the function of the integumentary system.	0	1	1	1	1	0
I believe that the dermis contributes to the functions of the integumentary system because that is where most of the action happens, but the epidermis is just a layer that is mainly for aspect of physical appearance. The dermis is a layer of skin where sweat glands can be found and hair follicles, so majority of the functions of the integumentary system happen here.	0	0	1	0	1	0
keratinocytes provide an important function of the integumentary system by providing strength against abrasion and water resistance.	1	1	0	0	0	0

of student responses in the training dataset. Then a final prediction is obtained by taking a weighted vote of the classifier predictions (Dietterich, 2000). The weighting of the individual classifiers for the ensemble is based on the probability that the prediction is correct or incorrect, precision, and specificity.

**Table 2.8.** Machine scoring algorithms in the ensemble.

Machine scoring algorithm	Description
Support Vector Machine	Constructs a hyperplane to maximize separation of data points based on a binary classifier, works well with binary data
Supervised Latent Dirichlet Allocation	Semantic model which assesses likelihood of co-occurrence of similar words, themes are detected during human scoring
LogitBoost	Logistic regression model that is iterative with weighting, works well with binary data
Classification Tree	Iterative model for multiple predictors which sorts data based on features from root to branches, works well with binary data
Bagging Classification Trees	Builds multiple classification trees by resampling and replacement, bagging works to reduce variance
Random Forest	Builds multiple classification trees using a random subset from data, tries to reduce correlations between predictions
Penalized Generalized Linear Model	Regression model which constrains regression coefficient to reduce variance, works well with data that contains multiple predictors
Maximum Entropy	Estimates probability distribution from lexical data

Predicted scoring of the training set was then compared to the human scoring by using Cohen's kappa to quantify the agreement between the human scoring and the computer scoring.

Cohen's kappa ranges from 0.0 to 1.0 and is commonly used to quantify agreement between human and computer ratings (Landis and Koch, 1977). The levels of agreement are: values between 0.21 and 0.40 are considered "fair", between 0.41 and 0.60 are considered "moderate", between 0.61 and 0.8 are "substantial", and between 0.81 and 1.0 are "almost perfect" (Landis and Koch, 1977).

During analysis of the training set, the ensemble builds a computational model to account for the patterns detected. The same student responses and human scoring were evaluated again to determine the performance of that model by using leave-one-out cross-validation. Leave-one-out cross-validation means that the machine scoring algorithms are trained on the data minus one data point, then the algorithms are tested on that one data point that was left out (Knox, 2018). In this study, a data point is a student response; the machine scoring algorithms were trained on 90% of the data in the training set, and validation was on the remaining 10%. The leave-one-out process was repeated for all combinations: each time, a different data point (10% of the student responses) was left out and then tested. This process simulates model performance expected on new data or in a real-world application.

The ensemble-generated scoring model was then applied to a new set of human-scored student responses (testing dataset) to determine if the model performs effectively with new student responses (i.e., the training model is tested). The performance of the model with the testing data set was measured by using a confusion matrix and Cohen's kappa.

### Student Interviews

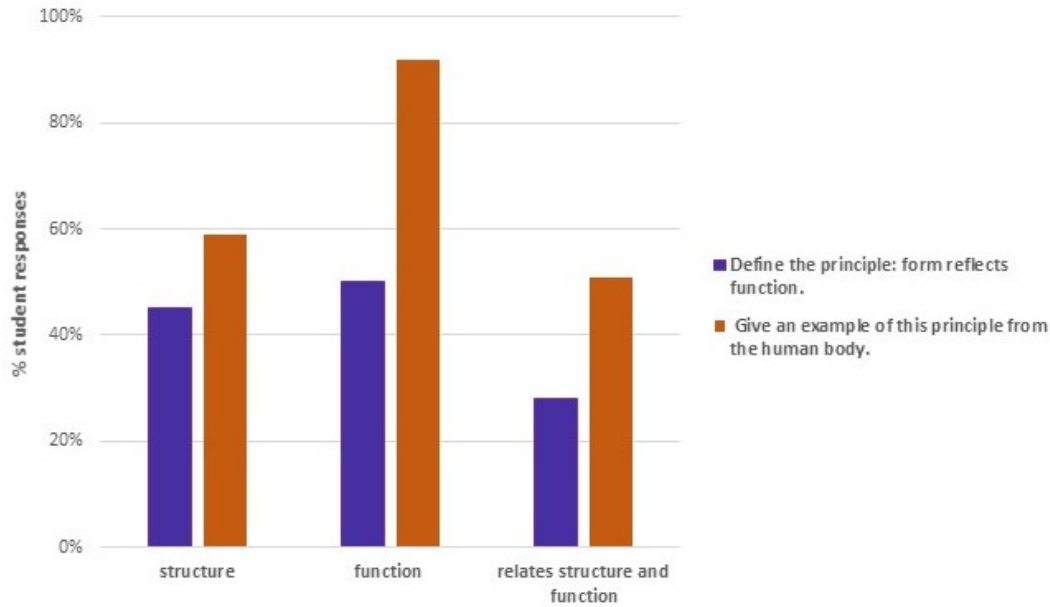
Seventeen students were interviewed to confirm that their written responses accurately reflected their thinking and to obtain feedback on the question prompts. The interviews occurred after the students had answered the questions via the course management system. Students from

both General Physiology and Human Anatomy and Physiology were invited to participate in the interviews. Each interview was approximately 90 minutes long and was audio recorded. Students were compensated for their participation. At the beginning of the interview, I used a think-aloud protocol in which each student was provided with a question prompt that he or she had previously answered and asked for a verbal response. Once he or she completed his or her verbal response, each was shown his or her written response and asked to compare the responses. Each student was also asked for feedback on the question prompts. Student responses and feedback on the question prompts were followed-up with probes to clarify any terms or explanations. The interviews were qualitatively analyzed for general themes. Details of the interview protocol are found in the Appendix.

## **Results**

### Human Scoring of Define and Give Example Questions

Human scoring of the student responses assessed the percentage of students who used structure, function, or the structure-function relationship in their responses. When asked to define the core principle of structure and function, 45% of students identified structures, 50% identified functions and 28% related structure and function. When asked to give an example of the core principle, 59% of students identified structures, 92% identified functions and 51% were able to link the two concepts (Fig. 2.1).



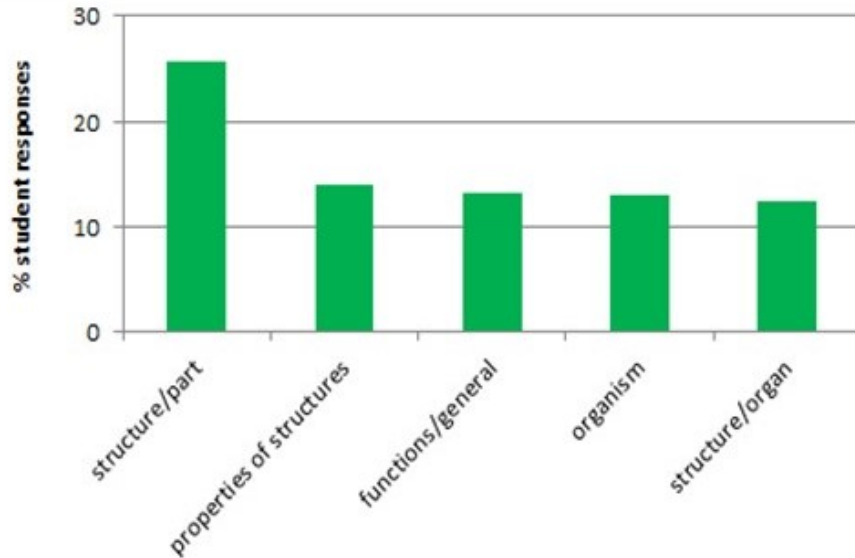
**Figure 2.1.** Human scoring of student responses to “Define” and “Give Example” questions. N=541 “Define”, N=803 “Give Example”.

### Lexical Analysis

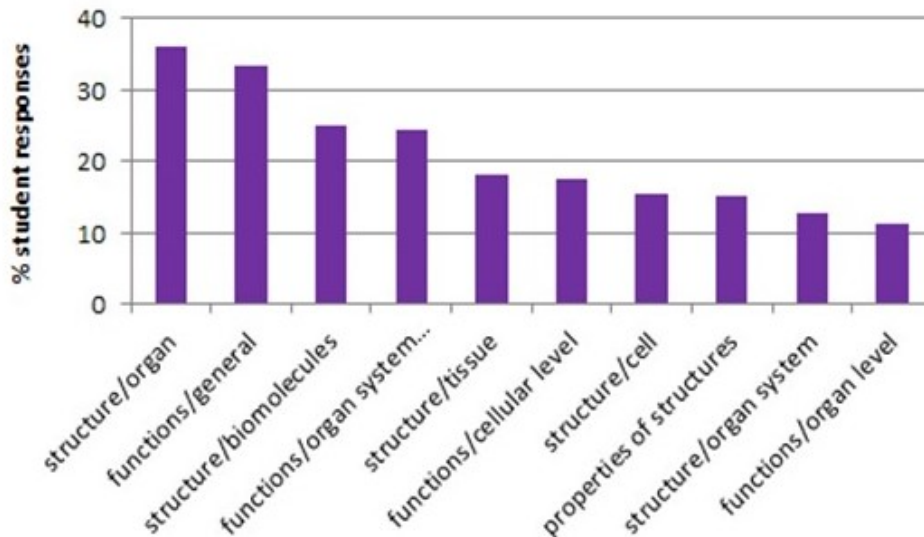
Lexical analysis using SPSS Modeler produced 22 biologically relevant categories from student responses. When students were asked to define the core principle, responses frequently included the categories of part, general functions, organism, and organ (Fig. 2.2).

When students were asked to give an example of the core principle, categories used frequently included structure/organs, general functions, structure/biomolecules, organ system functions, and structure/tissue (Fig. 2.3). Student responses were more heterogeneous in their thinking when they were asked to provide examples of the concept as demonstrated by the diversity of lexical categories.





**Figure 2.2.** Categories from lexical analysis of student responses to “Define the principle”. A total of 22 lexical categories were formed based on student responses. Only categories in more than 10% of student responses are shown.



**Figure 2.3.** Categories from lexical analysis of student responses to “Give an example of the principle”. A total of 22 lexical categories were formed based on student responses. Only categories in more than 10% of student responses are shown.

#### Model performance: Logistic regression

The categories obtained via lexical analysis were used in logistic regression models to predict human scoring of student responses. The model used a forward stepwise method to

identify the lexical categories that predicted the presence or absence of structure, function, or the relationship of structure and function within student responses. A total of six logistic regression models were built. For each of the two questions, “Define the principle: form reflects function” and “Give an example of the principle form reflects function from the human body”, I built three models: one model predicts the presence or absence of structure, one predicts function, and the third predicts structure relating to function.

In Table 9, “Define the principle-Structure” the lexical analysis category *Structure/Biomolecules* has a regression coefficient of 2.270, which means that with a one-unit increase (going from 0 to 1) in *Structure/Biomolecules*, I expect a 2.270 increase in the log-odds of Structure, while holding the other lexical analysis categories constant. The exponent of the regression coefficient ( $\beta$ ) provides the odds ratio. The odds ratios describe the likelihood of the lexical analysis category predicting the presence or absence of structure, function, or the relation of structure and function. For example, in Table 9, “Define the principle-Structure”, the lexical analysis category *Structure/Part* has an odds ratio of 191.371, meaning that the presence of *Structure/Part* in a student response leads to the response being 191 times more likely to be predicted as a one for Structure. Lexical categories with a negative regression coefficient and odds ratio less than one are less likely to contribute to the predictive model. For example, in Table 9, “Define the principle-Relates structure and function”, the lexical analysis category of *process* has a regression coefficient of -0.902. If students use “process” in their response, they are 0.41 times less likely to be predicted as a 1 in the model for relating structure and function. Logistic regression results for the question “Define” are shown in Table 2.9, and results for the question “Give Example” are shown in Table 2.10.

**Table 2.9.** Logistic regression model for the question “Define the principle: form reflects function”. N=254. (\* p<.05, \*\*p<.005).

Define the principle-Structure			
	Lexical Category	$\beta$	Odds Ratio
	<i>Dynamics</i>	2.176	8.807*
	<i>Structure/biomolecules</i>	2.270	9.675**
	<i>Structure/part</i>	5.254	191.371**
Define the principle-Function			
	<i>Functions/general</i>	1.416	4.120*
	<i>Structure/part</i>	1.003	2.726*
Define the principle-Relates structure function			
	<i>Process</i>	-0.902	0.406*
	<i>Structure/biomolecules</i>	2.095	8.125**
	<i>Structure/part</i>	3.628	37.622**

**Table 2.10.** Logistic regression model for the question “Give an example of the principle form reflects function from the human body”. N=517 (\* p<.05, \*\*p<.005)

Give an example-Structure			
	Lexical Category	$\beta$	Odds Ratio
	<i>Function</i>	-2.684	0.068**
	<i>Functions/cellular level</i>	-1.139	0.320**
	<i>Functions/organ level</i>	1.952	7.042*
	<i>Structure</i>	2.954	19.181**
	<i>Structure/cell</i>	2.145	8.543**
	<i>Structure/complex structures</i>	2.042	7.709*
	<i>Structure/organ</i>	2.846	17.227**
	<i>Structure/organ system</i>	1.910	6.754**
Give an example-Function			
	<i>Functions/cellular level</i>	1.706	5.504*
	<i>Functions/general</i>	0.970	2.639*
	<i>Mechanism</i>	0.912	2.490*
	<i>Structure</i>	-2.784	0.062**
	<i>Structure/cell</i>	1.542	4.673*
	<i>Structure/organ</i>	1.174	3.234*
Give an example-Relates structure function			
	<i>Function</i>	-3.077	0.046**
	<i>Functions/general</i>	0.996	2.707**
	<i>Mechanism</i>	0.739	2.093*
	<i>Process</i>	1.626	5.085**
	<i>Structure</i>	1.996	7.360**
	<i>Structure/cell</i>	2.615	13.667**
	<i>Structure/cell components</i>	3.416	30.440**
	<i>Structure/complex</i>	1.828	6.221*
	<i>Structure/organ</i>	3.102	22.247**
	<i>Structure/organ system</i>	1.529	4.614**
	<i>Structure/tissue</i>	1.826	6.210**

### Model performance: Accuracy of Human-Computer Agreement

Overall, there was a 91.7% agreement between human coding and logistic model predictions. The confusion matrix contains information on the agreement between the actual classification (from human coding) and predicted classifications from the logistic regression model (Table 2.5). For example, for “Define the principle-Structure”, human coding and computer prediction agreed on the presence of structure in 40 cases and the absence of structure in 193 cases. The confusion matrix also displays disagreement between human coding and computer predictions. For example, on six occasions, the computer predicted that structure was present when the response was coded as absent, i.e., false positive. Human coding of structure being absent (coded as 0) was correctly predicted by the model in 97% of cases (Table 2.5). Human coding of structure being present (coded as 1) was correctly predicted in 72.7% of cases.

The chi-square goodness of fit test was performed for each predictive model. The three logistic regression models (structure, function, and relates structure and function) for “Define the principle” demonstrate accuracy of 0.917, 0.594, and 0.894, respectively (Table 2.11). The three logistic regression models (structure, function, and relates structure and function) for “Give an example of the principle” indicate accuracy of 0.89, 0.921 and 0.876, respectively (Table 2.11). I used kappa coefficient as a measure of accuracy, which takes into consideration chance agreement. A kappa coefficient of 0.7 was used as the level of acceptable agreement (Cohen, 1960). The chi-square goodness of fit test evaluates the human coding and computer prediction for a significant model.

**Table 2.11.** Accuracy and goodness of fit of logistic regression models for “Define the principle” and “Give an example of the principle”.

A. Define the principle form reflects function (n=254)		
Predictive model	Accuracy (kappa)	Chi-Square (p)
Structure	0.917	58.504 (0.567)
Function	0.594	78.881 (0.730)
Relates structure and function	0.894	113.666 (0.000)*

B. Give an example of the principle form reflects function (n=517)		
Predictive model	Accuracy (kappa)	Chi-Square (p)
Structure	0.890	403.777 (0.000)*
Function	0.921	129.703 (0.844)
Relates structure and function	0.876	251.780 (0.029)*

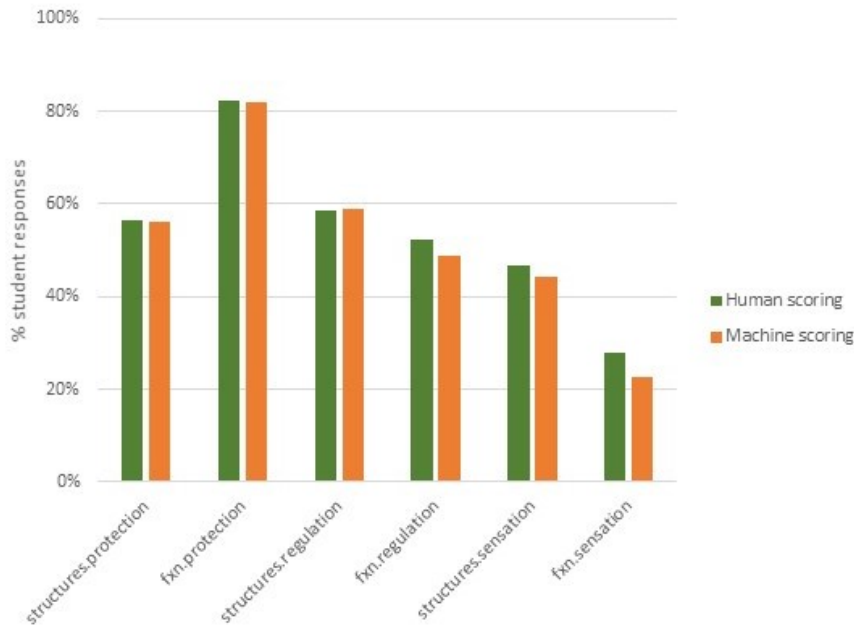
Accuracy was determined based on the confusion matrix for each logistic regression model. Chi-square goodness of fit is based on the comparison between the human coding and predictive model. (\* denotes significance  $p < 0.05$ ).

### Human Scoring of Remaining Questions

Human scoring by using a conceptual rubric revealed the percentage of students who mentioned specific concepts in their responses about the structure-function relationship. For example, in the “two layers of skin” question, 56% of students mentioned structures in the skin responsible for protection, and 82% mentioned the function of protection (Fig. 2.4).

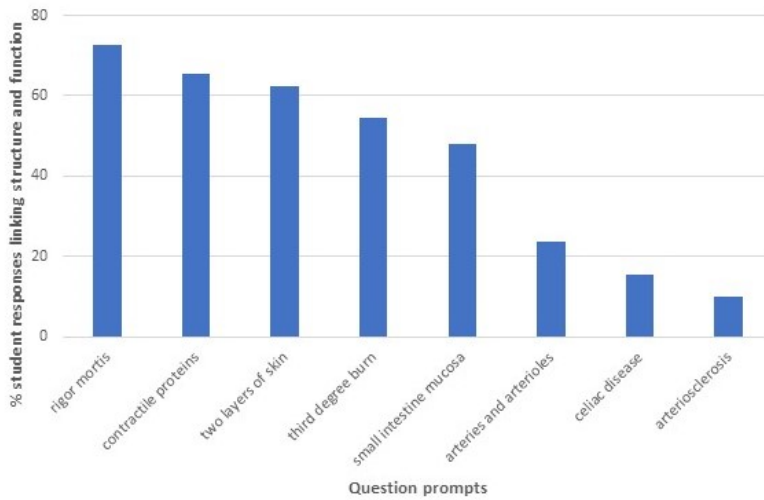
### Machine Scoring

The results of human-scoring were used for a training data set, and a testing data set for machine-scoring using an ensemble method. As shown in Figure 2.4, the human-identified frequencies of concepts (green bars) are similar to machine-scoring frequencies (orange bars).



**Figure 2.4.** Frequency of occurrence of concepts scored as “1” (present) between human scored and machine scored explanations of the structure-function relationship in the “two layers of skin” question.

The remaining constructed response questions were scored for the presence or absence of the structure-function concepts (Fig. 2.5 and Appendix, Table A.1). The “rigor mortis” question had the highest percentage (72%) of students linking structure and function in their responses, while the “arteriosclerosis” question had the lowest percentage (9.9%). For all of the questions, an average of 44% of student responses linked structure and function in their responses.



**Figure 2.5.** Frequency of occurrence of students linking structure and function in their responses for the eight constructed response questions.

#### Model Performance: Confusion Matrix

Each category in the training and testing datasets was evaluated for agreement with a confusion matrix. The confusion matrix compared the human scoring to the machine scoring (predicted scores). For example, the category *function protection* has the machine scoring agree with the human scoring that 480 of the responses included the concept. However, 11 responses demonstrated that the concept was not detected by machine scoring (Table 2.12).

**Table 2.12.** Confusion matrix of student responses to “two layers of skin” and *function protection* category.

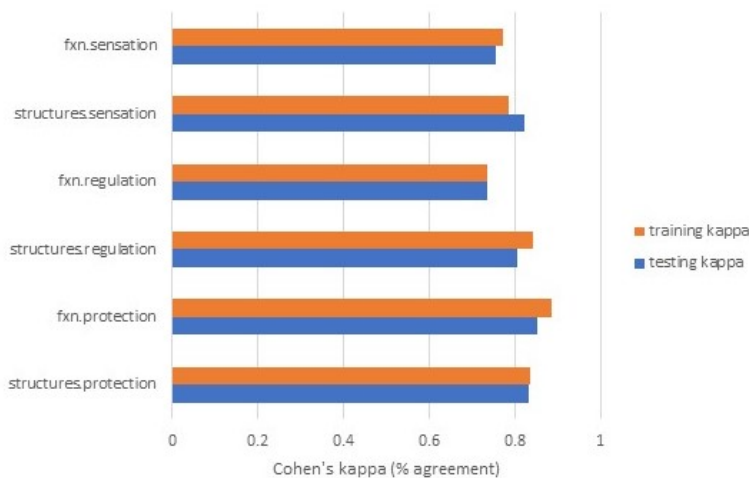
		Predicted (machine scored)	
		Present (1)	Absent (0)
Actual (human scored)	Present (1)	480	11
	Absent (0)	9	96

## Model Performance: Cohen's Kappa

The performance of machine scoring using the ensemble method was evaluated by using Cohen's kappa. Cohen's kappa ranges from 0.0 to 1.0 and is commonly used to quantify agreement between human and computer scoring (Landis and Koch, 1977). In the training dataset, Cohen's kappa for all six categories was above 0.7 (Fig. 2.6). For example, with the “two layers of skin” question, the category *function protection* demonstrated a kappa of 0.885 in the training dataset. The model generated from the training dataset was then used to build a predictive model with new human-scored responses, or the testing dataset (Fig. 2.6).

Performance of the predictive model was evaluated with a confusion matrix and Cohen's kappa. All kappa values for the testing data were above 0.7.

In this study, machine scoring detected patterns in student responses and predicted human scoring. The kappa values were similar to the human-human agreement (Appendix, Table A.2). For the “two layers of skin” question, the kappa values for all six categories were above 0.7 (Fig. 2.6). The precision and recall for the “two layers of skin” question were above 0.7 for all six



**Figure 2.6.** “Two layers of skin” categories with training kappa and testing kappa values as a measure of model performance.



categories (Appendix, Table A.2). For the “rigor mortis” question, the kappa values, precision and recall, were above 0.7 for all six categories (Appendix, Table A.2). However, the “celiac disease” question had a kappa value above 0.7 for only one category. The remaining five categories for the “celiac disease” question were below the benchmark of 0.7. For the machine scoring questions, there were a total of 49 structure and function categories. Of the 49 categories, 34 categories in the training data and 30 categories in the testing data exceeded our kappa benchmark of 0.7 (Appendix, Table A.2).

#### Model Performance: Precision and Recall

Precision is how often the model predicts a positive case correctly and is calculated from the confusion matrix. Of the 49 machine-scoring categories, 46 categories had precision above 0.7 (Appendix, Table A.2). Recall is a measure of the correctly identified positive cases. Of the 49 machine scoring categories, 36 categories had recall above 0.7 (Appendix, Table A.2).

#### Student Interviews

##### “Two layers of skin” question

Students were interviewed to determine if their responses to the short answer questions accurately reflected their thinking. At the beginning of the interview, each student was asked to provide an answer to the questions he or she had previously answered in homework. After providing a verbal answer, the student was shown his or her previously written response. Two students were interviewed about the “two layers of skin” question. Their verbal explanations closely aligned with their written explanations. For example, one student was consistent in her mention of the structures and function of protection and regulation, but not sensation:

*Student 1 written:* The epidermis has multiple layers of cells. Since stratified squamous tissue is the tissue found in the epidermis, this makes sense. This type of tissue is meant to withstand abrasion and provide protection, which is one of the primary purposes of the

skin. Since the surface layers of the epidermis are keratinized, dead cells, this allows them to flake off successfully. The dermis has adipose tissues which help the skin to insulate which is another function of the integumentary system.

*Student 1 verbal:* Let's see. Which structures of these layers contribute to the functions ... Your dermis has a lot of adipose in it if I recall. That also serves one of the functions of the integumentary system of insulating your body. Fat insulates, there we go, there's that. Protection, insulation, those are the main two things that I remember about the integumentary system. That's how ... You have oil glands also in your skin and they, your sebaceous glands, they secrete oil that help to, I guess, moisturize your skin. But they also, I feel like one of the glands might secrete something that keeps away bacteria or something along those lines and that's how it's protective.

The other student was also consistent in her explanation of structures and function of protection in both her written and verbal responses. However, in her verbal response, she also mentions “blood vessels” and “growing cells”, which are structures involved in regulation, but she does not mention the function of regulation:

*Student 2 written:* The epidermis is composed of 4-5 layers (depending on if you're talking about thick or thin skin) that all have specific structures within that contribute to its overall protective function. For instance, the stratum corneum has dead keratinized squamous epithelium cells that serve as a first line of defense against the abrasion involved in daily activities.

*Student 2 verbal:* So, I remember that epidermis is the upper layer and the dermis is underneath it. And then with the epidermis, like for your skin, you can have stratified squamous of the epithelial cells. Which help the integumentary system do its job because it's supposed to protect the body from like outside pathogens or chemicals. And so, that stratified squamous epithelial cells basically gives your body a nice layered defense against the outside world. And then the dermis just supported the epidermis with blood vessels and growing cells, making sure that the epidermis has all the layers of the stratified squamous epithelial cells.

Both students were asked for feedback on the question prompt, and if any part of the question prompt was either helpful or confusing in answering the question:

*Student 1:* I think that this was actually the first question was a really great introduction into this experiment, if you will. Because you didn't just say, "Consider the integumentary system, how does form meet function?" That would have been like, wow,

that's a lot to think about because you can also think of nails, you can think of hair. But you said, "Consider the two layer of skin," you named the two layers of skin and then you said, "What structures of these layers contribute to the function?" And I think that's about as good as you could have worded that, and explain your reasoning behind it. You're not just saying, "This form meets this function," you're also telling them to explain as much as possible.

*Student 2:* I believe that number two was well put. It definitely is straight forward, it is just asking to recall some information that at the time might seem basic for the dermis, epidermis. It should be pretty fundamental if you are writing about the integumentary system. So, I think this was a good question. Because it was straight forward. And wasn't any tricks to it. It was just kinda like remember. So, it was a good question.

I used the same protocol to interview 17 students about the remaining questions to determine if their responses to the short answer questions accurately reflected their thinking. Students' verbal responses to the question prompts were similar to their original written response. The majority of the students (16 out of 17) stated that their written response was much more detailed than their verbal response when asked to compare the two.

Students found the wording of the question prompts helpful for eliciting their responses. No students found the wording to be confusing. Many of the students stated that the question prompts provided details that were helpful in answering the question. For example, these students provided feedback on the question prompt, "The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction", and the students claimed that the use of the "sliding filament model" in the prompt was helpful:

*Student 3:* "I think describing ... Putting in there about the sliding filament model you know that in some fashion they have to move along the filament then, and I think that helps you visualize what's happening, and again, it could help you even if you don't know, make an educated guess, or have an idea of what's going on."

*Student 4:* "For this one, I guess the actin and myosin. I think the "sliding filament model of the muscular contraction" really helped because at first I forgot what actin myosin were, but then I remembered sliding filament. Then I pictured a video I saw in class about the two sliding over ... One coming to connect. So that helped me remember."

Other students noted that the scenario in the question prompt was helpful in answering the question because it provided a visualization. For example, these students answered the question, “A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.”

*Student 5:* “Well, I like how before they tell you that the patient is in rigor mortis that they kind give you the symptoms and in your mind your already starting to think about maybe what is this? ... Where it says the victim is clutching a syringe in one hand and that gives you a good visualization of what's going on. I like that because you can kinda get a picture in your mind, if you maybe learn visually and stuff and that gives you an idea of... It gives you a picture that you can work off of in your mind.”

*Student 6:* “I actually like how you gave a scene, so a person could relate it to actual life. I don't think there was any part of it, actually, that was confusing or unnecessary. It helps to give an image. It helps a person visualize the scene. Saying that the victim is clutching the syringe, so you know that it's unable to actually be removed. Then just stating 'actin and myosin' itself, so you know that 'Oh, we're talking about actin and myosin and the role with muscle contraction and relaxation.' It helps me visualize what's going on.”

## Discussion

This study demonstrated that lexical analysis and machine scoring can be used to identify student ideas about the structure-function relationship as has been shown for questions in evolution (Ha et al., 2011), photosynthesis (Weston et al., 2015), and genetics (Prevost et al., 2016). In this study, lexical analysis and machine scoring were used to build predictive models that can analyze student thinking about the structure-function relationship in anatomy and physiology. Therefore, this study was designed 1) to build and test the efficacy of computer-automated scoring tools to predict human scoring and 2) to examine student understanding of the core concept of structure-function.

Research Question 1:

*How can computer-automated scoring methods, such as lexical analysis and machine scoring, be used to build predictive models that mimic human scoring of structure-function formative assessment questions?*

### Model Performance

The categories obtained via lexical analysis were used in logistic regression models to predict human scoring of student responses. A total of six logistic regression models were built: three models for “Define the principle: form reflects function” and three models for “Give an example of the principle form reflects function from the human body”. The three logistic regression models (structure, function, and relates structure and function) for “Define the principle” demonstrate accuracy of 0.917, 0.594, and 0.894, respectively. The three logistic regression models (structure, function, and relates structure and function) for “Give an example of the principle” indicate accuracy of 0.89, 0.921 and 0.876, respectively (Table 2.11).

Human scoring was used to inform the computer-automated scoring models. To build predictive models, the supervised computer-scoring models learn the scoring rules from the human scoring. This process is similar to the scoring of evolution questions to detect patterns associated with the presence or absence of concepts (Ha et al., 2011). In this study, it was important to incorporate human-scoring into the computer-scoring because human-scoring could recognize multiple ways to describe functions, and student responses to “Define the principle” were varied in their terminology to describe functions. Human scorers could easily comprehend meaning, interpret lexical expressions, and recognize the equivalence in these definitions. The human scoring of these definitions was then used to train the computer scoring model:

This principle means that the shape of the body part reflects what that body part does or is used for. (Define)

The shape of a particular cell, organ system or any other living structure can give way to its purpose and how it serves the body. (Define)

This means that body structures take a particular form or shape which helps them best perform their function (Define)

For the “Define” question, the predictive model for *function* demonstrated a kappa value of 0.594 (Table 2.11). This low kappa coefficient could be due to the varied linguistic expressions students use to discuss function as shown in the above student responses. Human scoring of these “Define” responses interpreted the expressions to describe functions. Lexical analysis would detect all of these words within student responses, but human scoring would form the categories that contain similar terms and synonyms. For example, “does”, “used for”, “purpose”, and “serves” indicate function. Human scorers were able to interpret this meaning, and this information was used to form categories and train the lexical analysis models. The predictive models are therefore more robust because they were trained with human scoring data.

In this study, machine scoring successfully detected patterns in student responses and predicted human scoring. For the “two layers of skin” question, the kappa values for all six categories were above 0.7 (Fig. 2.6). For the remaining seven questions, there were a total of 43 categories with 34 categories that exceeded our kappa benchmark of 0.7 in the training data and 30 categories in the testing data (Appendix, Table A.2). Two questions, “two layers of skin” and “rigor mortis”, had high kappa values ( $>0.7$ ), while “small intestine mucosa” and “celiac disease” had categories with low kappa values ( $<0.7$ ). These results may be explained by examining the limitations of model performance.

## Limitations of Model Performance

The machine scoring algorithms were generally highly effective (training: 23/43 > 0.8 kappa, 11/43 > 0.7 kappa, 7/43 > 0.5 kappa; testing: 23/43 > 0.8 kappa, 7/43 > 0.7 kappa, 9/43 > 0.5 kappa, Appendix, Table A.2) at scoring responses to the structure-function short answer responses. However, a few limitations were revealed and are summarized in Table 2.13. The factors that limited the effectiveness of machine scoring included: uncommon concept frequencies, the diversity of expressions students used to represent concepts, and misspellings.

The first factor that may limit the effectiveness of machine scoring is uncommon concept frequencies. Prior research in machine scoring has demonstrated the frequency of occurrence of specific concepts is as important as overall sample size (Dumais, et al., 1998; Ha et al., 2011). Machine scoring is based on the frequency of occurrence of cases, or the positive instances, and this value is incorporated into the accuracy of machine scoring algorithms (Dumais, et al., 1998). A larger number of cases will have a greater effect on performance. In another study, the concept “competition” was rarely used by students to explain evolutionary change and therefore the machine scoring algorithms did not have sufficient positive cases of student responses containing the concept to build a predictive model (Ha et al., 2011). In this study, example (1) in Table 2.13 demonstrates a disagreement, a positive machine score for structures protection and a negative human score, due to the presence of an uncommon term. In this situation, the term “stratum basale” describes the bottom layer of the epidermis, which is important for regeneration. The term was used in 19 student responses (<5%) in the training data set. However, the other strata of the epidermis are important in protection, and many student responses (>25%) mention the other strata. The machine scoring algorithm recognized the term “stratum” in the response but was unable to differentiate between the various types of strata, so stratum basale

was uncommon. A potential solution to this limitation is to increase the frequency of occurrence of this term, possibly by using duplicate responses in the training dataset. The duplicate responses may produce different results with the predictive model because the repetition yields more influence on the resulting model (Witten & Frank, 2005).

A second limitation to the effectiveness of machine scoring was the diversity of expressions students used to denote concepts. The student response in Example (2) of Table 2.13 includes the phrase “disposable sloughing” to refer to the regeneration of the skin. Other student responses included “cell division”, “making more skin cells”, “pushing them upward”, and “sloughed off”. These phrases were detected by human scorers as representing regeneration but were not detected by machine scoring. If the total number of student responses is increased in the training dataset, it is possible to identify other students who use these types of expression. If a phrase occurs in more responses, there is a greater likelihood that it will be detected by machine scoring.

The third type of limitation to the effectiveness of machine scoring is misspelling. Ha and Nehm (2016) found that misspelled words do not have an impact on machine scoring of evolution responses. However, they point out that the effect of misspelled words depends on which words are misspelled, and whether the misspelled word is a key evolution concept (Ha and Nehm, 2016). In this study, misspelled words are shown in Example (3) in Table 2.13. The student refers to a “dendrite cell” in his or her response. Since a dendrite is a nerve cell process, the student may have been referring to a dendritic cell but misspelled the word. Human scorers recognized the word and assumed that the meaning that the student implied was a dendritic cell. Machine scoring did not recognize the term “dendrite cell” as a structure involved in protection. To minimize misspellings, spell-check software could be added during the homework data



collection or in pre-processing. A potential problem with this solution is that the dictionaries used by spell-check programs often lack discipline-specific words (Ha & Nehm, 2016). Thus, correctly spelled words from an anatomy and physiology course might be incorrectly labeled as misspelled. For example, misspelled words in student responses to this question included “Merkle” [Merkel], “kerhatinnized” [keratinized], and “squamus” [squamous], which led to misclassifications. Another potential solution to misspelled words suggested by Ha and Nehm (2016) is to identify commonly misspelled words and include them in the training data for the machine-scoring model. In general, misspelled words occurred with such low frequency that they did not have a meaningful impact on their computer scoring models (Ha & Nehm, 2016). In this study, misspelled words occurred with low frequency (1-2%) and do not appear to affect the predictive models. However, as computer-automated scoring usage continues to increase, further studies with other student populations and in other disciplines are warranted.

Another potential limitation in this study is question administration: Low-stakes formative assessment in this format can help to increase student confidence through feedback and to allow students to explore their ideas. However, the effort students direct towards the assessment task is related to how important they perceive the task to be (Wise & DeMars, 2005). If students do not perceive the value in a formative assessment task, it may affect their effort, yet effort is difficult to measure. However, this approach has been used successfully to investigate student understanding in many contexts (Carter & Prevost, 2018; Haudek et al., 2012; Prevost et al., 2016). In this study, all questions were administered online as low-stakes homework outside of class with no time limit. Students were awarded a small number of points for completion rather than correctness and were encouraged to give their best effort.

**Table 2.13.** Examples of types of disagreements between human-scored and machine-scored explanations.

Limitation	Scoring disagreement	Examples	Solutions to correct disagreement
Low frequency of a concept OR a concept is common	Negative human score but positive machine score for <i>structure protection</i>	(1) The basal stratum is one of the layers which contributes to the functions of the integumentary system because it is the only layer capable of cell division which pushes up cells and helps them replenish the outer layer which is constantly shedding dead cells.	Increase frequency of occurrence of term in training samples, possibly through the use of duplicate responses
Diversity of expressions used to represent concept	Positive human score for <i>function regulation</i> but negative machine score	(2) Its not so much one layer or the other that contributes more, its more of a combination of both. The epidermis provides a kind of shielding and disposable sloughing that allows for protection, whereas the dermis provides most of the nervous functions as well as vascularity. Neither of these would function correctly without the roles of the other.	Increase number of student responses in training data so rare expressions may become more common
Misspellings	Positive human score for <i>structure protection</i> and <i>function sensation</i> but negative machine score	(3) Our skin protects us from anything harmful that exists everywhere we go. As soon as I think of this, I think how it must fight off intruders and protect us. A particular cell in our skin layers helps us with this, a dendrite cell. Also, if anything was to happen, we need to feel this happening.	Use of spell check software during homework data collection, increase number of student responses so rare expression may become more common

Furthermore, the computer-assisted scoring programs used in this study are time-intensive. Lexical analysis and machine scoring use different approaches for automated scoring. Lexical analysis extracts words and phrases prior to building a predictive model. Machine scoring detects patterns from human-scored responses while building a predictive model. Both tools require human scoring of responses to build predictive models, although lexical analysis may be used on unscored responses to explore student word choices. All student responses were human scored, which involved time for training expert graders, time to achieve interrater reliability, and time to score the responses. For the lexical analysis model, time was also spent building the lexical library of anatomy and physiology terms.

Additionally, the cost associated with the lexical analysis software IBM SPSS Modeler is substantial, which makes it somewhat cost-prohibitive. However, the machine scoring models are housed on the Automated Analysis of Constructed Response (AACR) server and are freely accessible. Instructors interested in using these questions, or other biology related questions, may visit the AACR research group website at <https://create4stem.msu.edu/project/aacr>.

Research Question 2:

*What do the predictive models built from computer-automated scoring demonstrate about student conceptual understanding of the structure-function relationship in physiology?*

### Conceptual Understanding of the Structure-Function Relationship

Conceptual learning serves as a foundation for understanding physiology and provides students with a tool to connect fragments of factual information. Traditional classroom learning in anatomy and physiology involves the rote memorization of facts with minimal time spent on conceptual understanding (Michael, 2007). Rote memorization does not exemplify understanding, but a lack of conceptual understanding (Pines & West, 1986). With rote

memorization, the student attempts to mentally organize the factual knowledge without an existing framework, and the memorized facts are fragments of knowledge. With conceptual understanding, a student learns a concept, such as the structure-function relationship, and then through instruction learns examples of how this concept may be applied. The concept then becomes a framework for a student to mentally organize information. One way in which students can demonstrate their conceptual understanding is to apply this understanding to a new context. In this study, we asked students to apply their knowledge of the structure-function relationship to the integumentary, muscular, digestive, and cardiovascular systems.

My results suggest that students have difficulty in applying the structure-function relationship for these constructed response questions. For the question “Define the principle: form reflects function”, only 28% of students related structure and function, while for the question “Give an example of the principle: form reflects function”, 51% were able to link structure and function. A similar result with students having difficulty relating structure to function was observed in a smaller study in which the order of the two questions was manipulated (Carter and Prevost, 2018). The structure-function relationship is explicitly taught in the Anatomy and Physiology course, but it is implied in the General Physiology course. Students learn the structure-function core concept at various points in the curriculum. In interviews, students were asked if they were familiar with the structure-function relationship. Although student responses confirmed that they were introduced to and were familiar with the structure-function relationship, students had difficulty applying this concept to both the definition and example questions (Fig. 2.1).

Yeah, we actually, gosh, have heard that [structure-function relationship]so many times. I've heard it in anatomy, we heard it ... I don't even know why we were talking about it in biochem, but we were talking about it in biochem and medical botany, and bio one and two, and I feel like I've heard that phrase a lot. (Student 7)

I've heard of it [structure-function relationship] before. I can't pinpoint where it was from. I feel like more Bio II was the way things look, like the finches' beaks reflects what they would eat and stuff. That kind of stuff, everything looks a certain way for a reason. (Student 8)

On average, 44% of students were able to relate structure-function in the eight questions in which students were asked to apply the concept to a specific physiological context. While responding to these questions, the students may have been reproducing facts they had memorized about each particular context rather than demonstrating conceptual understanding. Rote memorization may be exemplified by student responses to multiple questions during interviews; e.g., the following interview quotes come from student 10. In response to the prompt “Define the principle: form reflects function”, the student reiterates the question prompt, then provides an example without providing a definition of the core concept:

To me the principle form follows function means that you can't have one without the other. Function always follows structure or form and they are inseparable. Also what a structure can do or perform depends on the form its in. For example triceps can't perform what a biceps can and vice versa because they are two different forms that perform two different functions. (response to “Define” prompt)

When asked to “give an example”, the student again repeats the words “form” and “function”. The student does provide the functions “extend and rotate” associated with the muscle. However, the student attempts to further explain the description of the muscle and functions of the muscle as “big”, which is uninformative:

The form of your gluteus maximus is being the largest gluteal muscle because it performs the greater amount of functions. The gluteus maximus acts to extend and laterally rotate the hip joint and is a very powerful extensor. This is an example of form reflects function because a big muscle reflects big functions in a sense. (response to “Give example” prompt)

In this response to the “third degree burn” question prompt, the student broadly mentions the functions of protection, regulation, and sensation but only links structure and function with regards to sensation:

The epidermis and dermis are the outermost and second most superficial layers of the skin. When these are damaged the functions of touch is gone, the separation of your body from the outside world is gone, which makes you susceptible to infections and other harmful bacteria; the dermis regulates your body's temperature and with third degree burns this regulation is impaired and causes problems. Your sweat glands have been damaged, all of the skin's blood vessels and nerves, including sensory nerve endings that respond to touch, pressure, heat, cold, and pain have been damaged and your body is in a state of trouble and panic. (response to “Third degree burn” prompt)

The student’s response to the “rigor mortis” question includes the use of the terms “thick” and “thin” without further explanation of the contractile protein structures. The student recognizes that there is a lack of oxygen and ATP during rigor mortis but fails to identify ATP as necessary for the contractile proteins to detach:

The victim died clutching the syringe which is the stretching of a muscle which pulls the thick and thin filaments together. In order for the muscles to contract your body needs oxygen. Since the victim died and there wasn't any oxygen present to make ATP to help contract these muscles, the coroner couldn't pry open the victims hand releasing the syringe. (response to “Rigor mortis” prompt)

Movement often requires the contraction of a skeletal muscle. The sliding filament model describes the process used by muscles to contract. It is a cycle of repetitive events that causes actin and myosin myofilaments to slide over each other, contracting the sarcomere and generating tension in the muscle. (response to “Contractile proteins” prompt)

Overall, these interview responses demonstrate that as the student attempts to answer each question, the student struggles with the definition and merely reiterates the terms form and function. When the student provides an example, structures and functions are presented separately but are not linked explicitly. The student provides pieces of factual knowledge of structures and functions but is unsuccessful in connecting structures to functions. These

responses suggest that the student is repeating facts from what was learned in class about these topics rather than approaching these questions conceptually.

### Students Lack a Conceptual Framework for Structure and Function

Conceptual understanding necessitates a conceptual framework. A core concept is composed of multiple ideas that form a conceptual framework (McFarland et al., 2016; Michael et al., 2017). For example, the components of the structure-function relationship core concept are knowledge of structures and functions. Students must understand both of these terms to fully comprehend the link between the two. During the interviews, students were asked to define the individual terms, “structure” and “function”. Most students were able to define function but had difficulty with structure:

Wow. That's a good question. Structure is what something is, whether that be how it's shaped or what it looks like. I guess I could say what something looks like would be a better definition because when you get into what something is, that could get all philosophical I guess. (Student 9)

The same student responded thus when asked to define function:

What something does, that was the easiest. (Student 9)

Students found it challenging to define structures, which might explain why students had difficulty with defining the structure-function relationship. Although students mentioned learning about the structure-function relationship in their prior classes, they had difficulty using their knowledge to provide a definition as well as to apply it to the example. Such difficulty may be due to a lack of a conceptual framework. Although conceptual frameworks are useful in designing concept inventories, they can also provide a scaffold for student learning about the components that underlie a core concept (Michael et al., 2017).

## Levels of Organization

Lexical analysis, machine scoring, and student interviews indicate that students have difficulty with certain levels of organization, which is a core concept in physiology (Michael et al., 2009). In physiology education, all levels of organization -- molecules, cells, tissues, organs, and organ systems -- are included in the curriculum. It is important for students to be able to recognize that physiological processes occur at multiple levels of organization simultaneously (Lira & Gardner, 2017).

Lexical analysis categories revealed that students frequently referred to only a few levels of organization. The lexical analysis software used in this study, IBM SPSS Modeler, builds text categories which can contain multiple terms and synonyms. The categories can be honed by a subject matter expert, and in this case, the categories were designed to reflect the biological levels of organization. Students drew on only a few levels of organization, primarily at the organ and organ system levels (Figs. 2.2 and 2.3).

Eleven out of seventeen students who were interviewed identified organs as the level of organization that they draw from, three students mentioned cells, and three students described organ systems. In interviews, students explained that the macroscopic nature of organs makes organs much more tangible. This perspective is exemplified in these student responses:

I think I probably think about things I can touch. So I think about the bones because I can feel the features of my bones, if I touch my arm. I guess, yeah because like I said if I'm going to think about something really small, I feel like I need to see it in a microscope. But even then I can't touch it. And so I feel like if it's something that's big enough that I can feel or I can easily visualize the features, that's a lot easier for me to think about. It feels more natural. (Student 11)

I would say organs almost immediately. Just because that's very much ... The human body, and I guess especially limited exposure people have to medicine, just in their daily lives it's, "Oh, something was wrong with this particular organ." "Oh, something was wrong with this." And so almost always people this of ... Because I think of just organs in general. (Student 12)



Machine scoring also shows that students have difficulty with the questions that required students to include “properties of structures” in their response. Properties of structures are words that describe a structure, such as “flat”, “long”, or “elastic”. Students linked structure and function less frequently with the question prompts “blood pressure” (23.8%), “celiac disease”, (15.3%) and “arteriosclerosis” (9.9%).

Less than half of the student responses in this study demonstrated conceptual understanding of the structure-function relationship. A possible reason for performance on these structure-function questions may be inherent to the student populations from which data was collected. Differences in students’ academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance on these questions. The short answer questions were administered to students in a junior-level General Physiology course and a sophomore-level Human Anatomy and Physiology course at a large public research university (moderately selective), and to students in Human Anatomy and Physiology at two Southeastern two-year colleges. Both two-year institutions in this study are open-access, meaning that students may attend without any academic qualifications. This situation will be explored further in chapter 3 of this dissertation.

Another potential reason for student demonstration of conceptual understanding is the cognitive level of the structure-function questions. The cognitive level of the question prompts may have an effect on students demonstrating conceptual understanding. The short answer questions are from the first three levels of Bloom’s taxonomy: remember, understand, and apply (Anderson et al., 2001). “Remember” refers to retrieving information from long-term memory and is typically associated with recall or recognition tasks. “Understand” goes beyond simply remembering material, and “apply” refers to using the information in a different context

(Anderson et al., 2001). The question prompts from each of these levels ask for different types of conceptual understanding from the students, and this requirement may have an effect on performance. This possibility will be explored further in chapter 4 of this dissertation.

### Implications for Teaching

Lexical analysis and interviews showed that students draw on only a few levels of organization. Instructors should discuss examples of the structure-function relationship within a variety of levels of organization to help students apply the concept and reason across multiple levels of organization. Students enrolled in Human Anatomy and Physiology and General Physiology courses primarily intend to work in healthcare, where thorough knowledge of the human body is necessary. Therefore, students need to recognize the structure-function relationship from molecular to the organismal level. One way that instructors may address students' difficulty with molecular and cellular levels is to incorporate more examples at these levels to enhance student familiarity. My results also suggest that formative assessment tasks need to be designed to reflect multiple levels of organization and the structure-function relationship: e.g., discussing examples of the molecular structure of proteins and how such structure influences their function, or the shape of nerve cells and how the shape enables the function of communication. If formative assessment is solely targeted at the organ level, it may not identify student conceptions, or misconceptions, at the cellular level.

### Conclusion

My research demonstrates that automated scoring can successfully evaluate a large number of student responses in Human Anatomy and Physiology and General Physiology courses. Automated scoring alleviates some of the barriers to the use of constructed response questions as formative assessment, which is important for revealing student conceptual

understanding and their heterogeneous ideas. Prior work in conceptual understanding in physiology education has encouraged the use of multiple-choice assessments (Michael et al., 2009). Multiple-choice questions allow for guessing and response elimination strategies, while constructed-response questions require students to use their own knowledge to construct their responses rather than choose from a list of options like in multiple choice questions (Kuechler & Simkin, 2010; Martinez, 1991). Automated scoring of written assessment provides an avenue with which to focus on student understanding of the core concepts in undergraduate physiology education.

## References

- American Association for the Advancement of Science (2011). Vision and change in undergraduate biology education: A call to action.
- Abu-Mostafa, Y.S. (2012). Machines that think for themselves. *Scientific American*, 307(1), 78-81
- Anderson L.W., Krathwohl D.R., Bloom B.S., Bloom B.S. (2001). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman.
- Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater[R] V.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1-31.
- Bell, B. & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85, 536-553.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. Annual Meeting-Association for Computational Linguistics, 206-210.
- Carter, K.P. & Prevost, L.B. (2018) Question order and student understanding of structure and function. *Advances in Physiology Education*, 42(4),576-585.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1, 37-46.
- Cronbach, L.J. (1984). A research worker's treasure chest. *Multivariate Behavioral Research*, 19, 223-240.

Dietterich, T.G. (2000). Ensemble methods in machine learning. *In Multiple Classifier Systems Ensemble Methods in Machine Learning*, Springer.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management* (pp. 148-155).

Espinoza, F., Hamfors, O., Karlgren, J., Olsson, F., Persson, P. et al. (2018) Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction In: *Proceedings of Conference on Human Information Interaction & Retrieval*

Foltz, P.W., Laham, D. & Landauer, T.K. (1999). The Intelligent Essay Assessor: Application to educational technology. *Interactive Multimedia Electronic Journal of Computer Enhanced Learning* 1-6.

Foltz, P.W., Streeter, L.A., Lochbaum, K.E., & Landauer, T.K. (2013). Implementation and applications of the Intelligent Essay Assessor. *In Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Eds Shermis, M.D. and Burstein, J. Routledge, UK.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory*. Aldine Publishing Company, Hawthorne, NY.

Ha, M. Nehm, R. Urban-Lurain, M & Merrill, J.E. (2011). Applying computerized scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE-Life Sciences Education*, 10, 379-393.

Ha, M., & Nehm, R. H. (2016). The Impact of Misspelled Words on Automated Computer Scoring: A Case Study of Scientific Explanations. *Journal of Science Education and Technology*, 25, 358-374.

Haudek, K.C., Prevost, L. B., Moscarella, R.A., Merrill, J. & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE-Life Sciences Education*, 11, 283-293.

Haudek, K. C., Kaplan, J. J., Knight, J., Long, T., Merrill, J., Munn, A., Nehm, R., Smith, M. & Urban-Lurain, M. (2011). Harnessing Technology to Improve Formative Assessment of Student Conceptions in STEM: Forging a National Network. *CBE - Life Sciences Education* 10(2), 149-155.

Knox, S. W. (2018). *Machine learning: a concise introduction*. Hoboken, New Jersey: John Wiley & Sons, 2018.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence*, 18, 411-426.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3), 249.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37, 331-344.

Kuechler, W.L. & Simkin, M.G. (2010). Why is performance on multiple choice tests and constructed response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, 8, 55-73.

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

Lira, M.E. & Gardner, S.M. (2017). Structure-function relations in physiology education: Where's the mechanism? *Advances in Physiology Education*, 41, 270-278.

Livne, N.L., Livne, O.E. & Wight, C.A. (2007). Can automated scoring surpass hand grading of students' constructed responses and errors in mathematics? *Journal of Online Teaching and Learning*, 3(3), 295-306.

Martinez, M. (1991). A comparison of multiple-choice and constructed response figural items. *Journal of Educational Measurement*, 28(2), 131-145.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.

McFarland, J., Wenderoth, M. P., Michael, J., Cliff, W., Wright, A., & Modell, H. (2016). A conceptual framework for homeostasis: development and validation. *Advances in Physiology Education*, 40(2), 213-222.

Menard, S. (2002) Applied logistic regression. Thousand Oaks: Sage

Michael, J. (2007). What makes physiology hard for students to learn? Results of a faculty survey. *Advances in Physiology Education*, 31, 34-40.

Michael, J., Modell, H. McFarland, J. & Cliff, W. 2009 The "core principles" of physiology: what should students understand? *Advances in Physiology Education*, 33, 10-16.

Michael, J., Martinkova, P., McFarland, J., Wright, A., Cliff, W., Modell, H., and M. P. Wenderoth. (2017). Validating a conceptual framework for the core concept of "cell-cell communication". *Advances in Physiology Education*, 41(2), 260-265.

Nehm, R., Ha, M., & Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology*, 21(1), 183-196.

Nehm, R.H. & Haertig, H. (2012) Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, 21, 56-73.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127-42.

Pines, A. L., & West, L. H. T. (1986). Conceptual understanding and science learning: an interpretation of research within a sources-of-knowledge framework. *Science Education*, 70, 583–604.

Prevost, L.B., Knight, J.K., & Smith, M.K. (2016). Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE-Life Sciences Education*, 15(4), ar65.

SPSS IBM Modeler (2013). IBM SPSS Modeler v15, IBM corporation.

Weston, J., Haudek, K.C., Prevost, L., Urban-Lurain, M. & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed response questions on photosynthesis. *CBE-Life Sciences Education*, 14, 1-12.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, 10(1), 1–17.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Amsterdam; Boston, MA: Morgan Kaufman.

Zhang, Y., Jin, R., & Zhou, Z. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine learning and Cybernetics*, 1, 43-52.

## CHAPTER 3 COMPARISON OF TWO-YEAR AND FOUR-YEAR STUDENTS

### **Abstract**

Anatomy and physiology is taught at a variety of institutions, including 2-year community colleges and 4-year research universities. Regardless of the type of institution offering anatomy and physiology, conceptual understanding of the structure-function relationship is necessary to understand physiological processes. The focus of my research was to compare conceptual understanding of 2-year versus 4-year anatomy and physiology students by using written formative assessment. I hypothesize that differences in students' academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance. Based on prior research, I predict that there will be a difference in conceptual understanding of the core concept structure and function between two-year and four-year students in anatomy and physiology, and that the students at the two-year institution will not perform as well as the students at the four-year institution, as measured by performance on the constructed response questions. Responses to eight short answer essay questions were collected at both types of institutions from 890 students in human anatomy and physiology over six semesters. My results demonstrated that there is a difference in conceptual understanding of the structure-function relationship between 2-year and 4-year students in anatomy and physiology with more 4-year students mentioning SRF concepts in their responses compared to the 2-year students. A potential reason for this difference may be college readiness. There was no difference in performance between institution types on structure-function concepts examined in the A&P II

course. My results suggested that students may benefit from a focus on core concepts within the content of anatomy and physiology courses. This focus should occur in both the first and second semesters of anatomy and physiology. Instructors can use written formative assessment to allow students to demonstrate their conceptual understanding within the organ systems.

## **Introduction**

Anatomy and physiology is taught at a variety of institutions, including 2-year community colleges and 4-year research universities. Students who take anatomy and physiology at 2-year community colleges are pursuing a variety of health programs, such as nursing, physical therapy, or radiologic technician, and some students transfer to 4-year institutions to complete a bachelor's degree. Students who take anatomy and physiology at 4-year institutions are pursuing a bachelor's degree and are interested in careers such as nursing, physical therapy, physician's assistant, or medical doctor. Health programs are currently a focus of higher education since there is a shortage of nurses and other allied health workers, yet there are few studies about anatomy and physiology courses (MacDowell et al., 2009; American Association of Colleges of Nursing, 2014; Forgey, 2016).

Despite the need for anatomy and physiology courses, there is high attrition at 2-year institutions, which leads to a lack of college readiness within nursing and allied health programs, or an inability to transfer to 4-year institutions. The success rate of students in anatomy and physiology courses at 2-year institutions [defined as 70% (C) or better] is typically near 50% (Hopper, 2011; Forgey, 2016). The high attrition rate may be due to the conceptual difficulty of the course content (Davis, 2010), the amount of terminology used in anatomy and physiology courses (Sturges & Maurer, 2013), or due to student requirements to synthesize information across scientific disciplines (Feder, 2005).



Many students take anatomy and physiology at 2-year institutions, but studies in the literature related to academic preparedness at community colleges for nursing, allied health programs, or to transfer to 4-year institutions are few. Performance in anatomy and physiology courses at community colleges appears to predict success in nursing and allied health programs (Newton et al., 2007). Melguizo & Dowd (2009) found that students who transfer from a community college to a university tend to be less academically prepared.

Since anatomy and physiology is a gatekeeper course for a multitude of health programs, the focus should be on increasing student success in the course (Forgey, 2016). One avenue to increase student success is the development of conceptual understanding. Often, students resort to rote memorization of facts rather than conceptual understanding (Michael et al., 2007). There has been a movement within the biology and anatomy and physiology communities towards conceptual understanding of the core concepts (Michael et al., 2009). The core concepts serve as a foundational learning tool for students; one core concept in anatomy and physiology is the structure-function relationship.

Regardless of the type of institution offering anatomy and physiology, conceptual understanding of the structure-function relationship is necessary to understand anatomical and physiological processes. However, the type of institution (2-year or 4-year) may influence conceptual understanding. The focus of my research is to compare conceptual understanding of structure-function in 2-year versus 4-year anatomy and physiology students by using written formative assessment and constructed response questions with responses collected from students at both types of institutions.

Research Question:

*Is there a difference in anatomy and physiology students' conceptual understanding of the structure-function relationship between 2-year and 4-year institutions?*

Research Hypothesis:

*I hypothesize that differences in students' academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance.*

## **Methods**

### Question Development and Administration

Eight short answer questions based on the core concept of “structure-function” were administered to students in Human Anatomy and Physiology at one Southeastern 4-year college and two Southeastern 2-year colleges (Table 3.1). At all three institutions, the Human Anatomy and Physiology course is a two-semester course. At the 4-year institution, two semesters of General Biology and one semester of General Chemistry are prerequisites. However, at the 2-year institutions, there are no prerequisites.

The questions were administered throughout the semester as part of regular online homework via the course management system. Administration of each question occurred after the relevant topic was discussed in class. Students were asked to explain their answer to the best of their ability without the use of outside resources.

I collected 1,491 responses over five semesters from 437 students at the two-year institutions and 1,438 responses over six semesters from 453 students at the 4-year institution (Table 3.2). Responses were collected from the classrooms of five faculty at the two-year institutions and two faculty at the four-year institution. All three institutions use the same textbook.

**Table 3.1** Short-answer structure-function questions administered at one 4-year institution and two 2-year institutions.

Topic	Question name	Question prompt
Integumentary system/Skin layers	Two layers of skin	Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.
	Third degree burn	Victims of third degree, or full thickness, burns have their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin.
Muscular system/Skeletal muscle contraction	Contractile proteins	The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction.
	Rigor mortis	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.
Digestive system/Small intestine	Small intestine mucosa	Consider the mucosa of the small intestine. Based on form reflecting function, explain how this layer contributes to the functions of the digestive system.
	Celiac disease	Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system.
Cardiovascular system/Blood vessels	Arteries/arterioles	Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation.
	Arteriosclerosis	Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.

**Table 3.2.** Number of responses collected for short answer structure-function questions administered at one 4-year institution and two 2-year institutions.

Topic	Question name	N 4-year	N 2-year
Integumentary system/Skin layers	Two layers of skin	322	274
	Third degree burn	264	194
Muscular system/Skeletal muscle contraction	Contractile proteins	118	261
	Rigor mortis	173	255
Digestive system/Small intestine	Small intestine mucosa	190	124
	Celiac disease	139	129
Cardiovascular system/Blood vessels	Arteries/arterioles	145	129
	Arteriosclerosis	87	125
Totals		1438	1491

### Human Scoring

As discussed in chapter 2 of this dissertation, the student responses to the eight short-answer questions were scored by using a conceptual rubric designed for each question (Appendix, Table A.1). The human scoring rubric for each question identifies structures, functions, or a concept that links structure and function. Four coders scored a subset of responses and achieved an inter-rater reliability (Cronbach's alpha) of 0.7 or higher for each concept (Cronbach, 1984). Each rater was then assigned a subset of responses to code with at least two coders assigned to each response. After this round of independent coding, I resolved any disagreements. Each response was then coded for structure relates function (SRF) concepts (see Table 2.7, chapter 2).

## Statistical Analyses

To examine differences in conceptual understanding between the 2-year and 4-year student populations, I compared the SRF concepts for each question with a chi-square test for homogeneity (Marascuilo & McSweeney, 1977). The chi-square test for homogeneity is used to determine if a difference exists between two independent groups based on a binary dependent variable. In this study, both the independent and dependent variables are binary. The independent variable for the chi-square test is the institution, which has two values, 2-year or 4-year institution. The dependent variable is the presence or absence of structure-function concepts in the student responses. I compared the proportion of student responses that included the structure-function concept by institution; e.g., I compared the number of responses to the integument questions, which included the structure-function concept *protection*, between the 2-year and 4-year institutions. I compared the proportion of SRF codes between institutions for 14 structure-function (SRF) concepts by using the chi-square test of homogeneity. Because I performed multiple statistical comparisons, I applied the Bonferroni correction and lowered the critical p value from 0.05 to 0.01 to reject my null hypothesis (Shaffer, 1995).

## Results

My results demonstrate that there is a difference in conceptual understanding of the structure-function relationship between 2-year and 4-year students in anatomy and physiology with more 4-year students mentioning SRF concepts in their responses compared to the 2-year students. However, on average, less than 50% of students linked structure and function in their responses regardless of question topic or institution. For each topic, I will present a comparison of SRF concepts between institutions.

## Topic 1: Integumentary System/Skin layers

The integumentary system questions include “Two layers of skin” and “Third degree burn” (Table 3.1). The SRF concepts for both integument questions are *sensation*, *protection*, and *regulation* (Appendix, Table A.3). Overall, students wrote about the structure-function (SRF) concepts in less than 60% of their responses (Fig. 3.1).

### SRF Concept 1: Sensation

For the SRF concept of *sensation*, significantly more 4-year responses contained the idea regardless of which question was asked. The structure-function relationship for *Sensation\_Two layers of skin* was mentioned by 19% of the 2-year students and 31.1% of the 4-year students, a statistically significant difference ( $p=0.001$ ). The structure-function relationship for *Sensation\_Third degree burn* was noted by 40.2% of the 2-year students and 54% of the 4-year students, which is a statistically significant difference ( $p=0.002$ ) (Fig. 3.1).

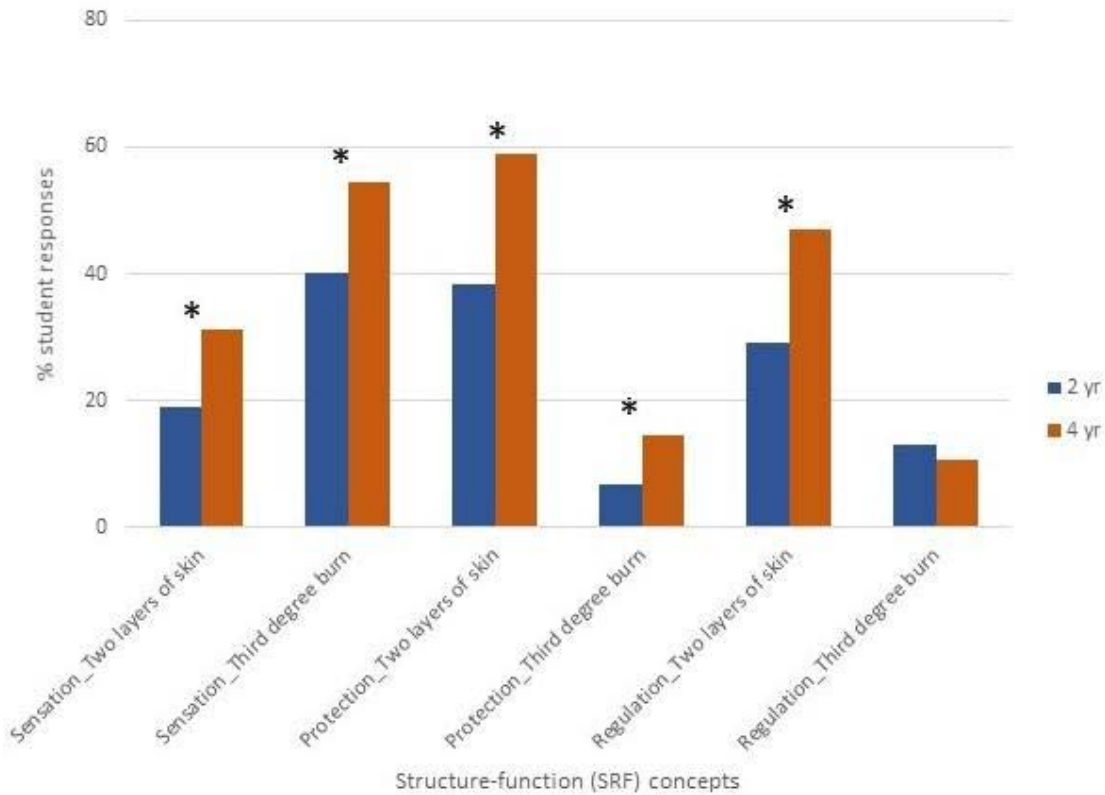
### SRF Concept 2: Protection

For the SRF concept of *protection*, there was a significant difference in the number of responses that included the idea between the two varieties of institution regardless of question. For *Protection\_Two layers of skin*, 38.3% of the 2-year students mentioned the structure-function relationship, while 59% of the 4-year students mentioned it, a statistically significant difference ( $p=0.000$ ). *Protection\_Third degree burn* was noted in 6.7% of the 2-year student responses and in 14.4% of the 4-year responses ( $p=0.010$ ) (Fig. 3.1).

### SRF Concept 3: Regulation

For the SRF concept of *regulation*, there was a significant difference in the number of responses that included the idea between institutions for the “Two layers of skin” question but not for the “Third degree burn” question. For *Regulation\_Two layers of skin*, 29.2% of the 2-year

responses mentioned the structure-function relationship, while 46.9% of the 4-year responses included it, a statistically significant difference ( $p=0.000$ ). For *Regulation\_Third degree burn*, 12.9% of the 2-year students wrote about the structure-function relationship, while 10.6% of the 4-year responses included it, which was not a significant difference (Fig. 3.1).



**Figure 3.1.** Percentage of student responses from 2-year and 4-year institutions that included integument structure-function concepts. \* significant  $p$  value  $<0.01$ .

## Topic 2: Muscular System/Skeletal muscle contraction

Four SRF concepts (*ATP necessary for contraction to end*, *myosin binds to actin*, *muscle contracts due to calcium*, and *sarcomere contractile unit*) were shared between two short-answer questions on muscle contraction that included the “contractile proteins” and “rigor mortis” questions (Table 3.1). The SRF concept of *ATP no longer available* was unique to the rigor

mortis question. The SRF concept of *muscle shortening* was unique to the contractile proteins question (Appendix, Table A.3). Overall, students wrote about the structure-function (SRF) concepts in less than 66% of their responses (Fig. 3.2).

#### SRF Concept 4: ATP Necessary for Contraction to End

For the SRF concept of *ATP necessary for contraction to end*, there was not a significant difference between the institution types for either question (Fig. 3.2). For the contractile proteins question, 4.6% of the 2-year students wrote about ATP being necessary for contraction to end, while 11% of the 4-year students wrote about it. For the rigor mortis question, 51.8% of the 2-year students and 54.3% of the 4-year students wrote about ATP being necessary for contraction to end (Fig. 3.2).

#### SRF Concept 5: Myosin Binds to Actin

For the SRF concept of *myosin binds to actin*, there was a significant difference in the number of responses that included the idea between the two institution types regardless of question. For *myosin binds to actin\_contractile proteins*, 51.3% of the 2-year students wrote about the idea in their responses, while 66.1% of the 4-year students wrote about it, a statistically significant difference ( $p=0.007$ ). For *myosin binds to actin\_rigor mortis*, 24.7% of the 2-year students wrote about the idea, whereas 49.7% of the 4-year students wrote about it, a statistically significant difference ( $p=0.000$ ) (Fig. 3.2).

#### SRF Concept 6: Muscle Contracts due to Calcium

For the SRF concept of *muscle contracts due to calcium*, there was a significant difference in the number of responses that included this idea between the two institutions regardless of question. For *muscle contracts due to calcium\_contractile proteins*, 12.6% of the 2-year students wrote about this idea in their responses, while 32.2% of the 4-year students wrote



about it, a statistically significant difference ( $p=0.000$ ). For *muscle contracts due to calcium \_rigor mortis*, 9.4% of the 2-year students wrote about this idea, and 21.4% of the 4-year students wrote about it, a statistically significant difference ( $p=0.001$ ) (Fig. 3.2).

#### SRF Concept 7: Sarcomere Contractile Unit

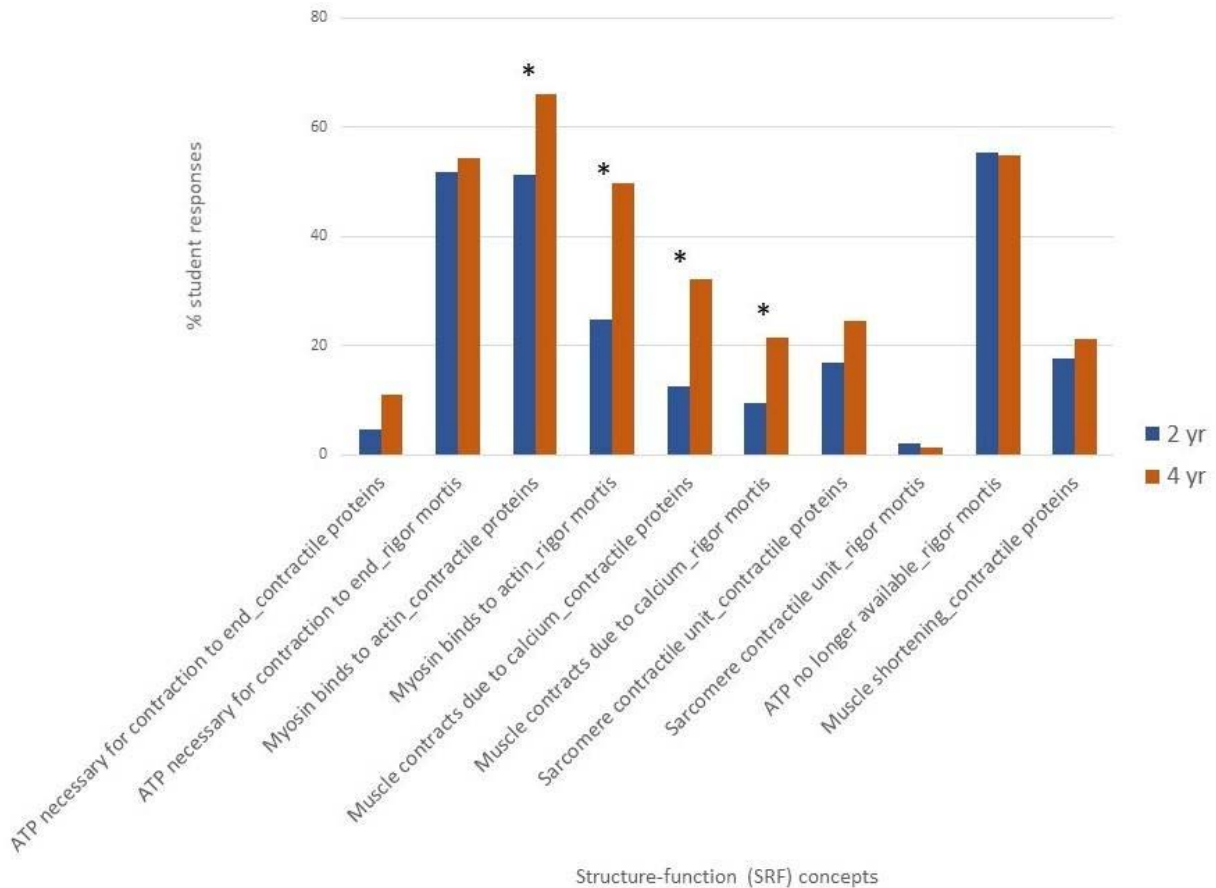
For the SRF concept of *sarcomere contractile unit* there was not a significant difference between the institutions for either question (Fig. 3.2). For the contractile proteins question, 16.9% of the 2-year students wrote about the sarcomere being the contractile unit, while 24.6% of the 4-year students wrote about it. For the rigor mortis question, 2% of the 2-year students and 1.2% of the 4-year students wrote about the sarcomere as the contractile unit (Fig. 3.2).

#### SRF Concept 8: ATP no Longer Available

The SRF concept of *ATP no longer available* was evaluated only for the rigor mortis question; there was not a significant difference between institution types. Among the 2-year students, 55.3% wrote about ATP no longer being available, while 54.9% of the 4-year students wrote about it (Fig. 3.2).

#### SRF Concept 9: Muscle Shortening

The SRF concept of *muscle shortening* was evaluated only for the contractile proteins question, and there was not a significant difference between institution types. Among the 2-year students, 17.6% wrote about muscle shortening, while 21.2% of the 4-year students wrote about it (Fig. 3.2).



**Figure 3.2.** Percentage of students' responses from 2-year and 4-year institutions that included muscle contraction structure-function concepts. \* significant p value <0.01.

### Topic 3: Digestive System/Small Intestine

Two SRF concepts, *absorption* and *digestion*, were shared between two small intestine short-answer questions, which included small intestine mucosa and celiac disease questions (Table 3.3). The SRF concepts of *secretion* and *protection* were unique to the small intestine mucosa question (Appendix, Table A.3). Overall, students wrote about the structure-function (SRF) concepts in less than 37% of their responses (Fig. 3.3).

#### SRF Concept 10: Absorption

For the SRF concept of *absorption*, there was not a significant difference between the institutions for either question (Fig. 3.3). For the small intestine mucosa question, 32.3% of the

2-year students wrote about the small intestine mucosa, while 36.8% of the 4-year students wrote about it. For the celiac disease question, 14% of the 2-year students and 18.7% of the 4-year students wrote about the absorption structure-function relationship in the small intestine mucosa (Fig. 3.3).

#### SRF Concept 11: Digestion

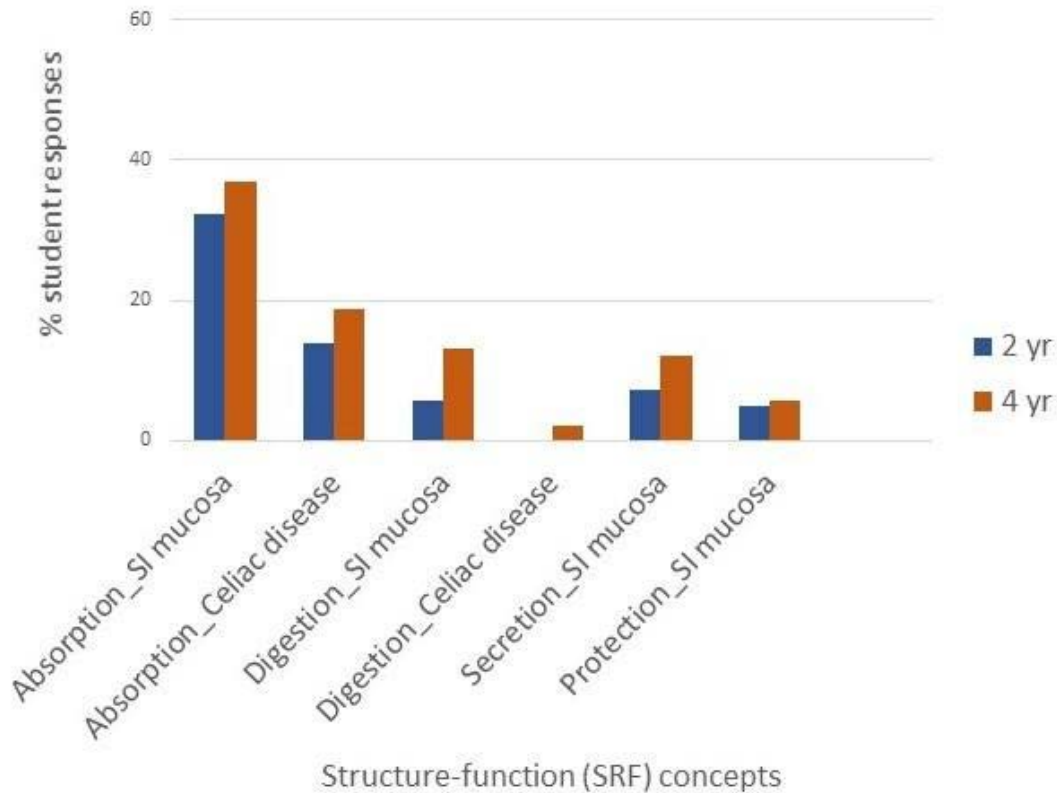
For the SRF concept of *digestion*, there was not a significant difference between the institution types for either question (Fig. 3.3). For the small intestine mucosa question, 5.6% of the 2-year students wrote about the structure-function of digestion in the small intestine mucosa, while 13.2% of the 4-year students wrote about it. For the celiac disease question, none of the 2-year students yet 2.2% of the 4-year students wrote about the digestion structure-function relationship in the small intestine (Fig. 3.3).

#### SRF Concept 12: Secretion

The SRF concept of *secretion* was evaluated only for the small intestine mucosa question, and there was not a significant difference between the institution types. Among the 2-year students, 7.3% wrote about the secretion structure-function relationship, while 12.1% of the 4-year students wrote about it (Fig. 3.3).

#### SRF Concept 13: Protection

The SRF concept of *protection* was evaluated only for the small intestine mucosa question, and there was not a significant difference between the institution types. Among the 2-year students, 4.8% wrote about the protection structure-function relationship, while 5.8% of the 4-year students wrote about it (Fig. 3.3).



**Figure 3.3.** Percentage of students' responses from 2-year and 4-year institutions that included small intestine structure-function concepts. No significant differences between the institutions.

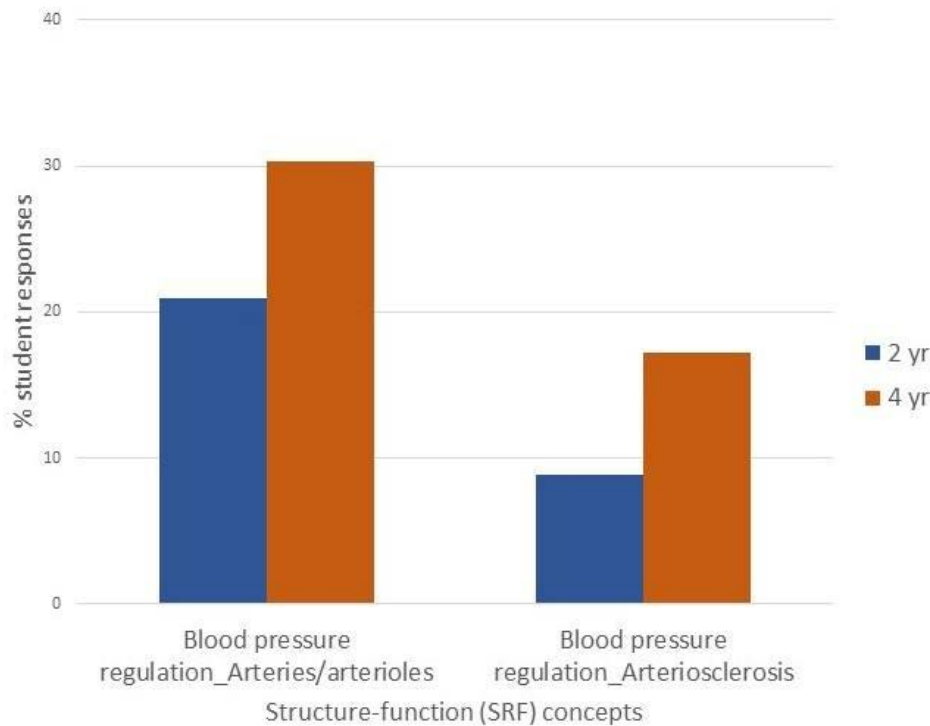
#### Topic 4: Cardiovascular System/Blood vessels

One SRF concept, *blood pressure regulation* (Appendix, Table A.3), was shared between the two short-answer questions on the blood vessels, which included the arteries/arterioles and arteriosclerosis questions (Table 3.1). Overall, students wrote about the structure-function (SRF) concept in less than 31% of their responses (Fig. 3.4).

#### SRF Concept 14: Blood Pressure Regulation

For the SRF concept *blood pressure regulation*, there was not a significant difference between the institution types for either question (Fig. 3.4). For the arteries/arterioles question, 20.9% of the 2-year students wrote about the blood pressure regulation structure-function

relationship, while 30.3% of the 4-year students wrote about it. For the arteriosclerosis question, 8.8% of the 2-year students and 17.2% of the 4-year students wrote about the blood pressure regulation structure-function relationship (Fig. 3.4).



**Figure 3.4.** Percentage of students' responses from 2-year and 4-year institutions that included blood pressure regulation structure-function concepts. No significant difference between the institutions.

## Discussion

Regardless of the type of institution offering anatomy and physiology, conceptual understanding of the structure-function relationship is necessary to understand anatomical and physiological processes. However, the type of institution (2-year or 4-year) may influence conceptual understanding. The focus of my research is to compare conceptual understanding of structure-function in 2-year versus 4-year anatomy and physiology students by using written

formative assessment and constructed response questions with responses collected from students at both types of institutions.

Research Question:

*Is there a difference in anatomy and physiology students' conceptual understanding of the structure-function relationship between 2-year and 4-year institutions?*

Research Hypothesis:

*I hypothesize that differences in students' academic readiness between two-year and four-year institutions may affect conceptual understanding and student performance.*

Conceptual understanding of the structure-function relationship is necessary to understand physiological processes, and yet, my results suggest that there is a difference in conceptual understanding based on institution. I found that the 4-year students mentioned 5 of the 14 SRF concepts more often than the 2-year students. These differences occurred only in the Anatomy & Physiology (A&P) I courses and not in A&PII. There was no difference in performance between institutions in A&PII. Potential reasons for this difference may be college readiness or academic integration.

College readiness

College readiness includes academic preparedness (e.g., GPA and standardized test scores) as well as behaviors related to success such as critical thinking, study skills, time management, and self-regulation (Barnes et al., 2010). Characteristics of underprepared community college students identified in one study include a lower expectation of achievement, greater test anxiety, and a lower course completion rate (Grimes, 2006), while another study found gender, race and GPA to be predictors of student success (Mamiseishvili & Deggs, 2013). In a longitudinal study that analyzed demographic and academic characteristics of students who

persisted to degree completion, time between high school and community college, and GPA were the significant predictors (Craig & Ward, 2008).

Anatomy and physiology is a gatekeeper course to health careers at 2-year and 4-year institutions. Since this course is offered at both types of institutions, it offers an opportunity to study preparedness between 2-year and 4-year institutions. Students enrolling in A&P I at a 2-year institution may be less college ready than their counterparts at a 4-year institution. In general, 2-year institutions have an open access policy, so that all students are able to attend with a high school diploma or GED being the only academic qualification, although dual enrollment high school students are also able to attend (Cohen & Brawer, 2003; Grubb, 1999). In contrast to the 2-year institutions, according to the National Association for College Admission Counseling, 4-year institutions are more selective, requiring standardized test scores and specific GPA thresholds, with average acceptance rates of 65% (NACAC, 2018). The 4-year institution factored into this study requires standardized test scores in the top 30% nationally and a high GPA, with an overall acceptance rate of 47%, and is moderately selective.

Compared to the 2-year students, the 4-year students have had more opportunities to be exposed to science content and to develop self-regulatory skills before taking A&P I. For example, 4-year students typically take a year of biology and one semester of chemistry as prerequisites for anatomy and physiology. By taking these college courses first, the 4-year students have potentially assimilated college readiness skills such as critical thinking, study habits, and time management to help them be more successful in anatomy and physiology. In contrast, students at a 2-year institution may be taking anatomy and physiology as their first college course because there are no prerequisites.

## Academic integration

The concept of academic integration is closely linked with students' academic performance in college (Mamiseishvili and Deggs, 2013), and academic and social integration are intertwined (Karp et al., 2010). Academic integration refers to students becoming attached to the intellectual component of college, while social integration is the relationships and connections outside of the classroom (Karp et al., 2010). Integration, or a sense of belonging, was correlated with persistence to the second year in community college students (Karp et al., 2010). However, Mamiseishvili and Deggs (2013) found academic integration to contribute to persistence while social integration did not have an effect on the likelihood of persistence in community college students.

Compared to the 2-year students, the 4-year students have had more opportunities to be academically and socially integrated before taking A&P I. For example, 4-year students typically reside on campus during their first years of college and have more opportunities to participate in student organizations and clubs (Pichon, 2015). The 4-year students have potentially fostered a sense of belonging and social integration to be more successful in anatomy and physiology. In contrast, students at a 2-year institution may be commuter students with limited opportunities for involvement outside the classroom (Pichon, 2016).

One way to develop college readiness skills and academic integration at 2-year institutions is to assign prerequisites for anatomy and physiology. However, the few studies on prerequisites for anatomy and physiology have had mixed conclusions. Sturges & Maurer (2013) identified previous coursework in biology and chemistry as being positively correlated with student success in anatomy and physiology at a 4-year university. On the other hand, Forgey (2016) described a natural science prerequisite course as having a negative correlation to student



success in anatomy and physiology, while a general biology prerequisite course had a positive correlation to student success at a 2-year college. A further example of a positively correlated prerequisite is an immersion general chemistry course designed to facilitate student success (Lloyd & Eckhardt, 2010). Prerequisite courses which employ active learning and collaboration allow students to connect with peers and instructors and thereby facilitating both social and academic integration (Karp et al., 2010). Although there appears to be mixed results from adding prerequisites for anatomy and physiology, a general biology course, or even a medical terminology course, could provide college readiness skills and academic integration to 2-year students, which may increase their preparedness for the course.

On the other hand, adding prerequisites would add more classes to the 2-year curriculum compared to the 4-year curriculum, thus causing students to take more time to graduate. Taking more time to complete a degree increases the chance that students may not finish a degree (Forgey, 2016). Currently, attrition in anatomy and physiology courses at 2-year institutions is around 50% (Harris et al., 2004). With a high attrition rate already, adding more classes to the curriculum could potentially increase the attrition rate further. If the prerequisite classes were designed to facilitate college readiness skills, then they may decrease the attrition rate. Furthermore, beyond taking longer to graduate, adding prerequisites may cause students to face financial burdens by paying for the additional courses (Forgey, 2016). Adding prerequisites may be an additional expense but doing so may ensure that students matriculate through the class at the 2-year institution on the first attempt rather than having to take the class a second time.

#### Conceptual Understanding in A&P II

There was no difference in performance between institution types on structure-function concepts examined in the A&P II course. However, conceptual understanding of the structure-

function relationship was lower for A&P II students (digestive and cardiovascular systems) compared to the A&P I students. Less conceptual understanding in A&P II students is perplexing as these students should be more prepared after having completed one semester of anatomy and physiology.

A possible reason why A&P II responses showed fewer instances of conceptual understanding may be that students find the particular organ systems targeted in the questions to be challenging. The short-answer questions administered to A&P II courses examined 4 SRF concepts related to the digestive system (absorption, digestion, secretion, and protection) and 1 SRF concept related to the cardiovascular system (blood pressure regulation). Other studies have demonstrated students have difficulty with these organ systems. Prokop & Fancovicova (2006) evaluated first-year undergraduates from a 4-year institution for their knowledge of concepts related to all of the organ systems and found that 50% of students were successful with concepts related to the digestive system, and 60% were successful with concepts related to the cardiovascular system. Michael et. al. (2002) found that students from 2-year and 4-year institutions have a number of conceptual difficulties regarding the cardiovascular system, including pressure/flow/resistance relationships and blood pressure regulation. The prevalence of conceptual difficulties related to the digestive and cardiovascular systems appears to be uniform across diverse student populations. Future studies should include formative assessment of additional structure-function concepts for these two organ systems, such as mechanisms of absorption and pressure/flow/resistance relationships, to determine whether having multiple opportunities and varying contexts in which to apply the structure-function relationship aids in student learning of the core concept. In addition to the structure-function core concept, conceptual understanding of these organ systems necessitates knowledge of other core concepts,

such as homeostasis, information flow, matter/energy transfer/transformation, and levels of organization (Michael et al., 2009), which have also been demonstrated to be difficult for students. Therefore, further examination of these concepts and their connections within the digestive and cardiovascular systems may help to inform the difficulty with conceptual understanding observed in this study.

### Conceptual Understanding of Structure-Function

On average, less than half of students in both A&P I and A&P II demonstrated conceptual understanding regardless of institution. Prior studies have shown that students have difficulty understanding the structure-function relationship (Carter & Prevost, 2018; Lira & Gardner, 2017). A possible reason for such performance on these structure-function questions may be levels of organization although the question topics in this study are at the molecular, cellular, tissue and organ levels. The short-answer questions in this study were intentionally designed around multiple levels of organization. In chapter 2 of this dissertation, I found that students were more comfortable with macroscopic levels of organization, with fewer students referring to the molecular and cellular levels of organization in their responses. Some of the conceptual difficulty with these questions may be due to the focus on molecular and cellular levels (e.g., muscular system with actin and myosin). However, students struggled with the questions at the tissue and organ levels of organization. Therefore, levels of organization may not be contributing to the difficulty with conceptual understanding of the structure-function relationship observed in this study. Further research is necessary to explicitly focus on the role of levels of organization in student conceptual understanding.

Another possible reason for students lacking conceptual understanding may be the cognitive levels of the short-answer question prompts. The questions in this portion of the study

are at the “understand” and “apply” levels, which are the second and third levels of Bloom’s taxonomy, respectively (Anderson et al., 2001). This possibility will be explored further in chapter 4 of this dissertation.

### Implications for Teaching

If students from 2-year institutions are lacking college readiness skills and academic integration to succeed in Anatomy & Physiology, they may not be able to complete an allied health degree or attain a bachelor's degree. Two-year colleges are an essential component of higher education because almost one half of students who receive a bachelor's degree in science and engineering attend a community college at some point in their education (Olson & Labov, 2012). Community colleges have the potential to provide groundwork by focusing on college readiness skills, such as critical thinking, study skills, and time management (Wang, 2015).

Two-year institutions may want to consider adding an A&P boot camp to help students prepare academically for anatomy and physiology (Garrett, 2012). Similar endeavors with pre-semester week-long boot camps have led to increased student success and retention for biology majors (Wischusen & Wischusen, 2007; Wheeler & Wischusen, 2014) and for STEM majors (Findley-Van Nostrand & Pollenz, 2017). The Biology Intensive Orientation for Students (BIOS) biology boot camp has shown to be effective at retention by increasing student success through higher grades on class exams and final course grades (Wischusen & Wischusen, 2007) and by developing self-efficacy, self-regulation, and a sense of belonging (Wheeler & Wischusen, 2014). The STEM Academy has also been shown to be effective at retention by increasing a sense of belonging, enhancing students’ science identity, and by developing self-efficacy (Findley-Van Nostrand & Pollenz, 2017). An A&P boot camp could be a week-long experience that occurs prior to the start of the semester, and it could be designed to prepare the students for

the rigor of anatomy and physiology. During this period, students could be exposed to anatomy and physiology course materials, basic study skills, anatomical terminology, and interactive quizzes prior to the start of the semester. A&P boot camp activities could include listening to lectures, participation in active learning exercises, and a laboratory activity. Such a boot camp could be provided for nominal costs, which include materials and textbooks. I recommend a boot camp textbook to help students to think critically and to focus on core concepts, such as the structure-function relationship, in a low-stakes format, and an example of such a textbook is *Get Ready for A&P* (Garrett, 2012). An A&P boot camp for students at 2-year institutions has the potential to facilitate the development of conceptual understanding and lead to college readiness skills in a pre-semester format, providing some of the benefits of a prerequisite course (e.g., critical thinking, time management skills, anatomy terminology, etc.), without increasing time to graduation/completion and financial cost.

These results suggest that students may benefit from a focus on core concepts within the content of anatomy and physiology courses. This focus should occur in both the first and second semesters of anatomy and physiology. For example, the structure-function relationship should be introduced early in the first semester of anatomy and physiology, then reinforced as each organ system is encountered through the first and second semesters. Instructors can use written formative assessment to allow students to demonstrate their conceptual understanding within the organ systems.

## Conclusion

In summary, there is a difference in conceptual understanding of the structure-function relationship between 2-year and 4-year students in anatomy and physiology. These differences occurred only in the Anatomy & Physiology (A&P) I courses and not in A&P II. There was no

difference in performance between institutions in A&P II. However, conceptual understanding of the structure-function relationship was lower for A&P II students (digestive and cardiovascular systems) compared to the A&P I students. On average, less than 50% of students linked structure and function in their responses regardless of question topic or institution. My results from written formative assessment suggest that both 2-year and 4-year college students have difficulty with conceptual understanding of the structure-function relationship.

## References

- American Association of Colleges of Nursing (2014). Nursing shortage fact sheet. Retrieved from <https://www.aacnnursing.org/News-Information/Fact-Sheets/Nursing-Shortage>
- Anderson L.W., Krathwohl D.R., Bloom B.S., Bloom B.S. (2001). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman.
- Barnes, W., Slate, J., & Rojas-LeBouef, A. (2010) College-readiness and academic preparedness: The same concepts? *Current Issues in Education* 13(4): 1-28.
- Carter, K.P. & Prevost, L.B. (2018) Question order and student understanding of structure and function. *Advances in Physiology Education* 42(4):576-585.
- Cohen, A. M., & Brawer, F. B. (2003). *The American community college*. John Wiley & Sons.
- Craig, A. J., & Ward, C. V. L. (2007). Retention of Community College Students: Related Student and Institutional Characteristics. *Journal of College Student Retention: Research, Theory & Practice*, 9(4), 505–517.
- Cronbach, L.J. (1984). A research worker's treasure chest. *Multivariate Behavioral Research* 19:223-240.
- Davis, G.M. (2010). What is provided and what the registered nurse needs-bioscience learning through the pre-registration curriculum. *Nurse education today*, 30(8):707-712.
- Feder, M.E. (2005). Aims of undergraduate physiology education: A view from the University of Chicago. *Advances in Physiology Education* 29(1):3-10.
- Findley-Van Nostrand, D. & Pollenz, R.S. (2017). Evaluating Psychosocial Mechanisms Underlying STEM Persistence in Undergraduates: Evidence of Impact from a Six-Day Pre-College Engagement STEM Academy Program. *CBE—Life Sciences Education* 16(2):1-15.

Forgey, S.B. (2016). An evaluation of pathways to community college success in anatomy and physiology I. ProQuest Dissertations & Theses Global. (1874351428). Retrieved from <https://search.proquest.com/docview/1874351428?accountid=14745>

Garrett, L.K. (2012). *Get Ready for A&P*. Boston, MA: Pearson

Grimes, S. (2006) Underprepared Community College Students: Characteristics, Persistence, And Academic Success. *Community College Journal of Research and Practice* 21(1):47-56.

Grubb, N.W. (1999). Learning and earning in the middle: The economic benefits of sub-baccalaureate education. Occasional paper. New York: Community College Research Center, Teachers College, Columbia University.

Harris, D.E., Hannum, L., & Gupta, S. (2004). Contributing factors to student success in Anatomy and Physiology: Lower outside workload and better preparation. *The American Biology Teacher* 66(3):168-175.

Hopper, M. (2011). Student enrollment in a supplement course for anatomy and physiology results in improved retention and success. *Journal of College Science Teaching* 40(3):70-79.

Karp, M. M., Hughes, K. L., & O’Gara, L. (2010). An exploration of Tinto’s integration framework for community college students. *Journal of College Student Retention: Research, Theory and Practice*, 12(1), 69–86.

Lira, M.E. & Gardner, S.M. (2017). Structure-function relations in physiology education: Where’s the mechanism? *Advances in Physiology Education* 41: 270-278.

Lloyd, P.M., & Eckhardt, R.A. (2010). Strategies for Improving Retention of Community College Students in the Sciences. *Science Educator* 19(1):33-41.

MacDowell, M., Glasser, M., Fitts, M., Fratzke, M. & Peters, K. (2009). Perspectives on rural health workforce issues: Illinois-Arkansas comparison. *The Journal of Rural Health* 25(2): 135-140.

Mamiseishvili, K., & Deggs, D. M. (2013). Factors Affecting Persistence and Transfer of Low-Income Students at Public Two-Year Institutions. *Journal of College Student Retention: Research, Theory & Practice*, 15(3), 409–432.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Belmont, CA: Wadsworth Publishing Company.

Melguizo, T. & Dowd, A.C. (2009). Baccalaureate success of transfers and rising 4-year college juniors. *Teachers College Record* 111(1):55-89.

Michael, J., Wenderoth, M.P., Modell, H.I., Cliff, W., Horwitz, B., McHale, P., Richardson, D., Silverthorn, D., Williams, S. & Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education* 26(2):72-84.

Michael, J. (2007). What makes physiology hard for students to learn? Results of a faculty survey. *Advances in Physiology Education* 31:34-40.

Michael, J., Modell, H. McFarland, J. & Cliff, W. (2009). The “core principles” of physiology: what should students understand? *Advances in Physiology Education* 33:10-16.

National Association for College Admission Counseling (NACAC). (2018). *Chapter 1: College Applications*. Retrieved from <https://www.nacacnet.org/news--publications/publications/state-of-college-admission/soca-chapter1/>

Newton, S.E., Smith, L.H., Moore, G., & Magnan, M. (2007). Predicting early academic achievement in baccalaureate nursing program. *Journal of Professional Nursing* 23(3):144-149.

Olson, S. & Labov, J. B. (2012). *Community Colleges in the Evolving STEM Education Landscape*. National Academies Press.

Pichon, H. W. (2016). Developing a sense of belonging in the classroom: community college students taking courses on a four-year college campus. *Community College Journal of Research and Practice*, 40(1), 47-59.

Prokop, P. & Fancovicova, J. (2006). Students' ideas about the human body: Do they really draw what they know? *Journal of Baltic Science Education* 2(10):86-95.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology* 46: 561-584.

Sturges, D. & Maurer, T. (2013). Allied health students' perception of class difficulty: The undergraduate human anatomy and physiology class. *The Internet Journal of Allied Health Sciences and Practice* 11(4):1-10.

Wang, X. (2015). Pathway to a Baccalaureate in STEM Fields: Are Community Colleges a Viable Route and Does Early STEM Momentum Matter? *Educational Evaluation and Policy Analysis*, 37(3): 376-393.

Wheeler, E. R., & Wischusen, S. M. (2014). Developing Self-regulation and Self-efficacy: A Cognitive Mechanism for Success of Biology Boot Camps. *Electronic Journal of Science Education*, 18(1):1-15.

Wischusen, S. M., & Wischusen, E. W. (2007). Biology intensive orientation for students (BIOS): a biology “boot camp”. *CBE—Life Sciences Education*, 6(2): 172-178.



## CHAPTER FOUR COMPARISON OF QUESTION FEATURES

**A Note to Reader:** Portions of this chapter have been published by the American Physiological Society. Permission has been granted by the publisher. KPC is the first author of the published work. Luanna Prevost is second author and my Ph.D. advisor. Documentation of approval is in the Appendix.

### **Abstract**

Short answer essay questions contain features which are elements of the question which aid students in connecting the question to their existing knowledge. Varying the features of a question may be used to provide insight into the different stages of students' emerging biological expertise and differentiate novice students who have memorized an explanation from those who exhibit understanding. I am interested in examining the cognitive level of questions, the use of guiding context/references in question prompts, and the order of questions, and how these features elicit student explanations of the core concept structure-function in anatomy and physiology. I hypothesize that varying the features (cognitive level, guiding context and question order) of short answer questions may affect student explanations. Short answer questions based on the core concept 'structure-function' were administered to 767 students in a junior level General Physiology course and to 573 students in a sophomore level Human Anatomy and Physiology course at a large southeastern public university. Student responses were first human scored and then scored by using lexical analysis and machine scoring. Students were interviewed to examine their familiarity with levels of organization and to confirm their

interpretation of the questions. Students demonstrated more conceptual understanding of four of the structure-function concepts when answering the understand questions and more conceptual understanding of two structure-function concepts when answering the apply questions. The question prompts provided a different context which may have influenced student explanations. There was no difference in conceptual understanding of the structure-function relationship with and without the use of a guiding context in the wording of the question prompt. For question sequence, students performed better on the last questions in the sequence, regardless of whether the last question was easier or more difficult. Instructors should provide students with questions in varying contexts and cognitive levels will allow students to demonstrate their heterogeneous ideas about a concept.

## **Introduction**

Short answer essay questions contain features that are elements of the question which aid students in connecting the question to their existing knowledge (Goldstein, 2011). Question features are the superficial characteristics of a question prompt that can be changed without altering the underlying concept being assessed (Federer et al., 2015). Question features may influence student explanations by acting as a knowledge retrieval cue (Goldstein, 2011). Numerous studies have indicated that the features and formats of constructed response questions influence student explanations (Federer et al., 2015; Nehm & Ha, 2011; Opfer et al., 2012; Prevost et al., 2013). For example, in biology, the taxa used in the question may affect how students respond in some cases but not others. When students were asked to describe how natural selection may lead to the gain of a trait, students provided more complete responses when describing the gain of a trait in a familiar animal such as a cheetah than the gain of a trait in an unfamiliar animal such as a locust (Nehm & Ha, 2011). Students discussed varying types of

mutations with greater frequency when the question stem referred to animals compared to bacteria (Prevost et al., 2013). However, other studies have not found student explanations to be influenced by question features (Weston et al., 2015). For example, when students were asked to explain photosynthesis with different species in the question stems (corn vs. peanut), results were similar for both plant species. In each case, more than half of the students demonstrated correct conceptions of photosynthesis (Weston et al., 2015). Varying the features of a question may be used to provide insight into the different stages of students' emerging biological expertise and differentiate novice students who have memorized an explanation from those who exhibit understanding. This chapter investigates the effect, if any, of the question features on student responses.

As a student views a short answer question prompt, attention is focused on relevant pieces of information presented (Martinez, 1999). For example, an underlying concept in the question prompt might be recognized by the student, which would be processed by working memory. Once the information arrives in working memory, it is connected to existing knowledge, which is contained in both working memory and long-term memory, and this connection causes the information to become reorganized (Glynn & Muth, 1994; Martinez, 1999; Mayer, 1992). As expertise develops in a subject, the way in which the information is processed changes. Students take a more naive approach, categorizing problems based on recognizing question features, while experts categorize problems based on recognizing underlying core concepts (Chi et al., 1981; Opfer et al., 2012; Smith et al., 2013). Additionally, experts not only recognize underlying concepts, but they are more likely to be able to apply their knowledge (Chi et al., 1981). One of the goals of formative assessment is to help students make these connections and reorganize their knowledge as they move from novice to expert ideas and develop scientific

literacy (Bell & Cowie, 2001; Chi et al., 1981; Glynn & Muth, 1994). Although the effect of many question features on prompting student understanding can be investigated, I am interested in examining the cognitive level of questions, the use of guiding context/references in question prompts, the order of questions, and how these features elicit student explanations of the core concept structure-function in anatomy and physiology.

## Investigating Student Responses to Varying Question Features

### Cognitive Level

The Bloom taxonomy provides a framework for evaluating students' cognitive processes and can be used during the development of formative assessment short answer questions.

Bloom's taxonomy is a framework of hierarchical categories in which assessment methods and learning objectives are classified (Bloom, 1956; Anderson et al., 2001). Bloom's taxonomy has been used to evaluate learning outcomes and assessment in K-12 education since the 1960s but only in limited contexts in higher education (Crowe et al., 2008). There are six levels of cognition in Bloom's taxonomy, and my research will focus on the first three levels. The first level of Bloom's taxonomy is "remember", which refers to retrieving information from long term memory and is typically associated with recall or recognition tasks (Anderson et al., 2001). The second level is "understand", which goes beyond simply remembering material and refers to grasping the meaning and extrapolating information. Because the taxonomy is a hierarchical framework, remembering is necessary for understanding. The third level is "apply" in which existing knowledge is applied to a novel problem (Anderson et al., 2001). The final three levels of cognitive processes in the hierarchy are "analyze", "evaluate" and "create" (Anderson et al., 2001).

In order for meaningful learning to occur, students must use cognitive processes other than “remember.” Meaningful learning refers to a student being able to understand what they have learned and apply it in a novel context (Anderson et al., 2001). For example, formative assessment short answer questions at the “understand” level of Bloom’s taxonomy will encourage students to retrieve information related to terms and concepts, while short answer questions at the “apply” level will encourage students to engage in application-style thinking behaviors and to use their knowledge to solve a problem.

### Guiding Context

The context of a question prompt may influence student responses. For example, providing a specific context in a question prompt may cue the student to the salient elements and facilitate knowledge retrieval. If short answer questions are designed with either a reference in the question prompt to the core concept structure-function, or with no reference to the structure-function relationship, these questions can provide insight into the type of learning that is occurring.

Providing the core concept in the question prompt may assist novice learners in responding to the question. In the question with the reference to the core concept, knowledge in a specific context is being elicited (Duit, 1991). Novice level learners benefit from having the specific context provided in a question prompt as this context helps them to make the connection to their existing knowledge (Duit, 1991). For example, when students are asked to describe pressure differences in two scenarios, with and without a reference to atmospheric pressure in one of the scenarios, students who responded to the prompt with a reference to atmospheric pressure provided more partially correct responses, with few students provided a completely correct scientific explanation (Clough & Driver, 1986). Students who were not prompted with

atmospheric pressure had responses that included more incorrect alternative ideas (Clough & Driver, 1986). In the question without the reference to the concept, the student has to cognitively retrieve the information regarding the core concept, and then apply the information in a new context, which is more difficult and requires the student have access to their knowledge and exhibit comprehension (Anderson et al., 2001).

### Question Sequencing

Short answer questions that are ranked as “remember questions” require the student to access his or her knowledge and use it to answer the question, while “understand” questions provide a knowledge framework to the student (Anderson et al., 2001; Duit, 1991). A “remember” question, such as “Define the principle: form reflects function”, requires the student to retrieve the core concept without context, which is cognitively more difficult compared to an “understand” question. An “understand” question, such as “Give an example of the principle form reflects function from the human body”, directs the students’ attention to the core concept in a specific context rather than the student having to retrieve it. The order in which these questions are presented to the students may affect their ability to cognitively retrieve the information and apply the concept. Asking the “remember” question first (more difficult) may distract students’ attention from the core concept and the context. However, asking the “understand” question first (cognitively easier) will direct the students’ attention to the core concept and the context (Duit, 1991; Gentner & Toupin, 1986).

Question order may also elicit conceptual priming and affect student explanations. When students are asked a question, they search their memories to retrieve the information. The search is truncated as soon as enough information is found to answer the question. According to the theory of increased cognitive accessibility, their response to the next question will be based on

the information recently retrieved (Schwarz & Strack, 1991); this phenomenon is termed conceptual priming. Exposure to a concept acts as a prime, which then activates memories associated with the prime in subsequent questions. A larger number of preceding questions would increase the amount of potentially relevant information retrieved and may make subsequent questions cognitively easier (Carter & Prevost, 2018).

Prior research demonstrates mixed results for question order and conceptual priming; much of that research focused on multiple choice questions. Question order is more likely to have an effect on student performance when the multiple-choice assessment is given under a speed condition; students have a certain amount of time to complete the assessment (Leary & Dorans, 1985). However, question sequencing effects are not observed in all cases. Huck and Bowers (1972) found no difference in performance between two versions of a multiple-choice final examination delivered to an undergraduate introduction to psychology class, with the only difference being the arrangement of easy-to-hard or hard-to-easy items. Similar results were also found in an undergraduate educational psychology class with two versions (easy-to-hard/hard-to-easy) of multiple-choice examinations (Brenner, 1964).

The sequence of the short answer questions may affect student explanations, which should be considered when evaluating student understanding of a core concept. By developing short answer questions and varying the cognitive level, guiding context, and question sequencing, and by comparing student responses to these questions, my research investigates the breadth of student understanding of the relationship between structure and function in anatomy and physiology.

#### Research Questions:

1. *How does varying the features of short answer questions affect student explanations about the structure-function relationship in anatomy and physiology?*
2. *Do student responses to understand and apply level question reveal differences in their conceptual understanding of structure and function?*
3. *How do student descriptions of the structure-function relationship differ when answering a question prompt with reference to the core concept compared to students answering a question prompt without the reference?*
4. *How does varying the order of questions from different cognitive levels affect student explanations of the structure-function relationship?*

#### Research Hypotheses:

1. *I hypothesize that varying the features (cognitive level, guiding context and question order) of short answer questions may affect student explanations.*
2. *I hypothesize that there is a difference in conceptual understanding based on the cognitive level of the question prompts.*
3. *I hypothesize that there is a difference in conceptual understanding based on the reference to the core concept in the question prompt.*
4. *I hypothesize that there is a difference in conceptual understanding between the question orders.*

#### **Methods**

##### Question Development and Administration

Short answer questions based on the core concept structure-function were administered to 767 students in a junior level General Physiology course and to 573 students in a sophomore



level Human Anatomy and Physiology course at a large southeastern public university. The questions were administered throughout the semester as part of regular online homework via the course management system. Administration of each question occurred shortly after the relevant topic was discussed in class. Students were asked to explain their answer to the best of their ability without the use of outside resources. For a subset of each of the questions from the previous chapters, I modified the question prompts to examine how changes to the cognitive level, guiding context, or question sequence influenced student responses.

### Cognitive Level

I administered short answer questions at two cognitive levels: understand and apply (Anderson et al., 2001) (Table 4.1). For each topic, students answered an understand question followed by an apply question.

### Guiding Context

To compare how the presence or absence of guiding context influences student explanations, I varied the prompts of four short answer questions. Two versions of each question were administered: one with the phrase “Based on form reflecting function” in the question prompt and the other version without this phrase (Table 4.2). Each class was randomly split in half. Half the class received the question prompt with the reference to the structure-function relationship (version A) and the other half received version B. Responses were collected over three semesters.

**Table 4.1** Short answer questions administered to students in General Physiology and Human Anatomy and Physiology with question prompts at the understand and apply cognitive levels.

Topic	Question name	Cognitive level	Question prompt
Integumentary system/Skin layers	Two layers of skin	Understand	Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.
	Third degree burn	Apply	Victims of third degree, or full thickness, burns have their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin.
Muscular system/Skeletal muscle contraction	Contractile proteins	Understand	The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction.
	Rigor mortis	Apply	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.

### Question Sequencing

Students were asked to define and give examples of the concept structure-function. These questions were administered to students in a General Physiology course (Table 4.3; Carter & Prevost, 2018). The class was randomly split in half, and each half received the questions in a different order. Half of the students answered format DX (define followed by give an example), and the other half answered format XD (give an example followed by define) (Table 3). For each question format, students were asked to provide one definition and three examples. Students were not able to return to a question within the sequence.

## Human Scoring of Responses

Human scoring of all responses occurred as discussed in chapter 2 of this dissertation. The student responses to the eight short answer questions were scored using a conceptual rubric designed for each question (Appendix, Table A.1). As discussed in chapter 3, each response was then coded for structure relates function (SRF) concept. If the student included a structure (1) and a related function (1), the SRF concept would be (1).

## Computer-Automated Scoring

Responses to the define and give example questions were scored by using lexical analysis. Student responses to “Define the principle: form reflects function” and “Give an example of the principle: form reflects function” were analyzed by using IBM SPSS Modeler Text Analysis version 16 (SPSS, 2013). The steps of lexical analysis include extraction and categorization; thus, the software identified terms and phrases and grouped them into categories. The student’s written responses were categorized into zero, one, or more categories following extraction. The category grain size was hierarchical based on biological levels of organization from molecular to organ system (Carter & Prevost, 2018).

For the remaining questions, the results of human scoring were used for a training data set and a testing data set for machine scoring using an ensemble method as discussed in chapters 2 and 3. Ensemble methods combine machine scoring algorithms to obtain better predictive results than could be obtained by using only one machine algorithm. There are eight algorithms used in the ensemble method in this study, and each algorithm is used to predict the scoring of student responses in the dataset. Then, a final prediction is obtained by taking a weighted vote of the classifier predictions (Dietterich, 2000).

**Table 4.2.** Short answer questions administered to students in General Physiology and Human Anatomy and Physiology. Version A contained the structure-function relationship (guiding context, italicized) while Version B did not have the structure-function prompt.

Topic	Question # and version	Prompt
Muscular system/Skeletal muscle contraction/Rigor mortis	1A	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. <i>Based on form reflecting function</i> , explain the role of actin and myosin in the process of rigor mortis.
	1B	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Explain the role of actin and myosin in the process of rigor mortis.
Digestive system/Small intestine/Celiac disease	2A	Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. <i>Based on form reflecting function</i> , relate the damage of villi to the functions of the digestive system.
	2B	Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Relate the damage of the villi to the functions of the digestive system.
Cardiovascular system/Blood vessels/Arteries	3A	Arteries and arterioles are important in blood pressure regulation. <i>Based on structure reflecting function</i> , explain how the structure of these blood vessels contributes to blood pressure regulation.
	3B	Arteries and arterioles are important in blood pressure regulation. Explain how the structure of these blood vessels contributes to blood pressure regulation.
Cardiovascular system/Blood vessels/Arteriosclerosis	4A	Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. <i>Based on the principle form reflects function</i> , relate this diagnosis to the functions of the arteries and arterioles.
	4B	Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.

**Table 4.3.** Description of question format DX and XD. Each question format was administered to half of a General Physiology class. Students were asked to provide one definition and three examples (from Carter & Prevost, 2018).

Format	Description	Bloom's taxonomy
DX	Define the principle: form reflects function <i>followed by</i> Give an example of the principle: form reflects function from the human body	Remember <i>followed by</i> understand
XD	Give an example of the principle: form reflects function from the human body <i>followed by</i> Define the principle: form reflects function	Understand <i>followed by</i> remember

## Statistical Analyses

### Cognitive Level

To examine differences in conceptual understanding between the cognitive levels of question prompts for individual students, I compared the proportion of student responses mentioning SRF concepts for each question with a McNemar test (McNemar, 1947). The McNemar test is used to determine if there are differences on a dichotomous dependent variable between two related groups. In this study, both the independent and dependent variables are binary. The independent variable for the McNemar test is cognitive level, which has two values: understand and apply. The dependent variable is the presence or absence of structure-function concepts in the student responses. I compared the proportion of student responses that included the structure-function concept by cognitive level for each question topic. For example, I compared the number of responses to the integument questions that included the structure-function concept *protection* between the understand and apply cognitive levels. I compared the proportion of SRF codes between cognitive levels for each of seven structure-function (SRF) concepts using the McNemar test. Because I performed multiple statistical comparisons, I

applied the Bonferroni correction and lowered the critical p value from 0.05 to 0.01 to reject my null hypothesis (Shaffer, 1995).

### Guiding Context

To examine differences in conceptual understanding between the question formats, I compared the proportion of student responses mentioning SRF concepts for each question with a chi-square test for homogeneity (Marascuilo & McSweeney, 1977). The chi-square test for homogeneity is used to determine if a difference exists between two independent groups on a binary dependent variable. In this study, both the independent and dependent variables are binary. The independent variable for the chi-square test is question prompt which has two values, SRF prompt (guiding context) or no SRF prompt. The dependent variable is the presence or absence of structure-function concepts in the student responses. I compared the proportion of student responses that included the structure-function concept by question prompt. For example, I compared the number of responses to the rigor mortis question which included the structure-function concept *ATP necessary for contraction to end* between the question prompts. I compared the proportion of SRF codes between question prompts for each of eight structure-function (SRF) concepts using the chi-square test of homogeneity. Because I performed multiple statistical comparisons, I applied the Bonferroni correction and lowered the critical p value from 0.05 to 0.01 to reject my null hypothesis (Shaffer, 1995).

### Question Sequencing

The length of the written responses was compared between question formats to assess if students were more verbose with a definition question followed by a give example question (format DX), or a give example question followed by a definition question (format XD). The length of student responses was analyzed by using a Mann-Whitney U test (Mann & Whitney,

1947; Wilcoxon, 1945). I also performed a comparison of the hierarchical structure and function lexical categories from SPSS Modeler for format DX and format XD to determine if students used different words and phrases when they were asked to define the core concept compared to giving an example of the core concept. A Fisher's Exact Test analysis was performed to compare the lexical analysis categories between the DX and XD question formats.

### Student Interviews

I conducted interviews with four students following the interview protocol used by Haudek et al. (2012). Interviews began with a think aloud protocol during which students answered the same questions for which they had provided written responses in their homework. Details of the interview protocol are found in the Appendix. I analyzed their verbal responses to confirm that students were interpreting the questions in the manner intended and compared the verbal and written responses. I coded their verbal responses by using the structure, function, and structure relates to function categories used for written responses and compared the coding for written and verbal responses. I then identified the categories used in the verbal responses and compared them to categories identified in written responses (Carter & Prevost, 2018).

In the second part of the interview, I examined students' familiarity with levels of organization and their interpretation of the question wording. Students were first asked if they could recall the levels of organization. Then, they were asked which level of organization they typically found themselves thinking of for examples to identify student preferences within the hierarchy. Students were asked for their feedback on the question prompts, specifically their interpretation of the wording of the prompts. Students were then asked to define structure and function (Carter & Prevost, 2018). Different portions of the interviews support the results in the guiding context and question sequence sections of my results.

## Results

### Cognitive Level

I collected a total of 279 responses over two semesters from students in General Physiology and Human Anatomy and Physiology (Table 4.4). The responses are a subset of responses collected in the study discussed in chapter 2 of this dissertation. In this study each student answered the understand question followed by the apply question. For each topic, I will present a comparison of SRF concepts between cognitive levels.

**Table 4.4.** Number of responses collected for short answer structure-function questions at the understand and apply cognitive levels from students in General Physiology and Human Anatomy and Physiology.

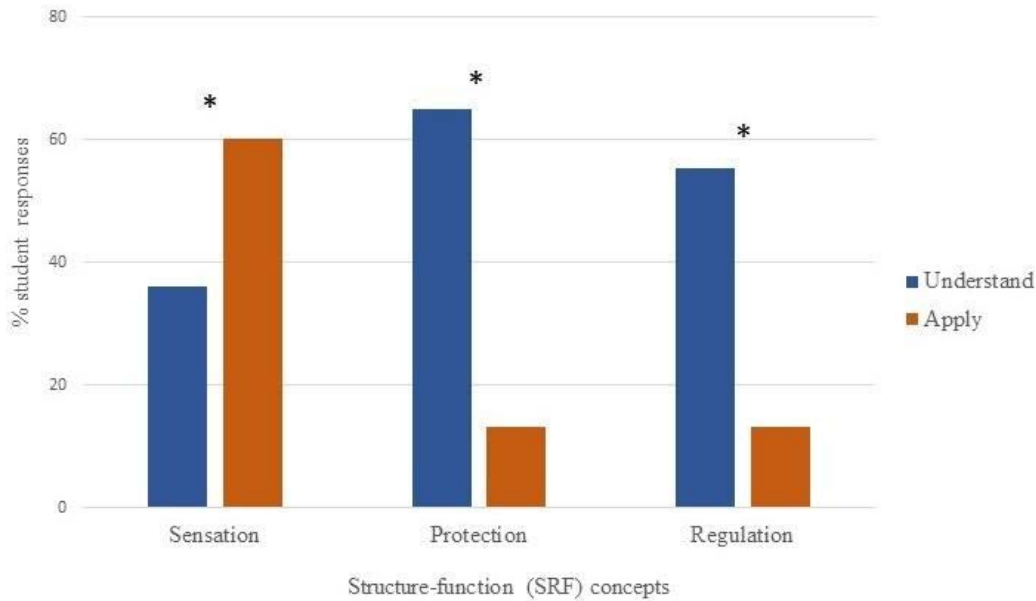
Topic	Question name	Cognitive level	N
Integumentary system/Skin layers	Two layers of skin	Understand	83
	Third degree burn	Apply	
Muscular system/Skeletal muscle contraction	Contractile proteins	Understand	196
	Rigor mortis	Apply	

#### Topic 1: Integumentary System/Skin Layers

The integumentary system questions used to examine the influence of cognitive level were “Two layers of skin” and “Third degree burn”. The “Two layers of skin” question is at the understand cognitive level while the “Third degree burn” question is at the apply level (Anderson et al., 2001). The SRF concepts identified in both of the integument questions are *sensation*, *protection* and *regulation* (Figure 4.1). The SRF concept *sensation* was mentioned in significantly more responses to the Apply question (60.2%) than the Understand question (36.1%)  $\chi^2$  (df=1, N=83)=7.848, p=0.000. However, significantly more responses contained the SRF concept *protection* in the Understand question (65.1%) than the Apply question (13.3%)  $\chi^2$  (df=1, N=83)=34.588, p=0.000 as well as the SRF concept *regulation* in the Understand question (55.4%)



than the Apply question (13.3%)  $\chi^2$  (df=1, N=83)=26.884, p=0.000, Figure 4.1, Appendix, Table A.4).

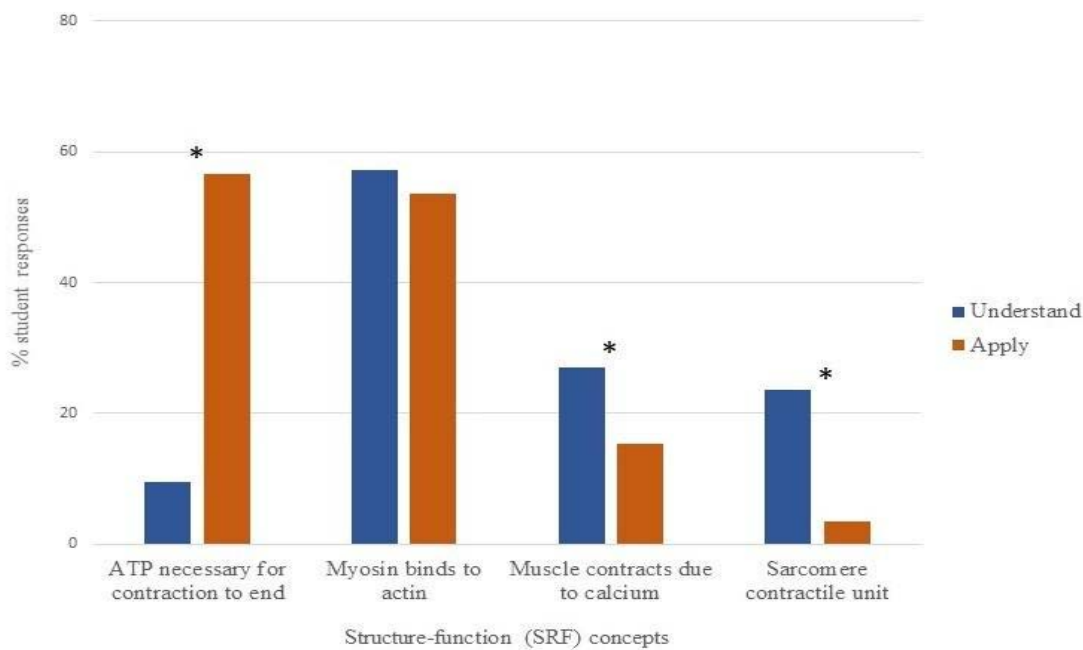


**Figure 4.1.** Percentage of student responses for “Two layers of skin” (Understand) and “Third degree burn” (Apply) questions that included integument structure-function concepts. \* significant p value <0.01.

### Topic 2: Muscular System/Skeletal Muscle Contraction

The muscle contraction questions used to examine the influence of cognitive level were “Contractile proteins” and “Rigor mortis”. The “Contractile proteins” question is at the understand cognitive level while the “Rigor mortis” question is at the apply level (Anderson et al., 2001). The SRF concepts identified in both of the muscle contraction questions are *ATP necessary for contraction to end, myosin binds to actin, muscle contracts due to calcium and sarcomere contractile unit* (Figure 4.2). The SRF concept *ATP necessary for contraction to end* was mentioned in significantly more responses to the Apply question (56.6%) than the Understand question (9.6%)  $\chi^2$  (df=1, N=196)=86.260, p=0.000. However, significantly more responses contained the SRF concept *muscle contracts due to calcium* in the Understand

question (27.1%) than the Apply question (15.3%)  $\chi^2$  (df=1, N=196)=9.490, p=.002 as well as the SRF concept *sarcomere contractile unit* in the Understand question (23.5%) than the Apply question (3.5%)  $\chi^2$  (df=1, N=196)=35.220, p=.000. There was not a significant difference between the cognitive levels for the SRF concept *myosin binds to actin* (Understand 57.1%, Apply 53.6%;  $\chi^2$  (df=1, N=196)=0.706, p=.401, Figure 4.2, Appendix, Table A.4).



**Figure 4.2.** Percentage of student responses for “Contractile proteins” (Understand) and “Rigor mortis” (Apply) questions that included skeletal muscle contraction structure-function concepts. \* significant p value <0.01.

### Guiding Context

For the guiding context comparison, I collected a total of 1,037 responses to four questions over three semesters from students in General Physiology and Human Anatomy and Physiology (Table 4.5). For each question, I will present a comparison of SRF concepts between question prompts. The data were initially separated by General Physiology and Human Anatomy

and Physiology courses, but there were no differences between the student responses, so the data were pooled (Appendix, Table A.6).

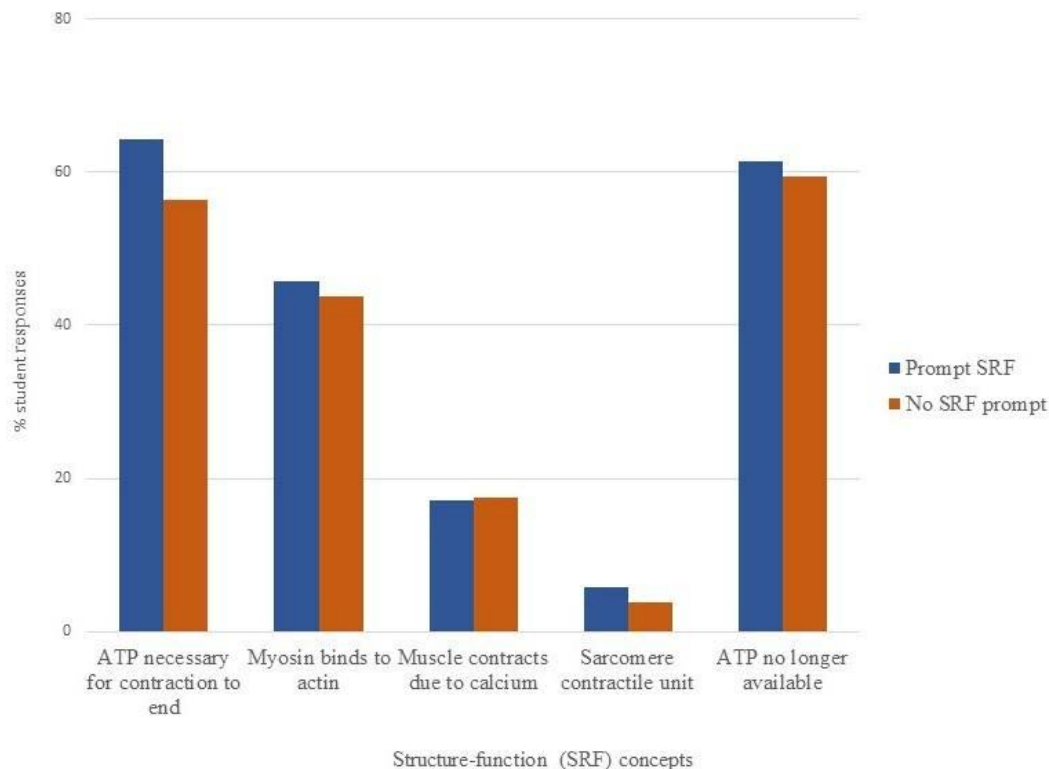
**Table 4.5.** Number of responses collected for short answer structure-function questions with either prompt or no prompt to the structure-function relationship.

Question # and version	Question prompt	N GP	N HAP
1A	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.	98	42
1B	A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Explain the role of actin and myosin in the process of rigor mortis.	123	38
2A	Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system.	57	45
2B	Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Relate the damage of the villi to the functions of the digestive system.	63	57
3A	Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation.	42	45
3B	Arteries and arterioles are important in blood pressure regulation. Explain how the structure of these blood vessels contributes to blood pressure regulation.	58	54
4A	Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Based on the principle form reflects function, relate this diagnosis to the functions of the arteries and arterioles.	97	45
4B	Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.	119	54
Totals		657	380

## Question 1: Muscular System/Skeletal Muscle Contraction/Rigor Mortis

The muscle contraction short answer question is the rigor mortis question (Table 4.1).

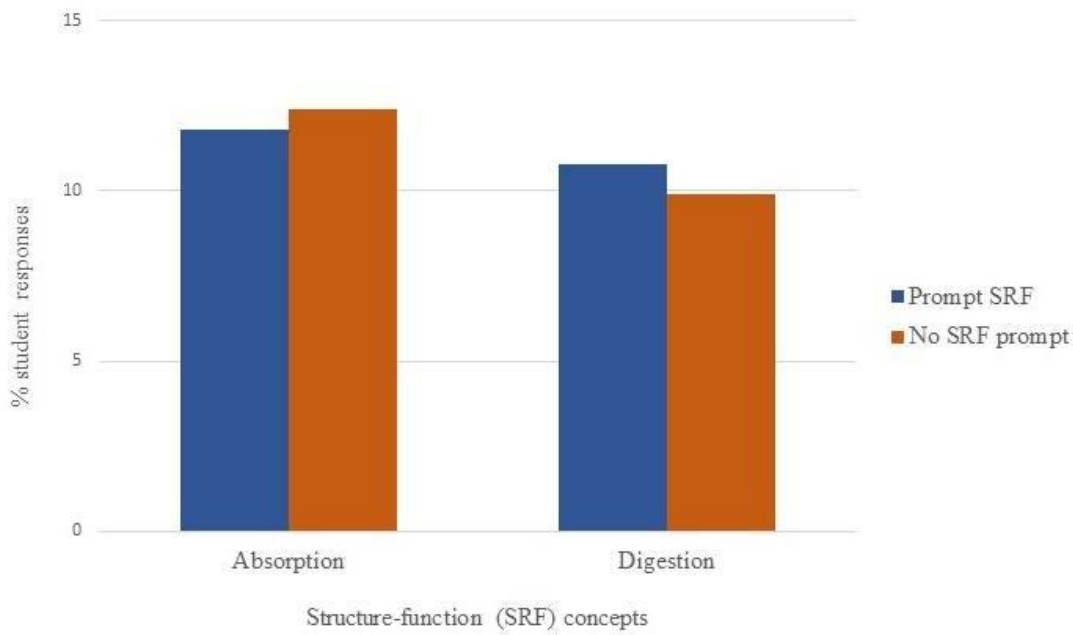
There are five SRF concepts within the rigor mortis question: *ATP necessary for contraction to end*, *myosin binds to actin*, *muscle contracts due to calcium*, and *sarcomere contractile unit* and *ATP no longer available* (Appendix, Table A.5). There was not a significant difference between the question prompts for the SRF concepts: *ATP necessary for contraction to end* ( $\chi^2_{(df=1, N=301)}=2.009, p=0.156$ ), *myosin binds to actin* ( $\chi^2_{(df=1, N=301)}=0.117, p=0.733$ ), *muscle contracts due to calcium* ( $\chi^2_{(df=1, N=301)}=0.007, p=0.935$ ), *sarcomere contractile unit* ( $\chi^2_{(df=1, N=301)}=0.648, p=0.421$ ) and *ATP no longer available* ( $\chi^2_{(df=1, N=301)}=0.132, p=0.717$ ) (Fig. 4.3).



**Figure 4.3.** Percentage of student responses from prompt SRF and no SRF prompt that included muscle contraction structure-function concepts.

### Question 2: Digestive System/Small Intestine/Celiac Disease

The small intestine short answer question is the celiac disease question (Table 4.1). There are two SRF concepts in the celiac disease question: *absorption* and *digestion* (Appendix, Table A.5). There was not a significant difference between the question prompts for the SRF concepts: *absorption* ( $\chi^2_{(df=1, N=222)}=0.021, p=0.885$ ) and *digestion* ( $\chi^2_{(df=1, N=222)}=0.045, p=0.832$ ) (Fig. 4.4).



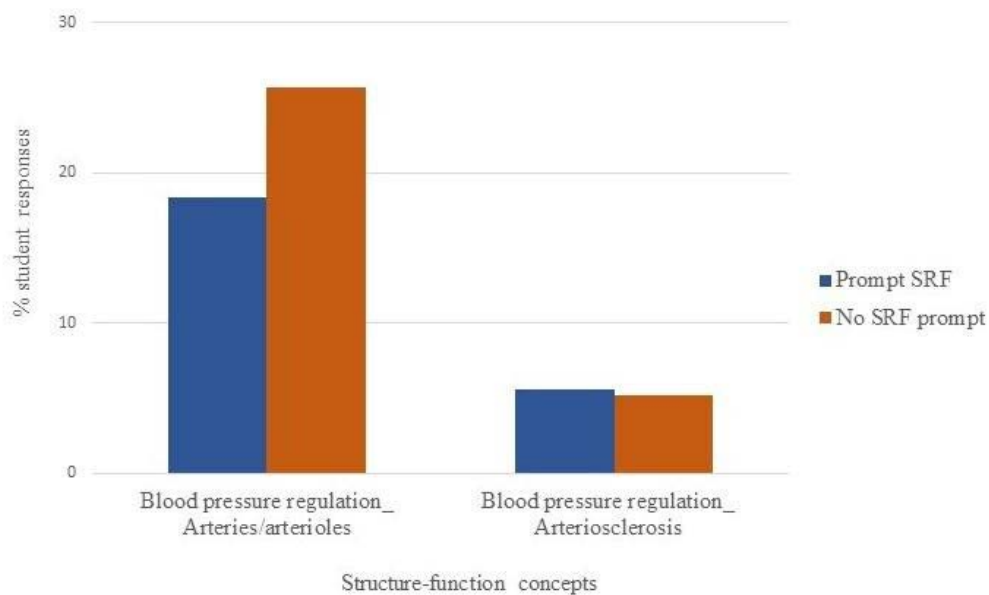
**Figure 4.4.** Percentage of student responses from prompt SRF and no SRF prompt that included small intestine structure-function concepts.

### Question 3: Cardiovascular System/Blood Vessels/Arteries and Arterioles

The blood vessels short answer questions included the arteries/arterioles question (Table 4.1). There was one SRF concept in the arteries and arterioles question: *blood pressure regulation* (Appendix, Table A.5). There was not a significant difference in student performance between the two question prompts the SRF concept ( $\chi^2_{(df=1, N=199)}=1.491, p=0.222$ ) (Fig. 4.5).

#### Question 4: Cardiovascular System/Blood Vessels/Arteriosclerosis

The blood vessels short answer questions included the arteriosclerosis question (Table 4.1). There was one SRF concept in the arteriosclerosis question: *blood pressure regulation* (Appendix, Table A.5). There was not a significant difference in student performance between the two question prompts the SRF concept ( $\chi^2_{(df=1, N=315)}=0.086, p=0.958$ )(Fig. 4.5).



**Figure 4.5.** Percentage of student responses from Arteries/arterioles and Arteriosclerosis questions with prompt SRF and no SRF prompt that included blood pressure regulation structure-function concepts.

#### Student Interviews Related to Guiding Context

Students were asked for their feedback on the question prompts. Some students found the wording of the structure-function relationship confusing (i.e., “Based on form reflecting function”). For example, the following student first answered the “celiac disease” question and then provided feedback on the question prompt. When providing feedback on the question prompt, she stated that she did not understand the word “reflects”:

I think when trying to figure out what you mean by relating ... reflecting function, that's kind of confusing, but generally it's like ... you know that you have to relate the structure to the function. I think it probably would be better saying based on damage to structure ... no, I don't know. I'm so bad at these. Maybe just instead of saying form just say structure. That doesn't sound good either. I don't know if that's confusing or if it's just hard to like ... I guess if it's hard to understand, it's technically confusing. (Ellen in response to “celiac disease” question)

The student was then asked to answer the “blood pressure” question: “Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation.”

I just ... I think personally I don't like the word reflecting, but I don't know what else to use there. I don't know. I think that reflecting ... I think it's the hard C in it. I don't know. I'm weird. I think it's just like the hard reflecting function since they're both really hard next to each other. I think it's just an off putting. (Ellen in response to “arteries/arterioles” question)

Following up on this student's response, the interviewer modified the question prompt by removing the structure-function relationship (i.e., “Based on form reflecting function”) and asked the student the same question but without the guiding context: “Arteries and arterioles are important in blood pressure regulation. Explain how the structure of these blood vessels contributes to blood pressure regulation.” The student responded as follows:

I think that's better because it also shortens it by a couple words. I think that helps like a jumpstart ... it's not just like a ... it's obviously not a multiple choice, but it's like you know that you already have to start explaining something and delving into what you know. I think that helps to get you ready for it. (Ellen in response to “arteries/arterioles” question with no SRF prompt)

Another student also mentioned they had difficulty with “the principle form reflects function” in the question prompts:

Yes, because I'm like wait, describe the principle. I'm like wait, what's the principle? Then reflecting function, really. It's just a little thing that tripped me up. Then I had to think okay form reflects function. I had to pick that apart and be like okay what does

form mean? What does function mean? Then I had to put it together and be like what do they mean together? (Agatha providing general feedback about question prompts)

However, this student also provided feedback on the question prompts in general and the structure-function relationship (“Based on form reflecting function”) in the prompts:

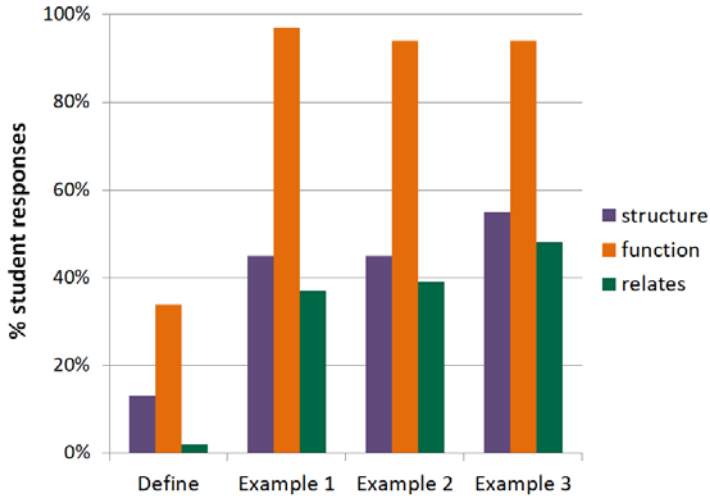
I think it's pretty clear because I think they both reflect one another, that's why it's a reflection. It's like they're both looking, if it were a mirror, function would be looking into form, form would be looking into function. Yeah because I would say form definitely reflects function. It affects how something is performed. But I also think how something is performed is a reflection in itself of the structure. So I don't think that's confusing because I think of reflection with a mirror. (Mary providing general feedback about the structure-function relationship in the question prompts)

### Question Sequencing

#### Human Scoring

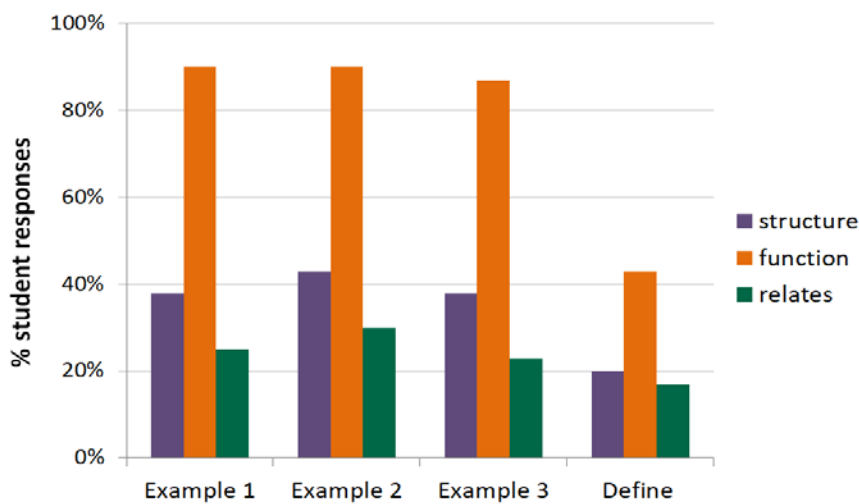
Human scoring of the responses revealed the percentage of students who used structure, function or related structure and function in their responses for each question version. When asked to define the core principle structure and function first (format DX), 13% of students identified structures, 34% identified functions, and 2% of students were able to link the two concepts (Figure 4.6). Students were asked to provide a total of three examples of the core principle. The identification of structures and functions were similar for the three examples, while relating structure and function increased from the first to the third example. By the third example, 48% of students related structure and function in their responses. Overall, students mentioned functions in their responses more often than structures. Within the examples, almost 100% of the student responses discussed functions.





**Figure 4.6.** Human scoring of student responses to format DX. N=62 (from Carter & Prevost, 2018).

When asked to define the core concept secondarily, after providing examples (format XD), 20% of students identified structures, 43% identified functions, and 17% of students were able to link structure and function in their definition (Figure 4.7). When asked to provide examples first, before giving a definition, less than 30% of students accurately related structure and function in any one of the three examples.



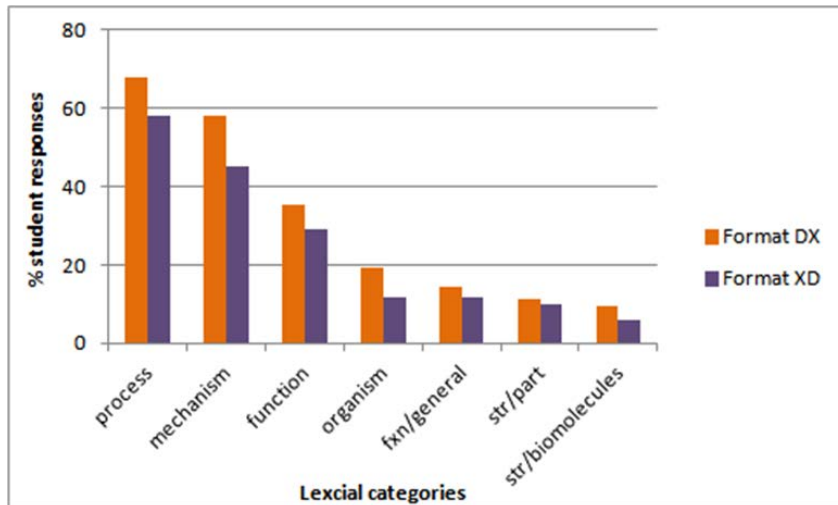
**Figure 4.7.** Human scoring of student responses to format XD. N=69 (from Carter & Prevost, 2018).

### Response Length

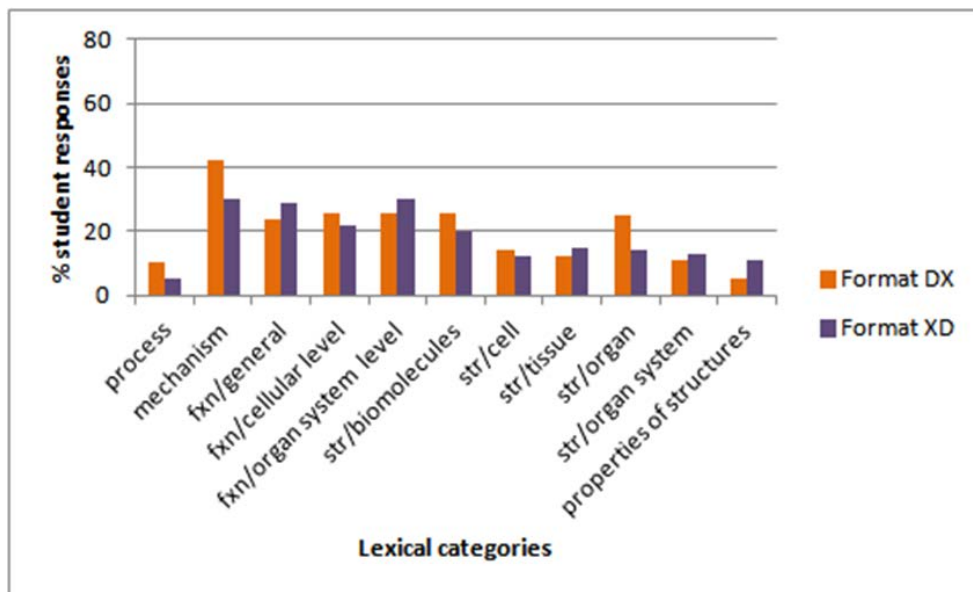
Student response length varied from one word to a short paragraph (102 words). There was not a significant difference in response length between question formats for define, give example 1, or give example 2. For the third example, response length was greater for format DX (median= 16.06) than for format XD give example 3 (median=15.69) (Mann-Whitney test,  $U=1717.00$ ,  $p=.05$ ,  $d= 0.019$ ) with an extremely small effect size (Cohen, 1988; Carter & Prevost, 2018).

### Lexical Analysis

Lexical analysis of the students' written responses produced 23 lexical categories (Table 2.4 from Chapter 2, and Carter & Prevost, 2018). I compared lexical categories between the two question formats. For the question "Define the principle: form reflects function", I identified 10 categories in student responses to the format DX, and 13 in the format XD responses. Figure 4.8 shows the seven most commonly used categories in student responses. These seven categories were found in more than 10% of student responses. For the question "Give an example of the principle form reflects function", both the format DX and format XD responses contained 20 categories, although only 11 categories are shown. These 11 categories were found in more than 10% of student responses (Fig. 4.9). Chi-square analysis of the lexical categories for the "Define" question between the two question formats demonstrated no significant difference between the question formats ( $X^2_{\text{Define}} (df=14, N=273)=13.61$ ,  $p=.479$ ; Fig. 4.8). Similarly, chi-square analysis of the lexical categories for the "Give example" question also demonstrated no significant difference ( $X^2_{\text{Give Example}} (df=19, N=953)=28.89$ ,  $p=.068$ ; Fig. 4.9).



**Figure 4.8.** Lexical categories contained in student responses to “Define the principle form reflects function”. Only categories found in more than 10% of the student responses are shown. (fxn=function, str=structure). Format DX N=62, format XD N=69 (from Carter & Prevost, 2018).



**Figure 4.9.** Lexical categories contained in student responses to “Give an example of the principle form reflects function”. Only categories found in more than 10% of the student responses are shown. (fxn=function, str=structure). Format DX N=62, format XD N=69 (from Carter & Prevost, 2018).

A Fisher’s Exact Test analysis of format DX revealed a significant difference in the number of structure and function lexical categories used in student responses when comparing

“Define” and “Give Example 1”, “Define” and “Give Example 2” and “Define” and “Give Example 3” ( $p < .05$ ) (Tables 4.6 and 4.7). However, there was not a significant difference in the structure and function lexical categories among the examples.

A Fisher’s Exact Test analysis of format XD revealed a significant difference in the structure and function lexical categories between “Define” and “Give Example 1”, “Define” and “Give Example 2” and “Define” and “Give Example 3” ( $p < 0.05$ ) (Tables 4.8 and 4.9). However, there was not a significant difference in the structure and function lexical categories among the examples.

**Table 4.6.** Format DX structure lexical categories with frequency by question prompt and Fisher’s Exact Test results comparing structure lexical categories by question prompt.

A. Format DX Structure lexical categories	Prompt/Frequency			
	Define	Give Example 1	Give Example 2	Give Example 3
<i>structure</i>	0	0	0	0
<i>biomolecules</i>	6	15	15	18
<i>cell</i>	1	6	7	13
<i>cell components</i>	0	1	2	1
<i>complex structures</i>	0	4	2	0
<i>organ</i>	0	14	16	17
<i>organ components</i>	0	0	1	6
<i>organ system</i>	0	8	6	6
<i>tissue</i>	0	12	4	6
<i>tissue components</i>	0	1	1	1
<i>part</i>	7	0	0	0
B. Format DX				
Structure lexical categories	Fisher’s Exact Test	Significance ( $p < .05$ )		
Define: Give Example 1	31.394	0.000		
Define: Give Example 2	28.499	0.000		
Define: Give Example 3	31.126	0.000		
Example 1: Example 2	6.367	NS		
Example 2: Example 3	6.744	NS		

**Table 4.7.** Format DX function lexical categories with frequency by question prompt and Fisher's Exact Test results comparing function lexical categories by question prompt.

A. Format DX <b>Function</b> lexical categories	Prompt/Frequency			
	Define	Give Example 1	Give Example 2	Give Example 3
<i>function</i>	22	1	4	2
<i>cellular level</i>	3	15	16	17
<i>disorders</i>	0	1	3	5
<i>general</i>	9	14	11	19
<i>organ level</i>	0	3	7	5
<i>organ system level</i>	0	27	12	9
<i>organism level</i>	1	1	0	1
<b>B. Format DX</b>				
<b>Function</b> lexical categories	Fisher's Exact Test	Significance (p<.05)		
Define: Give Example 1	61.710	0.000		
Define: Give Example 2	43.650	0.000		
Define: Give Example 3	46.637	0.000		
Example 1: Example 2	10.579	NS		
Example 2: Example 3	4.818	NS		

**Table 4.8.** Format XD structure lexical categories with frequency by question prompt and Fisher's Exact Test results comparing structure lexical categories by question prompt.

A. Format XD <b>Structure</b> lexical categories	Prompt/Frequency			
	Define	Give Example 1	Give Example 2	Give Example 3
<i>properties of structures</i>	0	9	8	6
<i>structure</i>	2	0	0	0
<i>biomolecules</i>	4	16	9	17
<i>cell</i>	2	6	8	11
<i>cell components</i>	0	2	4	4
<i>complex structures</i>	0	2	2	2
<i>organ</i>	0	9	12	9
<i>organ components</i>	0	3	3	1
<i>organ system</i>	0	5	13	9
<i>tissue</i>	1	6	8	9
<i>tissue components</i>	0	2	2	1
<i>part</i>	7	0	0	0
<b>B. Format XD</b>				
<b>Structure</b> lexical categories	Fisher's Exact Test	Significance (p<.05)		
Define: Give Example 1	31.394	0.000		
Define: Give Example 2	28.499	0.000		
Define: Give Example 3	31.126	0.000		
Example 1: Example 2	6.367	NS		
Example 2: Example 3	6.744	NS		

**Table 4.9.** Format XD function lexical categories with frequency by question prompt and Fisher's Exact Test results comparing function lexical categories by question prompt.

A. Format XD Function lexical categories	Prompt/Frequency			
	Define	Give Example 1	Give Example 2	Give Example 3
<i>function</i>	20	0	0	1
<i>cellular level</i>	0	19	14	12
<i>disorders</i>	1	2	1	5
<i>general</i>	8	18	19	22
<i>organ level</i>	0	4	6	4
<i>organ system level</i>	3	28	18	17
<i>organism level</i>	0	0	3	0

B. Format XD Function lexical categories	Fisher's Exact Test	Significance (p<.05)
Define: Give Example 1	63.200	0.000
Define: Give Example 2	56.722	0.000
Define: Give Example 3	47.608	0.000
Example 1: Example 2	5.654	NS
Example 2: Example 3	6.881	NS

## Student Interviews

Four students were interviewed for their interpretation of the questions and to provide a definition of structure and function along with examples. Student responses were coded in the same manner as written responses for structure, function, and structure relates to function. One student related structure and function in her definition and examples for both verbal and written responses. Another student identified structures only when prompted to define the core concept in both written and verbal responses but related structure and function in examples of the core concept in both written and verbal responses. The other two students identified only structures in their written responses to the definition question yet during the interview related structure and function in their responses to both the definition and give examples questions. The students also used similar lexical categories in their verbal and written responses. For the definition question, student responses were assigned to the same structure, structure/organ, and function categories in both verbal and written responses. For the example question, three of the four student responses

were assigned to the same categories: structure/organ, function/organ, structure/tissue, function/tissue.

During interviews, I examined students' use of structures and functions within the levels of organizations. When asked which level of organization they typically thought of as examples of structures and functions, three out of four students interviewed identified the organ level. The interviewer then asked why they thought of organs as examples of structures and functions. Students described the tangible nature of organs and being able to identify a clear purpose as reasons for focusing at the organ level, as exemplified in the following student responses

Because, I feel like they [organs] incorporate enough of tissues and cells. I feel like that's the level that you get to where you're actually, you have something that has a purpose. Obviously tissues have a purpose too, but I feel like tissues just build up organs, that's what their main thing, whereas like your organ does something. - student 1

Because they (organs) are the ones I remember the most, are bigger things, and not the smaller things, 'cause I can't wrap my mind around the smaller things. I can see the bigger things clearly, and I can dissect it better than I could do with the smaller things- student 2

I think it's just easier for me to find an example when it's something that I can visibly see, or I've seen before, rather than going to the atomic level, or the molecular level- student 3

The fourth student thought of cells when thinking of examples and mentioned red blood cells as an example. The student then described their thinking about cells.

Because a lot of cells in our bodies have organelles and they perform a lot of different functions. They provide a broad spec of functions and they ... A cell is like, reproduce, divide, busy, busy bodies. I feel like my head's, a million thoughts is always going through it. So I guess I would consider myself ... A cell is equipped to do certain functions. So I guess I make sure I'm equipped with certain information or I have to know a certain amount of information on a topic. Honestly, there's a picture of a little cell in my head that's running around and it's talking to little organs and organ system. -student 4

## Discussion

In this study, human scoring and computer-assisted scoring were used to examine how varying features of question prompts may affect student writing about the structure-function

relationship (SRF) in anatomy and physiology. This study was designed 1) to assess student understanding of the structure-function relationship when answering questions at different cognitive levels, 2) to examine if the presence or absence of the core concept structure-function in the question prompt influences student explanations of the core concept, and 3) to determine if varying the order of questions from different cognitive levels affects student performance on structure-function short answer questions.

Overall research question:

*1. How does varying the features of short answer questions affect student explanations about the structure-function relationship in anatomy and physiology?*

Overall research hypothesis:

*1. I hypothesize that varying the features (cognitive level, guiding context and question order) of short answer questions may affect student explanations.*

I have demonstrated that question features can influence student explanations of the structure-function relationship. I will discuss cognitive level, guiding context and question order separately.

Cognitive Level

Research Question:

*2. Do student responses to understand and apply level question reveal differences in their conceptual understanding of structure and function?*

Research Hypothesis:

*2. I hypothesize that there is difference in conceptual understanding based on the cognitive level of the question prompts.*



In this study, students show varied responses when prompted to demonstrate conceptual understanding of SRF with questions from different cognitive levels. There is a difference in conceptual understanding of SRF based on the cognitive level of the question prompt for six of the seven SRF concepts. However, there is no clear pattern of the cognitive level of the question prompt affecting student explanations. Students demonstrated more conceptual understanding of four of the SRF concepts when answering the understand questions and more conceptual understanding of two SRF concepts when answering the apply questions. The questions at the apply level provided a different context, which may have influenced student explanations.

My results show that context affects student explanations. These results align with prior studies that found that when students are asked to apply their knowledge in different contexts, they may produce varying explanations. In a study of thermal equilibrium and the transfer of heat energy between objects of different temperatures, Clark (2006) found that the context of the question influenced students' responses. Students were asked about thermal equilibrium in the context of wood, metal, and glass in a refrigerator, in an oven, and in a hot trunk and if the objects become the same temperature or remain different. The varying contexts elicited different types of student responses, both correct and incorrect. Nehm & Ha (2011) explored open-ended evolution questions and found that students' use of core concepts of natural selection differed significantly in relation to hierarchical level within-species or between-species and trait gain or trait loss. Although context affected student explanations, there was not a consistent pattern among the varying contexts. They found that explanations of trait loss included a greater number of naive ideas while explanations of trait gain included a greater number of evolutionary key concepts. Additionally, they found that within-species contexts demonstrated explanations with more natural selection concepts, and between-species contexts had less concepts and more naive

ideas (Nehm & Ha, 2011). In an evaluation of student conceptions about pressure, heat, and evolution across different contexts, Clough and Driver (1986) found that context influences performance with context affecting student explanations although there was not a consistent pattern among the different contexts. For example, students included more correct ideas in response to a question about pressure on a submarine and how that pressure changes with varying depth. When asked about pressure differences on the submarine, such as pressure across or pressure downwards, students had less correct ideas (Clough & Driver, 1986). Similarly, my results with the “understand” and “apply” level questions, which provide varying contexts in the question prompts, corroborate these findings: context affects student explanations.

Furthermore, it may not be only the context that affects student explanations, but whether the context is familiar or unfamiliar to the student. In a genetics education study, students mentioned more lexical categories in a question about genetic variation in animals (familiar organism) compared to bacteria (unfamiliar organism) (Prevost et al., 2013). Similarly, in a study of student conceptions of natural selection, students identified more natural selection core concepts in questions about animals (familiar) compared to plants (less familiar) (Heredia et al., 2016). In this study, familiarity of the question context may also have influenced student response. Students identified different SRF concepts when responding to an understand question than when responding to the apply questions. In addition, students mentioned more SRF concepts when responding to the understanding questions “two layers of skin” and “contractile proteins” compared to the apply questions “third degree burn” and “rigor mortis” (Figs. 1 and 2). However, students mentioned one SRF concept more than the others in response to each of the apply questions. The context of these apply question prompts may have been familiar to students. Using a familiar context like burns or rigor mortis may cue students to mention specific SRF

concepts while omitting other relevant concepts. The context of the question prompts between the understand and apply questions seems to influence student explanations.

Guiding Context

Research Question:

- 3. How do student descriptions of the structure-function relationship differ when answering a question prompt with reference to the core concept compared to students answering a question prompt without the reference?*

Research Hypothesis:

- 3. I hypothesize that there is a difference in conceptual understanding based on the reference to the core concept in the question prompt.*

This study found no difference in conceptual understanding of the structure-function relationship with and without the use of a guiding context in the wording of the question prompt. This finding is contrary to previous studies which found guiding context useful. For example, Clough and Driver (1986) found that when asked about pressure differences on a submarine and “use the idea of atmospheric pressure to explain your response”, student responses included more correct ideas than when this phrase was excluded from the prompt (Clough & Driver, 1986). In a more recent study of students’ evolutionary explanations, Kampouris and Zogra (2008) asked students questions in which different types of information were provided for the student to base their explanations of differential survival and trait maintenance through reproduction. In one question (“task 3”), students were given no information about the initial state of the evolutionary process while in another question (“task 4”), they were given details in the question prompt about intraspecific variation and natural selection. Students provided more evolutionary explanations to task 4 and more teleological explanations to task 3 (Kampouris & Zogra, 2008).

In this study, the guiding prompt that stated the concept “form reflects function” may not have improved student performance because they were unable to interpret the prompt. This occurrence may be due to students not understanding the concept, or the terms “structure/form” and “function” that comprise the conceptual framework, which is necessary for conceptual understanding of the structure-function relationship (McFarland et al., 2016; Michael et al., 2017). As described in student interviews from chapter 2, most students were able to define function but had difficulty with structure. Students must understand both terms to fully comprehend the link between structure and function. Therefore, providing the structure-function relationship in the question prompt may not be helpful in eliciting student explanations. Guiding context in the question prompt may be more helpful if students have a firm grasp of the concept.

#### Question Sequencing

##### Research Question:

4. *How does varying the order of questions from different cognitive levels affect student explanations of the structure-function relationship?*

##### Research Hypothesis:

4. *I hypothesize that there is a difference in conceptual understanding between the question orders.*

To address my research question of whether question sequence matters for formative assessment, half of the students answered a “remember” (define) question first followed by three “give example” (understand) questions (format DX). Only 2% of the students in format DX related structure and function in their definition, while 48% related structure and function in their third example. The other half of the students answered three “give example” (understand) questions followed by a “remember” (define) question (format XD). Of the group that defined

the core concept last, 17% of students related structure and function in their definition, and 23% related structure and function in their third example. Students performed better (related structure to function) on the definition question when it followed the example questions and on the example questions when they followed the definition question. The performance of students in this population may be related to the implicit nature of the core concept within this General Physiology course. The concept is explicitly taught in the prerequisite anatomy and physiology courses, but our results suggest that few students transfer their understanding of the concept in this study. Transference is learning a concept in one context and applying it in another (Duit, 1991). The minimal transference observed in this study suggests the need for a curriculum that explicitly helps students to recognize the importance of this concept (Michael, et al., 2009), and to use the concept as a way to build connections between prior knowledge and newly introduced ideas (Carter & Prevost, 2018).

My results show that students performed better on the last questions in the sequence, regardless of whether the last question was easier or more difficult (Carter & Prevost, 2018). These results are in contrast to prior research. For example, in a sample of 103 veterinary science students in a timed exam, difficult multiple-choice items early in examinations were correlated with decreased performance compared with students who had easier items first (Marks & Cronje, 2008). When the difficult questions were first, students ran out of time before they reached the easier questions. The question order under a timed condition may lead to fatigue as well as priming. Positing the more difficult questions first may cause conceptual priming, but because the students are in a timed condition, they may not have the opportunity to demonstrate their knowledge among the easier questions. However, Huck and Bowers (1972) found no difference

in performance when varying multiple choice item sequences when students are in a power condition and have unlimited time to finish the assessment.

In this study, students in the format DX group performed better on the examples than students in the XD group, while students in the XD group performed better on the definition than did students in the DX group. This finding suggests that for the DX group, the definition acted as a prime for the example question, and for the XD group, the examples acted as a prime for the definition question. In each case, students had greater success in retrieving or applying the core concept after priming. Interestingly, the three “give example” questions in each format do not seem to act as a prime for each other (Figs. 6 and 7). However, previous studies have shown that when the cognitive task is similar, there is less likely to be priming (Strack, 1992). Therefore, the three example questions would not act as a prime for each other.

#### Implications for Teaching

Question features affected student explanations, especially cognitive level and question sequencing. Formative written assessment as part of instruction with varying question features allows for an examination of the depth of student understanding of the structure-function relationship. Thus, formative assessment tasks should be at different cognitive levels and in varying contexts to facilitate learning the core concept. Students may be able to demonstrate conceptual understanding in one context but not another. In questions with a familiar context, students may exhibit more conceptual understanding. However, this may not be an accurate reflection of their understanding. Providing students with questions in varying contexts and cognitive levels will allow students to demonstrate their heterogeneous ideas about a concept.

The purpose of formative assessment is to provide feedback to both the instructor and student. However, formative assessment must align with curriculum and instruction if it is to

support learning (NRC, 2001). Assessment, curriculum, and instruction should be at similar cognitive levels and provide varying contexts. For example, if instruction is at the remember cognitive level, it is unfair to expect students to understand or apply the information in a new context. The results from formative written assessment can not only enhance learning via feedback to instructors and students, but it may provide feedback to instructors about alignment of curriculum to learning.

## Conclusion

In summary, question features can influence student explanations of the structure-function relationship. There was no clear pattern of the cognitive level of the question prompt affecting student explanations. I found no difference in conceptual understanding of the structure-function relationship with and without the use of a reference to the core concept in the wording of the question prompt. My results show that students performed better on the last questions in the sequence, regardless of whether the last question was easier or more difficult, which may indicate conceptual priming. The context of a question prompt may influence students' explanations.

## References

- Anderson L.W., Krathwohl D.R., Bloom B.S., Bloom B.S. (2001). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman.
- Bell, B. & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85, 536-553.
- Bloom, B. S. (1956). Taxonomy of Educational Objectives: The Classification of Education Goals. Cognitive Domain. Handbook 1. Longman.
- Bowen, C. W. (1998). Item design considerations for computer-based testing of student learning in chemistry. *Journal of Chemical Education*, 75(9), 1172-1175

- Brenner, M.H. (1964). Test difficulty, reliability and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 48, 98-100.
- Carter, K.P. & Prevost, L.B. (2018) Question order and student understanding of structure and function. *Advances in Physiology Education*, 42(4), 576-585.
- Chi, M.T., Feltovich, P.J., & Glasner, R. (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152
- Clark, D.B. (2006). Longitudinal Conceptual Change in Students' Understanding of Thermal Equilibrium: An Examination of the Process of Conceptual Restructuring. *Cognition and Instruction*, 24(4), 467-563.
- Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education*, 70, 473-496.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd ed., Hillsdale, NJ:Lawrence Erlbaum.
- Crowe, A., Dirks, C., & Wenderoth, M.P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education*, 7, 368-381.
- Dietterich, TG. (2000). Ensemble methods in machine learning. *In Multiple Classifier Systems Ensemble Methods in Machine Learning*, Springer.
- Duit, R. 1991. On the role of analogies and metaphors in learning science. *Science Education*, 75(6), 649-672
- Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a Constructed-Response Instrument to Explore the Effects of Item Position and Item Features on the Assessment of Students' Written Scientific Explanations. *Research In Science Education*, 45, 527-553.
- Gentner, D. & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.
- Glynn, S.M. & Muth, K.D. (1994). Reading and writing to learn science: Achieving scientific literacy. *Journal of Research in Science Teaching*, 31(9), 1057-1073.
- Goldstein, E.B. (2011). Cognitive Psychology: Connecting Mind, Research and Everyday Experience. Wadsworth Cengage: Belmont, CA.
- Haudek, K.C., Prevost, L. B., Moscarella, R.A., Merrill, J. & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE-Life Sciences Education*, 11, 283-293.



Heredia, S. C., Furtak, E. M., & Morrison, D. (2016). Exploring the influence of plant and animal item contexts on student response patterns to natural selection multiple choice items. *Evolution: Education and Outreach* 9, 1-11.

Huck, S.W. & Bowers, N.D. (1972). Item difficulty and sequence effects in multiple choice achievement tests. *Journal of Educational Measurement*, 9, 105-111.

Kampourakis, K., & Zogza, V. (2009). Preliminary Evolutionary Explanations: A Basic Framework for Conceptual Change and Explanatory Coherence in Evolution. *Science And Education*, (10), 1313-1340.

Leary L.F. & Dorans N.J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research* 55(3), 387-413.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50-60.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Belmont, CA: Wadsworth Publishing Company.

Marks, A. M., & Cronje, J. C. (2008). Randomized Items in Computer-based Tests: Russian Roulette in Assessment? *Educational Technology & Society*, 11(4), 41–50.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.

Mayer, R. E. (1992). Cognition and instruction: their historic meeting within educational psychology. *Journal of Educational Psychology*, 4, 405-412.

McFarland, J., Wenderoth, M. P., Michael, J., Cliff, W., Wright, A., & Modell, H. (2016). A conceptual framework for homeostasis: development and validation. *Advances in Physiology Education*, 40(2), 213–222.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157.

Michael, J., Modell, H. McFarland, J. & Cliff, W. (2009) The “core principles” of physiology: what should students understand? *Advances in Physiology Education*, 33, 10-16.

Michael, J., Martinkova, P., McFarland, J., Wright, A., Cliff, W., Modell, H., and M. P. Wenderoth. (2017). Validating a conceptual framework for the core concept of “cell-cell communication”. *Advances in Physiology Education* 41(2), 260-265

Nehm, R.H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48, 237-256.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive Foundations for Science Assessment Design: Knowing What Students Know about Evolution. *Journal of Research In Science Teaching*, 49(6), 744-777.

Prevost, L. B., Knight, J.K., Smith, M.K., & Urban-Lurain, M. (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico, held April 2013.

Schwarz, N. & Strack, F. (1991). Context effects in attitude surveys: Applying cognitive theory to social research. *European Review of Social Psychology* 2(1):31-50.

Semsar, K., & Casagrand, J. (2017). Bloom's dichotomous key: a new tool for evaluating the cognitive difficulty of assessments. *Advances in Physiology Education*, 41(1), 170-177.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-584.

Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A., Hough, C.M., Nickell, A.E., Peer, A.G., Coley, J.D. & Tanner, K. D. (2013). Development of the Biology Card Sorting Task to Measure Conceptual Expertise in Biology. *CBE - Life Sciences Education*, 12(4), 628-644.

SPSS IBM Modeler (2013). IBM SPSS Modeler v15, IBM corporation.

Strack, F. (1992). "Order effects" in survey research: Activation and information functions of preceding questions. In N. Schwarz & S. Sudman (Eds), *Context Effects in Social and Psychological Research*. New York: Springer.

Weston, J., Haudek, K.C., Prevost, L., Urban-Lurain, M. & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed response questions on photosynthesis. *CBE-Life Sciences Education*, 14, 1-12

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.

## APPENDICES

## Appendix A: Tables

**Table A.1.** Description of conceptual rubric for each short answer question prompt.

Question prompt	Conceptual rubric	Description
Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.	Structures protection	Pigments, cells, glands and tissues that provide protection.
	Function protection	Protective barrier
	Structures regulation	Cells, glands and tissues that regulate temperature, blood supply and cell division
	Function regulation	Homeostasis, thermoregulation, repair, and regeneration
	Structures sensation	Cells and tissues which provide sensation
	Function sensation	Sense of touch and sensory perception.
Victims of third degree, or full thickness, burns have their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin.	Structures protection	Pigments, cells, glands and tissues that provide protection.
	Function protection	Protective barrier
	Structures regulation	Cells, glands and tissues that regulate temperature, blood supply and cell division
	Function regulation	Homeostasis, thermoregulation, repair, and regeneration
	Structures sensation	Cells and tissues which provide sensation
	Function sensation	Sense of touch and sensory perception.

**Table A.1 (Continued)** Description of conceptual rubric for each short answer question prompt.

The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction.	ATP necessary for contraction to end	ATP required for myosin heads to detach from actin.
	Myosin binds to actin	During cross bridge formation myosin heads bind to active binding sites on actin.
	Muscle contracts due to calcium, troponin and tropomyosin	Calcium released from SR and attaches to troponin, causes tropomyosin to move from active binding sites
	Sarcomere is contractile unit	Contraction occurs in sarcomeres.
	Specific structure of myosin	Myosin is thick filaments with myosin heads
	Specific structure of actin	Actin is thin filaments with active binding sites.
	Muscle shortening	Muscle shortens during contraction, myosin pulls on actin to shorten the muscle.
A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis.	ATP no longer available	After death ATP is no longer available.
	ATP necessary for contraction to end	ATP required for myosin heads to detach from actin.
	Myosin binds to actin	During cross bridge formation myosin heads bind to active binding sites on actin.
	Muscle contracts due to calcium, troponin and tropomyosin	Calcium released from SR and attaches to troponin, causes tropomyosin to move from active binding sites
	Sliding filament model of contraction	Either mentions “sliding filament model’ or actin and myosin slide past each other.
	Sarcomere is contractile unit	Contraction occurs in sarcomeres.

**Table A.1 (Continued)** Description of conceptual rubric for each short answer question prompt.

Consider the mucosa of the small intestine. Based on form reflecting function, explain how this layer contributes to the functions of the digestive system.	structures of mucosa absorption	The cells and tissues which are responsible for absorption.
	increase surface area	The mucosa increases surface area for absorption.
	function of mucosa absorption	The mucosa is responsible for absorption of nutrients.
	structures of mucosa digestion	The enzymes which break down food.
	function of mucosa digestion	The mucosa breaks down food into nutrients.
	structures of mucosa secretion	The cells and tissues which secrete digestive enzymes, hormones or mucus.
	function of mucosa secretion	The mucosa secretes digestive enzymes, hormones and mucus.
	structures of mucosa protection	The cells and tissues responsible for protection.
	function of mucosa protection	The mucosa provides a protective barrier and reduces friction.
	propulsion (misconception)	The mucosa propels the food bolus. (Should be the muscularis propels food bolus).
Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system.	Villi structure	The villi change shape or flatten.
	Change in surface area	There is a decrease in surface area when the villi flatten.
	Decrease absorption	There is a decrease in absorption of nutrients.
	Change in digestive enzymes	There is a decrease in digestive enzymes.
	Decrease digestion	There is a decrease in the ability to break down food into nutrients.
	Propulsion (misconception)	The mucosa propels the food bolus. (Should be the muscularis propels food bolus).
.		

**Table A.1 (Continued)** Description of conceptual rubric for each short answer question prompt.

Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation	smooth muscle in wall	The arterial wall is made of smooth muscle.
	vasoconstriction/vaso dilation	The smooth muscle contracts or relaxes.
	elasticity/flexibility of arterial wall	Arterial wall has the ability to stretch or change shape, and then returns to original shape.
	relationship between resistance/flow/pressure	Mentions 2 out of 3. Mentions there is a relationship between them.
Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.	smooth muscle in wall	The arterial wall is made of smooth muscle.
	regulation of blood pressure	The smooth muscle contracts or relaxes to regulate blood pressure.
	elasticity/flexibility of arterial wall	Arterial wall has the ability to stretch or change shape, and then returns to original shape.
	relationship between resistance/flow/pressure	Mentions 2 out of 3. Mentions there is a relationship between them.

**Table A.2.** Metrics of model performance for each conceptual model.

Question prompt	Conceptual model	Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning. training n=418, testing n=179	Structures protection	0.773	0.836	0.687	0.928	0.935
	Function protection	0.967	0.885	0.721	0.974	0.977
	Structures regulation	0.923	0.841	0.805	0.919	0.923
	Function regulation	0.724	0.735	0.676	0.887	0.858
	Structures sensation	0.903	0.787	0.712	0.941	0.871
	Function sensation	0.945	0.773	0.837	0.902	0.754
		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Victims of third degree, or full thickness, burns have their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin. Training n=425, testing n=182	Structures protection	0.944	0.593	0.536	0.905	0.535
	Function protection	0.972	0.793	0.912	0.925	0.869
	Structures regulation	0.955	0.860	0.835	0.926	0.808
	Function regulation	0.858	0.660	0.639	0.842	0.801
	Structures sensation	0.958	0.851	0.904	0.957	0.935
	Function sensation	0.876	0.794	0.809	0.917	0.903



**Table A.2 (Continued)** Metrics of model performance for each conceptual model.

		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction. training n=506, testing n=217	ATP necessary for contraction to end	0.959	0.602	0.706	0.950	0.452
	Myosin binds to actin	0.972	0.778	0.785	0.915	0.902
	Muscle contracts due to calcium, troponin and tropomyosin	1.000	0.915	0.890	0.959	0.877
	Sarcomere is contractile unit	0.976	0.939	0.963	0.979	0.906
	Specific structure of myosin	0.985	0.913	0.834	0.950	0.958
	Specific structure of actin	0.892	0.809	0.842	0.885	0.842
	Muscle shortening	0.907	0.765	0.725	0.881	0.755

**Table A.2 (Continued)** Metrics of model performance for each conceptual model.

		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the process of rigor mortis. Training n=680, testing n=292	ATP no longer available	0.936	0.865	0.931	0.924	0.965
	ATP necessary for contraction to end	0.867	0.829	0.937	0.876	0.931
	Myosin binds to actin	0.874	0.705	0.887	0.870	0.825
	Muscle contracts due to calcium, troponin and tropomyosin	0.893	0.833	0.987	0.893	0.844
	Sliding filament model of contraction	0.946	0.881	1.000	0.953	0.837
	Sarcomere is contractile unit	1.000	0.959	1.000	0.983	0.953

**Table A.2 (Continued)** Metrics of model performance for each conceptual model.

		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Consider the mucosa of the small intestine. Based on form reflecting function, explain how this layer contributes to the functions of the digestive system. training n=220 testing n= 188	increase surface area	1.000	0.959	0.954	0.961	0.986
	structures of mucosa absorption	0.984	0.910	0.935	0.973	0.911
	function of mucosa absorption	0.985	0.907	0.890	0.963	0.987
	structures of mucosa digestion	0.896	0.892	0.946	0.900	0.923
	function of mucosa digestion	0.978	0.427 (n=60)	0.471	0.758	0.417
	structures of mucosa secretion	0.822	0.475 (n=32)	0.687	0.857	0.375
	function of mucosa secretion	0.895	0.923	0.729	0.938	0.953
	structures of mucosa protection	0.921	0.686	0.166	0.875	0.583
	function of mucosa protection	1.000	0.732	0.642	0.941	0.667
	propulsion (misconception)	0.813	0.226 (n=20)	0.000	0.750	0.150

**Table A.2 (Continued)** Metrics of model performance for each conceptual model.

		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system. Training n=504, testing n=216	Villi structure	0.793	0.208 (n=25)	0.000	0.000	0.000
	Change in surface area	0.861	0.764	0.619	0.891	0.722
	Decrease absorption	0.811	0.542	0.486	0.870	0.982
	Change in digestive enzymes	1.000	0.105 (n=17)	0.000	0.500	0.062
	Decrease digestion	0.857	0.398	0.325	0.778	0.326
	Propulsion (misconception)	1.000	0.213 (n=45)	0.009	0.667	0.138
		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation. Training =495, testing =210	smooth muscle in wall	0.934	0.923	0.949	0.970	0.970
	elasticity/flexibility of arterial wall	1.000	0.912	0.926	0.979	0.995
	regulation of blood pressure	0.887	0.834	0.856	0.907	0.932
	relationship between resistance/flow/pressure	0.948	0.570	0.573	0.781	0.817

**Table A.2 (Continued)** Metrics of model performance for each conceptual model.

		Human scoring IRR	Training kappa	Testing kappa	Precision	Recall
Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles. Training n=542, testing n=232	smooth muscle in wall	0.791	0.283 (n=32)	0.417 (n=3)	0.750	0.188
	regulation of blood pressure	0.938	0.793	0.800	0.916	0.736
	elasticity/flexibility of arterial wall	0.945	0.855	0.772	0.962	0.813
	relationship between resistance/flow/pressure	0.850	0.588	0.532	0.793	0.620

**Table A3.** Chi-square analysis of structure-function concepts from institutional comparison, df=1. \* P-value significant <.01.

Structure-function concepts from integument questions: Two layers of skin and third degree burn					
SRF concept #	Structure-function (SRF) concept	% 2 yr	% 4 yr	Chi-square value (df=1)	P value or *
1	Sensation_Two layers of skin	19.0	31.1	11.366	0.001*
	Sensation_Third degree burn	40.2	54.5	9.206	0.002*
2	Protection_Two layers of skin	38.3	59.0	25.338	0.000*
	Protection_Third degree burn	6.7	14.4	6.688	0.010*
3	Regulation_Two layers of skin	29.2	46.9	19.533	0.000*
	Regulation_Third degree burn	12.9	10.6	0.568	0.451

Structure-function concepts from muscle contraction questions: Contractile proteins and rigor mortis					
SRF concept #	Structure-function (SRF) concept	% 2 yr	% 4 yr	Chi-square value	P value and sig*
4	ATP necessary for contraction to end_contractile proteins	4.6	11.0	5.435	0.020
	ATP necessary for contraction to end_rigor mortis	51.8	54.3	0.273	0.601
5	Myosin binds to actin_contractile proteins	51.3	66.1	7.183	0.007*
	Myosin binds to actin_rigor mortis	24.7	49.7	28.399	0.000*
6	Muscle contracts due to calcium_contractile proteins	12.6	32.2	20.421	0.000*
	Muscle contracts due to calcium_rigor mortis	9.4	21.4	12.095	0.001*
7	Sarcomere contractile unit_contractile proteins	16.9	24.6	3.113	0.078
	Sarcomere contractile unit_rigor mortis	2.0	1.2	0.415	0.519

**Table A3 (Continued)** Chi-square analysis of structure-function concepts from institutional comparison, df=1. \* P-value significant <.01.

8	ATP no longer available_rigor mortis	55.3	54.9	0.006	0.938
9	Muscle shortening_contractile proteins	17.6	21.2	0.677	0.411

Structure-function concepts from small intestine questions: Small intestine mucosa and celiac disease					
SRF concept #	Structure-function category	% 2 yr	% 4 yr	Chi-square value	P value and sig *
10	Absorption_SI mucosa	32.3	36.8	0.693	0.405
	Absorption_Celiac disease	14.0	18.7	1.101	0.294
11	Digestion_SI mucosa	5.6	13.2	4.627	0.031
	Digestion_Celiac disease	0	2.2	2.816	0.093
12	Secretion_SI mucosa	7.3	12.1	1.926	0.165
13	Protection_SI mucosa	4.8	5.8	0.132	0.716

Structure-function concepts from blood vessel questions					
SRF concept #	Structure-function category	% 2 yr	% 4 yr	Chi-square value	P value and sig *
14	Blood pressure regulation_Arteries /arterioles	20.9	30.3	3.152	0.076
	Blood pressure regulation_Arterio sclerosis	8.8	17.2	3.397	0.065

**Table A4.** McNemar analysis of cognitive level of structure-function concepts, df=1. \* P-value significant <.01. Fall 2017 N=83

Structure-function concepts from integument questions: Two layers of skin (understand) and third degree burn (apply).					
SRF concept #	Structure-function (SRF) concept	% Understand	% Apply	McNemar value (df=1)	P value or *
1	Sensation	36.1%	60.2%	7.848	0.000*
2	Protection	65.1%	13.3%	34.588	0.000*
3	Regulation	55.4%	13.3%	26.884	0.000*
	SRF sum	73.5%	66.3%	0.694	0.405 NS

Structure-function concepts from muscle contraction question: Contractile proteins (understand) and Rigor mortis (apply)					
SRF concept #	Structure-function (SRF) concept	% Understand	% Apply	McNemar value (df=1)	P value and sig*
4	ATP necessary for contraction to end	9.6%	56.6%	86.260	0.000*
5	Myosin binds to actin	57.1%	53.6%	0.706	0.401 NS
6	Muscle contracts due to calcium	27.1%	15.3%	9.490	0.002*
7	Sarcomere contractile unit	23.5%	3.5%	35.220	0.000*
	SRF sum	70.9%	81.1%	7.521	0.006*



**Table A5.** Chi-square analysis of guiding context of structure-function concepts, df=1. \* P-value significant <.01. P= question prompt refers to structure-function relationship, NP=prompt does not refer to structure-function relationship.

Structure-function concepts from muscle contraction question: Rigor mortis					
SRF concept #	Structure-function (SRF) concept	% P	% NP	Chi-square value	P value and sig*
4	ATP necessary for contraction to end_rigor mortis	64.3	56.3	2.009	0.156
5	Myosin binds to actin_rigor mortis	45.7	43.8	0.117	0.733
6	Muscle contracts due to calcium_rigor mortis	17.1	17.5	0.007	0.935
7	Sarcomere contractile unit_rigor mortis	5.7	3.8	0.648	0.421
8	ATP no longer available_rigor mortis	61.4	59.4	0.132	0.717

Structure-function concepts from small intestine question: Celiac disease					
SRF concept #	Structure-function category	%P	% NP	Chi-square value	P value and sig *
10	Absorption_Celiac disease	11.8	12.4	0.021	0.885
11	Digestion_Celiac disease	10.8	9.9	0.045	0.832

Structure-function concepts from blood vessel questions					
SRF concept #	Structure-function category	% P	% NP	Chi-square value	P value and sig *
14	Blood pressure regulation_Arteries /arterioles	18.4	25.7	1.491	0.222
	Blood pressure regulation_Arterio sclerosis	5.6	5.2	0.086	0.958

**Table A6.** Chi-square analysis of guiding context of structure-function concepts by course, df=1. \* P-value significant <.01. P= question prompt refers to structure-function relationship, NP=prompt does not refer to structure-function relationship. GP=General Physiology, HAP=Human Anatomy and Physiology

Structure-function concepts from muscle contraction question: Rigor mortis <b>GP only</b>					
SRF concept #	Structure-function (SRF) concept	% P	% NP	Chi-square value	P value and sig*
4	ATP necessary for contraction to end_rigor mortis	60.2	50.4	1.934	0.164
5	Myosin binds to actin_rigor mortis	40.81	39.02	0.049	0.825
6	Muscle contracts due to calcium_rigor mortis	12.24	17.07	1.052	0.305
7	Sarcomere contractile unit_rigor mortis	6.12	4.87	0.153	0.696
8	ATP no longer available_rigor mortis	58.16	54.47	0.233	0.630

Structure-function concepts from muscle contraction question: Rigor mortis <b>HAP only</b>					
SRF concept #	Structure-function (SRF) concept	% P	% NP	Chi-square value	P value and sig*
4	ATP necessary for contraction to end_rigor mortis	73.81	73.68	0.000	0.990
5	Myosin binds to actin_rigor mortis	57.14	57.89	0.005	0.946
6	Muscle contracts due to calcium_rigor mortis	28.57	18.42	1.135	0.287
7	Sarcomere contractile unit_rigor mortis	4.76	0	1.865	0.173
8	ATP no longer available_rigor mortis	69.05	73.68	0.209	0.647

Structure-function concepts from small intestine question: Celiac disease <b>GP only</b>					
SRF concept #	Structure-function category	%P	% NP	Chi-square value	P value and sig *
10	Absorption_Celiac disease	17.54	15.87	0.060	0.806
11	Digestion_Celiac disease	8.77	9.52	0.020	0.887

**Table A6 (Continued)** Chi-square analysis of guiding context of structure-function concepts by course, df=1. \* P-value significant <.01. P= question prompt refers to structure-function relationship, NP=prompt does not refer to structure-function relationship. GP=General Physiology, HAP=Human Anatomy and Physiology

Structure-function concepts from small intestine question: Celiac disease <b>HAP only</b>					
SRF concept #	Structure-function category	%P	% NP	Chi-square value	P value and sig *
10	Absorption_Celiac disease	4.44	8.77	0.698	0.404
11	Digestion_Celiac disease	13.33	10.53	0.220	0.639

Structure-function concepts from blood vessel questions <b>GP only</b>					
SRF concept #	Structure-function category	% P	% NP	Chi-square value	P value and sig *
14	Blood pressure regulation_Arteries /arterioles	7.1	18.96	2.828	0.093
14	Blood pressure regulation_Arterio sclerosis	6.18	5.04	0.133	0.715

Structure-function concepts from blood vessel questions <b>HAP only</b>					
SRF concept #	Structure-function category	% P	% NP	Chi-square value	P value and sig *
14	Blood pressure regulation_Arteries /arterioles	28.89	33.33	0.170	0.680
14	Blood pressure regulation_Arterio sclerosis	4.44	5.55	0.063	0.802

## Appendix B: Interview protocol for Anatomy and Physiology assessment

Example Interview Protocol—the interviews are research and are not part of the course requirements

Participants will be asked to answer aloud anatomy and physiology short answer essay questions that were previously given as an online homework assignment.

Example questions

1. Define the principle form reflects function.
2. Give an example of the principle form reflects function from the human body.
3. Consider the two layers of the skin, the dermis and the epidermis. Which structures of these layers contributes to the functions of the integumentary system? Explain your reasoning.
4. A victim of a third degree, or full thickness, burn has their epidermis and dermis damaged. Relate the loss of functions with losing these layers of the skin.
5. The contractile proteins actin and myosin are involved in the sliding filament model of muscle contraction. Based on the structure of actin and myosin describe their role in skeletal muscle contraction.
6. A medical examiner is called to a crime scene to investigate the circumstances of a recent death. The victim is clutching a syringe in one hand and the medical examiner is unable to remove it. Based on form reflecting function, explain the role of actin and myosin in the victim's grip on the syringe and the process of rigor mortis.
7. Your foot transmits the weight of your body to the ground and supports you in an upright position. Explain how your foot demonstrates that structure reflects function.
8. Each of the long bones of the body has a hollow canal running down its length. Based on the principle of complementarity, explain the function of this canal.
9. Consider the mucosa of the small intestine. Based on the principle of complementarity, explain how this layer contributes to the functions of the digestive system.

10. Your patient was recently diagnosed with celiac disease, which is an autoimmune disease in which gluten damages the villi of the small intestine. Based on form reflecting function, relate the damage of villi to the functions of the digestive system.

11. Arteries and arterioles are important in blood pressure regulation. Based on structure reflecting function, explain how the structure of these blood vessels contributes to blood pressure regulation.

12. Mr. Gallagher has been taken to the local emergency room with a complaint of chest pain. Further investigation reveals he has arteriosclerosis, or a hardening of the arterial walls. Relate this diagnosis to the functions of the arteries and arterioles.

The participant will be given a copy of her/his originally submitted (written) answer(s) and will be asked the following questions:

1) What changes, if any, do you notice in your written and verbal answers?

2) How would you define the following terms:

Process

Structure

Function

3) Can you recall your strategy for answering this question on this assignment?

4) What aspects of the course are most helpful in preparing you for this problem?

5) Did you draw on any concepts you learned in other classes?

6) What anatomy and/or physiology courses have you enrolled in/completed?

7) What other high school/college biology courses have you completed?

8) Let's talk about the question itself. Are you having any *problems* with any parts of the question? Are there any parts that are *confusing*?

9) What are the parts of the question that you felt are *most relevant* to your attempts to answer it?

10) Are there any parts that you feel are unnecessary, that is, parts that you consider *irrelevant* to the question that you can simply ignore?

11) Was any of the material on relating structure and function confusing to you?

12) Do you have any other questions or comments on the subject of the relationship between structure and function, and how it applies to the anatomy and physiology concepts you learned in class?

## Appendix C: IRB Approval Letters



RESEARCH INTEGRITY AND COMPLIANCE  
Institutional Review Boards, FWA No. 00001669  
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799  
(813) 974-5638 • FAX (813) 974-7091

November 4, 2016

Kelli Carter  
Integrative Biology  
4202 E. Fowler Ave.  
SCA110  
Tampa, FL 33620

RE: **Expedited Approval for Initial Review**

IRB#: Pro00027955

Title: Student conceptual understanding in Anatomy & Physiology: Investigating formative assessment and lexical analysis

**Study Approval Period: 11/3/2016 to 11/3/2017**

Dear Ms. Carter:

On 11/3/2016, the Institutional Review Board (IRB) reviewed and **APPROVED** the above application and all documents contained within, including those outlined below.

**Approved Item(s):**

**Protocol Document(s):**

[Carter\\_protocol\\_Sept\\_2016.docx](#)

**Consent/Assent Document(s)\*:**

[Carter\\_study\\_Sept\\_2016\\_Adult\\_Minimal\\_Risk.docx.pdf](#)

[Carter\\_study\\_Online\\_Consent\\_Form\\_\(No\\_Signature\\_Line\).docx](#) (not a stamped form).

\*Please use only the official IRB stamped informed consent/assent document(s) found under the "Attachments" tab. Please note, these consent/assent document(s) are only valid during the approval period indicated at the top of the form(s). Waivers are not stamped.

It was the determination of the IRB that your study qualified for expedited review which includes activities that (1) present no more than minimal risk to human subjects, and (2) involve only procedures listed in one or more of the categories outlined below. The IRB may review

research through the expedited review procedure authorized by 45CFR46.110. The research proposed in this study is categorized under the following expedited review category:

- (5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis).
- (6) Collection of data from voice, video, digital, or image recordings made for research purposes.
- (7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Your study qualifies for a waiver of the requirements for the documentation of informed consent as outlined in the federal regulations at 45CFR46.117(c) which states that an IRB may waive the requirement for the investigator to obtain a signed consent form for some or all subjects if it finds either: (1) That the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality. Each subject will be asked whether the subject wants documentation linking the subject with the research, and the subject's wishes will govern; or (2) That the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context. (Online consent).

As the principal investigator of this study, it is your responsibility to conduct this study in accordance with IRB policies and procedures and as approved by the IRB. Any changes to the approved research must be submitted to the IRB for review and approval via an amendment. Additionally, all unanticipated problems must be reported to the USF IRB within five (5) calendar days.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,



Kristen Salomon, Ph.D., Vice Chairperson  
USF Institutional Review Board





## Hillsborough Community College

www.hccfl.edu  
877.736.2575

HCC INSTITUTIONAL REVIEW BOARD  
39 Columbia Drive #415 • Tampa, FL 33606-3584  
(813) 253-7193

October 20, 2017

Kelli Carter

RE: HCC IRB #2017\_009

TITLE: Student conceptual understanding in Anatomy & Physiology: Investigating formative assessment and lexical analysis

Dear Ms. Carter:

On October 20, 2017, the HCC Institutional Review Board (IRB) determined that your research meets Hillsborough Community College's requirements and federal criteria for expedited status which includes activities that (1) present no more than minimal risk to human subjects [21 CFR 56.110], and (2) involve only procedures listed in one or more of the categories listed below:

- (1) Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. [45 CFR 46.101(b)(1)]
- (7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies. [45 CFR 46.110(a)]

As the Principal Investigator for this project at HCC, it is your responsibility to ensure that this research is conducted as detailed in your Hillsborough Community College IRB application and supporting documents and consistent with the ethical principles outlined in the Belmont Report and with HCC policies and procedures.

The HCC IRB will maintain your research proposal and expedited status approval for a period of one year from the date of this letter. If you wish to continue this research beyond one year, you

must submit a request for continuing review at least 60 days prior to the expiration date. If you complete the research prior to the end of the one-year period, you must submit a request to close the study. Please note that it is your responsibility to notify the IRB of the status of this study no later than one year from the date of this letter, or upon completion of the research, whichever is sooner.

If you have any questions concerning this information, please contact me at (813) 253-7193 or by email at [azujoVIC@hccfl.edu](mailto:azujoVIC@hccfl.edu).

Wishing you all the best.

Sincerely,



Alisa M. Žujović, M.S., Chairperson  
HCC Institutional Review Board



MEMORANDUM

To: Kelli Carter  
From: Donna Burdzinski, Ed.D.  
Date: September 7, 2016  
Subject: Approval to conduct research at Pasco-Hernando State College (PHSC)

Congratulations, Kelli! Pasco-Hernando State College's Independent Research Review Committee reviewed your research proposal and accompanying documents, and made a recommendation for approval by the President's Administrative Leadership Team (PALT). Your request to conduct research entitled "*Student Conceptual Understanding in Anatomy & Physiology; Investigating Formative Assessment and Lexical Analysis*" was presented to the President's Administrative Leadership Team on September 6, 2016 and your request to conduct this research study at Pasco-Hernando State College was *approved*.

You may proceed with your research as indicated in your IRR application. If you have not already done so, please be sure that you have discussed this research thoroughly with Dr. Hanak, Biological Sciences Department Chair, in advance of initiating your research.

Best wishes to you in your research. The College is very interested in knowing your results as they will include our students.

Cordially,

*Donna Burdzinski*

Donna Burdzinski, Ed.D.  
Chair, Independent Research Review Committee  
Provost, North Campus  
Pasco-Hernando State College  
11415 Ponce de Leon Blvd.  
Brooksville, FL 34601

**East Campus**  
36727 Blanton Road  
Dade City, FL 33523  
352.567.6701

**North Campus**  
11415 Ponce de Leon Boulevard  
Brooksville, FL 34601  
352.796.6726

**Porter Campus  
at Wiregrass Ranch**  
2727 Mansfield Boulevard  
Wesley Chapel, FL 33543  
813.527.6615

**Spring Hill Campus**  
450 Beverly Court  
Spring Hill, FL 34606  
352.688.8798

**West Campus/District Office**  
10230 Ridge Road  
New Port Richey, FL 34654  
727.847.2727

[www.phsc.edu](http://www.phsc.edu)

AN EQUAL ACCESS/EQUAL OPPORTUNITY INSTITUTION

## Appendix D: Publication consent from APS

Publication copyright permissions from Advances in Physiology Education. Retrieved from <https://www.physiology.org/author-info.permissions>

2/5/2019

Copyright and Permissions

HOME | JOURNALS ▾

### Copyright and Permissions

Reuse by Authors of Their Work Published by APS | Reuse by Non-authors of APS Published Content

Reuse in APS Publications of non-APS Published Content

#### Reuse by Authors of Their Work Published by APS

The APS Journals are copyrighted for the protection of authors and the Society. The Mandatory Submission Form serves as the Society's official copyright transfer form. Author's rights to reuse their APS-published work are described below:

Republishing in New Works	Authors may republish parts of their final-published work (e.g., figures, tables), without charge and without requesting permission, provided that full citation of the source is given in the new work.
Meeting Presentations and Conferences	Authors may use their work (in whole or in part) for presentations (e.g., at meetings and conferences). These presentations may be reproduced on any type of media in materials arising from the meeting or conference such as the proceedings of a meeting or conference. A copyright fee will apply if there is a charge to the user or if the materials arising are directly or indirectly commercially supported <sup>1</sup> . Full citation is required.
Theses and Dissertations	Authors may reproduce whole published articles in dissertations and post to thesis repositories without charge and without requesting permission. Full citation is required.
Open Courseware	Authors may post articles, chapters or parts thereof to a public access courseware website. Permission must be requested from the APS <sup>1</sup> . A copyright fee will apply to a book chapter and during the first 12 months of a journal article's publication. Full citation is required.

<https://www.physiology.org/author-info.permissions>

1/3

2/5/2019

Copyright and Permissions

HOME | JOURNALS ▾

	published <sup>1</sup> (see exception to authors' own institution's repository, as note below).
Institutional Repositories (non-theses)	Authors may deposit their accepted, peer-reviewed journal manuscripts into an institutional repository providing: <ul style="list-style-type: none"><li>the APS retains copyright to the article<sup>1</sup></li><li>a 12-month embargo period from the date of final publication of the article is observed by the institutional repository and the author</li><li>a link to the article published on the APS or publisher-partner website is prominently displayed alongside the article in the institutional repository</li><li>the article is not used for commercial purposes</li><li>self-archived articles posted to repositories are without warranty of any kind</li></ul>
	<sup>1</sup> Unless it is published under the APS Open Access ( <i>AuthorChoice</i> ) option, which allows for immediate public access under a Creative Commons license (CC BY 4.0) (See also the APS Policy on Depositing Articles in PMC.)

Go to top

< Back to Information for Authors

Sign up for alerts

SIGN UP



Privacy policy | Disclaimer

<https://www.physiology.org/author-info.permissions>

2/3

Carter, K.P. & Prevost, L.B. (2018) Question order and student understanding of structure and function. *Advances in Physiology Education*, 42(4), 576-585.

*Adv Physiol Educ* 42: 576–585, 2018;  
doi:10.1152/advan.00182.2017.

## HOW WE TEACH | *Generalizable Education Research*

### Question order and student understanding of structure and function

**Kelli P. Carter and Luanna B. Prevost**

*Department of Integrative Biology, University of South Florida, Tampa, Florida*

Submitted 7 December 2017; accepted in final form 1 August 2018

**Carter KP, Prevost LB.** Question order and student understanding of structure and function. *Adv Physiol Educ* 42: 576–585, 2018; doi:10.1152/advan.00182.2017.—The relationship between structure and function is a core concept in physiology education. Written formative assessments can provide insight into student learning of the structure and function relationship, which can then inform pedagogy. However, question order may influence student explanations. We explored how the order of questions from different cognitive levels affects student explanations. A junior level General Physiology class was randomly split in half. One-half of the students answered, “Define the principle: form reflects function,” followed by “Give an example of the principle: form reflects function” (format DX), whereas the other half answered, “Give an example of the principle: form reflects function,” followed by “Define the principle: form reflects function” (format XD). Human grading and computerized lexical analysis were used to evaluate student responses. Two percent of students in the format DX group related structure and function in their definition, whereas 48% of students related structure and function in their examples. In the format XD group, 17% related structure and function in their definition, and 26% related structure and function in their example of the principle. Overall, students performed better on the last question in the sequence, which may be evidence for conceptual priming. Computerized lexical analysis revealed that students draw on only a few levels of organization and may be used by instructors to quickly assess the levels of organization students use in their responses. Written assessment coupled with lexical analysis has the potential to reveal student understanding of core concepts in anatomy and physiology education.

assessment; lexical analysis; structure function relationship; undergraduate

#### INTRODUCTION

There is strong consensus from the national community of biologists, including physiologists, that undergraduate biology instruction should focus on student understanding and application of core concepts (1, 20). A core concept is an important foundation learning framework for students on which many topics can be built. The relationship between structure and function is a core concept in physiology education (20, 21). It is described in a variety of ways in research articles, education policy reports, and textbooks (Table 1). The predominant definition of the structure function relationship in biology education literature is, “the function of a cell, tissue, or organ is determined by its form” (21). However, at our institution, students are exposed to this concept in different ways in their

coursework. For example, students in Human Anatomy and Physiology see, “. . . function always reflects structure. That is, what a structure can do depends on its specific form” (17). In General Physiology, students see, “The function of a biological system typically cannot be understood without knowledge of its structure, and vice versa” (12). Despite the variation in wording, comprehension of the structure function core concept acts as a foundation in the learning process (1, 20) and helps to explain other core concepts, such as homeostasis or cell-to-cell communication, across all levels of organization (20). For example, the process of homeostasis requires a sensor, a control center, and an effector, all of which are structures that carry out particular functions (e.g., detecting and responding to environmental changes) for negative feedback and homeostasis. As a student understands the structure function relationship, this provides insight into the process of homeostasis.

**Cognitive levels.** Bloom’s taxonomy is a hierarchy of the cognitive levels for learning and assessment (2). The taxonomy may be used to evaluate learning outcomes and assessment methods (2, 5). There are six levels of cognition in Bloom’s taxonomy. The first level of Bloom’s taxonomy is “remember,” which refers to retrieving information from long-term memory and is typically associated with recall or recognition tasks (2). The second level is “understand,” which goes beyond simply remembering material and refers to grasping the meaning and extrapolating information from the learning context. Because the taxonomy is a hierarchical framework, remembering is necessary for understanding (2). Students must be able to move beyond “remember” for meaningful learning to occur.

**Facilitating written formative assessments.** Written formative assessments can provide insight into student learning of the structure and function relationship, which can then inform pedagogy. Written assessments are a form of constructed response questions and allow students to use their own knowledge to construct their response rather than choose from a list of options, like in a multiple-choice question (19). Constructed response questions reveal student thinking, since students use their own heterogeneous ideas to construct their written response (4, 22). Formative assessment provides feedback to both the instructor and student during the learning process (3). Instructors can use formative assessments to examine students’ existing conceptions and confusions about the structure function relationship within the context of physiology education and provide this feedback to students. Students can use feedback about their existing conceptions to modify their thinking and improve their learning. Instructors may also use feedback from formative assessments to reform instructional practices.

Despite the depth of information on student learning obtained from written assessments, there are drawbacks. The

Address for reprint requests and other correspondence: L. B. Prevost, Dept. of Integrative Biology, University of South Florida, 4202 E. Fowler Ave, SCA110, Tampa, FL 33620 (e-mail: prevost@usf.edu).

Table 1. Description of the core concept "structure function" in different types of publications

Description of Core Concept "Structure Function"	Type of Publication	Reference
"The function of a cell, tissue, or organ is determined by its form. Structure and function (from the molecular level to the organ system level) are intrinsically related to each other."	Physiology education research article	Michael and McFarland (21)
"Basic units of structure define the function of all living things."	Biology education policy report	AAAS (1)
"Mechanisms refers to the components of actual, living animals and the interactions among those components that enables the animals to perform as they do."	Physiology textbook	Hill et al. (12)
"... [F]unction always reflects structure. ... what a structure can do depends on its specific form."	Anatomy and physiology textbook	Marieb and Hoehn (17)

resource constraints of time, money, and expertise limit their use, and these constraints may delay formative feedback (10). The reliability and consistency of human grading is also a concern. The effectiveness of constructed response questions to provide insight into student thinking may be enhanced by using automated scoring (32). Automated scoring, such as lexical analysis, helps to alleviate resource constraints and inconsistency of human grading (22).

Lexical analysis involves linguistic-based analysis to identify, extract, and categorize text and can be used to quickly evaluate a large number of constructed responses (11, 22). Lexical analysis tools can incorporate human grading to reduce grading inconsistencies in computer scoring and augments the interpretation of student responses. Lexical analysis is a means for assessing student understanding, since the resulting lexical categories reveal terms and phrases students use in their responses (32). For physiology education, a specific application of lexical analysis could be determining the levels of organization students are using in their responses.

**Levels of organization.** Levels of organization is a core principle in physiology (20). It is necessary for students to grasp the concept of physiological processes at multiple scales, and to focus on emergent properties at all levels of organization (1, 33). Students should also be able to discern that organisms carry out physiological processes at multiple levels of organization simultaneously (16, 20). By including levels of organization into written formative assessments, instructors can get a much clearer picture of student comprehension of this important concept.

**Question order.** In psychology and survey literature, there is a myriad of research on question order, and whether questions, or the responses to the questions, are influenced by preceding questions (24, 26, 27). A response to a question may be due to the content of the question itself, or the response may be due to the order of the questions (26). For example, *question A* may be answered differently if it is asked before *question B* compared with the reverse order. Within educational assessments, questions are rarely offered in isolation but are part of a sequence.

Question order may elicit conceptual priming and affect student explanations. When students are asked a question, they search their memories to retrieve the information. The search is truncated as soon as enough information is found to answer the question. According to the theory of increased cognitive accessibility, their response to the next question will be based on the information recently retrieved (28); this phenomenon is termed conceptual priming. Exposure to a concept acts as a prime, which then activates memories associated with the prime in subsequent questions. A larger number of preceding

questions would increase the amount of potentially relevant information retrieved and may make subsequent questions cognitively easier.

Prior education research demonstrates mixed results for question order and conceptual priming; much of this research focused on multiple-choice questions. Question order is more likely to have an effect on student performance when the multiple-choice assessment is given under a speed condition: students have a certain amount of time to complete the assessment (15). In a sample of 103 veterinary science students in a timed exam, difficult multiple-choice items early in examinations were correlated with decreased performance compared with students who had easier items first (18). When the difficult questions are first, students run out of time before they get to the easier questions. The question order under a timed condition may lead to fatigue as well as priming. The more difficult questions first may cause conceptual priming, but, because the students are in a timed condition, they may not have the opportunity to demonstrate their knowledge in the easier questions. Huck and Bowers (13) found no difference in performance when varying multiple-choice item sequences when students are in a power condition and have unlimited time to finish the assessment. Singh (29) compared question orders with multiple-choice physics questions requiring symbolic or numerical calculations (quantitative) and conceptual reasoning (conceptual) questions, and found students tended to use the quantitative question if it was first to answer the conceptual question. However, students struggled with answering the conceptual question when it was before the quantitative question. Question order has an effect on multiple-choice questions under both speed and power conditions.

Few studies have examined the effect of question order on student performance in constructed response assessments. In a study using constructed response questions but also a power condition, Federer et al. (9) found a decrease in the use of evolutionary key concepts across constructed response item sequences, as well as a decrease in response length. At first glance, this result appears to negate the idea of conceptual priming. On deeper examination, the result may be due to the students being asked similar questions repeatedly and wanting to avoid redundancy (31).

Formative assessment with varying question orders, i.e., easy to hard, hard to easy, may help students learn how to recognize and apply concepts, regardless of context (29). By changing question order in formative assessment tasks, this allows students to begin to recognize underlying core concepts and transfer knowledge from one context to another. This recognition strategy is important as students move from novice

Table 2. Description of question format DX and XD

Format	Description	Bloom's Taxonomy
DX	"Define the principle: form reflects function," followed by, "give an example of the principle: form reflects function" from the human body.	"Remember" followed by "understand"
XD	"Give an example of the principle: form reflects function" from the human body, followed by, "define the principle: form reflects function."	"Understand" followed by "remember"

Each question format was administered to one-half the class. Students were asked to provide one definition and three examples.

to expert in conceptual reasoning (6). The general lack of research on the intersection of formative assessment and constructed response questions, and question order effects, namely conceptual priming, in physiology education motivated this study.

In this study, we explore how the order of questions from different cognitive levels affect student performance. Prior research has focused primarily on multiple-choice questions with varying results. Our focus is on written assessments, specifically constructed response questions. We then use lexical analysis to examine student thinking, including correct and incorrect ideas. We also examine on which levels of organization students draw when demonstrating their understanding of the structure function relationship.

#### METHODS

**Question development and administration.** We developed two short-answer questions to define and give examples of the concept "structure function" (Table 1). We developed the questions with feedback from three physiology instructors and two science education researchers to ensure that the questions were appropriate in content and structure for undergraduate anatomy and physiology courses. We also interviewed students for their interpretations of the questions (described below in the student interview section).

We administered the two short-answer questions to students in a junior level General Physiology course at a large southeastern public university. The General Physiology course is a survey of the structures and metabolic processes that vertebrate and invertebrate animals use to cope with their environments. The focus of the course is the evolution, ecology, and development of organismal functions, taking a comparative and integrative approach. The relationship between structure and function is an underlying concept of the course and is described within the course textbook (12). The concept was not explicitly stated during the class, but was implicit in the course content. For example, the structures of the small intestine responsible for absorption and the structures that produce light in fireflies were topics discussed in the class. Before administration of the questions, students had explored the structure and function relationship within several contexts, including digestion, nutrition, metabolism, and transport.

The questions were administered midway through the semester as part of regular online homework via the course management system.

The study protocol was approved under the Institutional Review Board (Pro00014285), and students provided consent before participation. Although nothing prevented students from using outside resources, students were instructed to explain their answer to the best of their ability without the use of outside resources. The following prompt was used: "For the following set of questions, please respond to the best of your knowledge. Please do not refer to any outside sources (textbooks, notes, internet, etc.)."

Students received one point for completion of each question. The class was randomly split in half, and each half received the questions in a different order. A total of 62 students answered format DX (define followed by give an example), and 69 students answered format XD (give an example followed by define) (Table 2). Based on Bloom's taxonomy, "Define the principle: form reflects function" is at the "remember" cognitive level, whereas "Give an example of the principle: form reflects function" is at the "understand" level. For each question format, students were asked to provide one definition and three examples. Students were not able to go back to a question within the sequence.

**Human scoring of student responses.** We used a three-bin rubric to code the presence or absence of the following concepts: structure, function, and structure relates function. We coded the student responses for the presence (1) or absence (0) of structure and function, and whether students provided a correct statement linking structure and function using a three-bin rubric (Table 3). For example, responses that mentioned "parts," "cardiovascular system," "vessels," "muscle," or "pupils" were coded 1 (present) for structure, as shown in bold font in Table 3. Responses that mentioned "produce a final product," "complete a task," and "transport blood" were coded 1 (present) for function, as shown in Table 3 with underlined text. Responses that provided a correct statement linking structure and function students were coded 1 (present) for structure relates function (Table 3). For example, in Table 3, "The cardiovascular system uses vessels to transport blood around the body" was coded as a 1 for structure (cardiovascular system/vessels), a 1 for function (transport), and a 1 for structure relates function, because the student demonstrates the connection between the two. However, responses that mention a structure and a function but do not give a correct statement linking the two were coded 0 for structure relates function. For example, in Table 3, "Transport across the membrane" was coded as a 1 for structure (membrane) and a 1 for function (transport), but as a 0 for structure relates function because there is no connection between the two.

To obtain interrater reliability, a subset of student responses (15%) were scored by both authors. An interrater reliability of  $>0.70$  ( $\kappa$ ) was

Table 3. Human scoring of student responses using three-bin rubric

Student Response	Structure	Function	Structure Relates Function
"... [A] pathway of <b>parts</b> that interact with one another in order to <u>produce a final product</u> "	1	1	1
"... [A] set of steps a system uses to <u>complete a task</u> "	0	1	0
"The <b>cardiovascular system</b> uses <u>vessels</u> to <u>transport blood</u> around the body."	1	1	1
" <u>Hormonal secretion</u> "	0	1	0
" <b>Muscle</b> functions"	1	0	0
"When our <b>pupils</b> <u>dilate</u> and <u>adjust</u> to lower light"	1	1	1
" <u>Transport</u> across the <b>membrane</b> "	1	1	0

Within student responses, structures are in bold and functions are underlined.

Table 4. Example student responses from “define” and “give example” questions, and categorization of student responses in SPSS Modeler

Response	Category					
	<u>Function/General</u>	<i>Mechanism</i>	<b>Nutrients</b>	<i>Disorder</i>	<b>Process</b>	<b><i>Structure/organ level</i></b>
“ <u>Nerve</u> damage and therefore <u>loss of feeling</u> and function will occur in these areas.”	X					X
“The loss of functions associated with a third-degree burn would be the loss of <u>sensation</u> , loss of some <u>movement</u> , and loss of <u>protection</u> .”	X					
“A <i>mechanism</i> is a set <b>process</b> by which a task is accomplished.”		X			X	
“A <b>process</b> by which a physiological reaction takes place in an organism.”					X	
“An example of a <i>mechanism</i> in the human body would be when the <u>heart rate is increased</u> or when the <u>respiratory rate is increased</u> in response to exercise, muscle activity, or anxiety.”	X	X				
“The <u>alveoli</u> of the <u>lungs</u> give a <u>high surface area</u> for <u>gas exchange</u> .”	X					X
“The main purpose of villi is for <u>absorption</u> , and the more villi present, the more <u>surface area</u> is available for more effective <u>absorption</u> . If the villi of the small intestine were damaged, <b>nutrients</b> from <b>food</b> will not be <u>absorbed</u> or <u>digested</u> properly, and an individual could become extremely <u>malnourished</u> .”	X		X	X		X

Functional/general words are underlined; mechanism words are in italics; nutrients words are in bold; disorder words are in italics and underlined; process words are in bold and underlined; and structure/organ level words are in bold, in italics, and underlined.

achieved (14), and the remaining responses were coded by KPC. Response length varied from one word to a short paragraph. Analysis of the human coding data consisted of determining the percentage of student responses that used structure, function, or related structure and function.

**Extraction and categorization using lexical analysis.** Student responses to the two short-answer questions were imported into and analyzed using IBM SPSS Modeler. Lexical analysis consists of extraction and categorization. During extraction, the software identified key terms and phrases from student responses using the lexical library that came with the software, as well as a “custom library” that was built to capture terms and phrases commonly used in physiology courses. A lexical library is similar to a dictionary of terms and phrases. The custom lexical library also includes synonyms, abbreviations, spelling variations, and misspellings. For example, synonyms and misspellings of “absorption” included “absorptive,” “absortion,” and “absorbition,” and all of these terms were added to the lexical library. Additional details on lexical libraries are described by Haudek and colleagues (11).

Examples of terms included in the libraries and extracted by the software are shown in Table 4. For example, in the first response in Table 4, the software recognizes the word “nerve” and the phrase “loss of feeling.” In the third response in Table 4, the software recognizes the words “mechanism” and “process,” whereas, in the fourth response, only “process” is recognized. In the fifth response in Table 4, the student mentions heart rate and respiratory rate increasing. These are both considered functions, as both heart and respiratory are adverbs modifying the word “rate.”

The second step of lexical analysis is categorization, in which the extracted terms and phrases are grouped into categories. A category contains all term and phrases in student responses that represent a homogeneous idea. For example, the category, structure, includes the terms “nerves,” “alveoli,” “lungs,” and “surface area,” which represent organs, components of organs, and properties of structures (i.e., surface area). The student’s responses were assigned to zero, one, or more categories following extraction. For example, the student response, “nerve damage and therefore loss of feeling and function” is categorized as function/general and structure/organ level (Table 4). We designed the category grain size in a hierarchical fashion to reflect the biological levels of organization from molecular to organ systems (Table 5). We also separated mechanism and process into two lexical categories.

**Question order comparison.** The length of the written responses were evaluated between question formats to assess if students were

more verbose with a definition question followed by a give-example question (format DX), or a give-example question followed by a definition question (format XD). The length of student responses was analyzed using a Mann-Whitney *U*-test. We also performed a comparison of the lexical categories for format DX and format XD to determine whether students used different words and phrases when they were asked to define the core concept compared with giving an example of the core concept. A Fisher’s exact test analysis was performed to compare the lexical analysis categories between the DX and XD question formats.

**Student interviews.** We conducted interviews with four students following the interview protocol used by Haudek et al. (11). Interviews began with a think-aloud protocol during which students answered the same questions for which they had provided written responses in their homework. We analyzed their verbal response to confirm that students were interpreting the questions in the manner intended and compare the verbal and written responses. We coded their verbal responses using the structure, function, and structure relates function categories used for written responses and compared the coding for written and verbal responses. We also identified the categories used in the verbal responses and compared them to categories identified in written responses.

In the second part of the interview, we examined students’ familiarity with levels of organization. Students were first asked if they could recall the levels of organization. Then they were asked about

Table 5. Hierarchical structure and function lexical categories from SPSS Modeler

Structure	Function	Other
Structure	Function	Dynamics
Structure/biomolecules	Function/cellular level	Mechanism
Structure/cell	Function/organ level	Organism
Structure/cell components	Function/organ system level	Process
Structure/tissue	Function/organism level	
Structure/tissue components	Function/general	
Structure/organ	Function/disorder	
Structure/organ components		
Structure/organ system		
Structure/part		
Structure/complex structures		
Properties of structures		



which level of organization they typically found themselves thinking for examples to identify student preferences within the hierarchy.

## RESULTS

**Human scoring.** Human scoring of the responses revealed the percentage of students who used structure, function, or related structure and function in their responses for each question version. When asked to define (cognitive level: remember) the core principle structure and function first (format DX), 13% of students identified structures, 34% identified functions, and 2% of students were able to link the two concepts (Fig. 1). Students were asked to provide a total of three examples of the core principle (cognitive level: understand). The identification of structures and functions were similar for the three examples, while relating structure and function increased from the first to the third example. By the third example, 48% of students related structure and function in their responses. Overall, students mentioned functions in their responses more often than structures. Within the examples, almost 100% of the student responses discussed functions.

When asked to define the core concept second, after providing examples (format XD), 20% of students identified structures, 43% identified functions, and 17% of students were able to link structure and function in their definition (Fig. 2). When asked to provide examples first, before giving a definition, <30% of students accurately related structure and function in any one of the three examples.

**Response length.** Student response length varied from one word to a short paragraph (102 words). There was not a significant difference in response length between question formats for define, give example 1, or give example 2. For the third example, response length was greater for format DX (median = 16.06) than for format XD, give example 3 (median = 15.69) (Mann-Whitney test,  $U = 1717.00$ ,  $P = 0.05$ ,  $d = 0.019$ ) with an extremely small effect size (7).

**Lexical analysis.** Lexical analysis of the students' written responses produced 23 lexical categories (Table 5). We compared lexical categories between the two question formats. For the question, "Define the principle: form reflects function," we identified 10 categories in student responses to the format DX, and 13 in the format XD responses. Figure 3 shows the seven

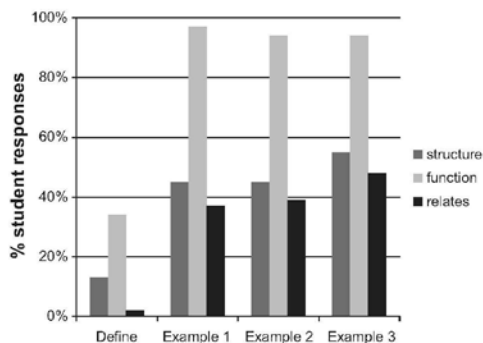


Fig. 1. Human scoring of student responses to format DX (define followed by give an example).  $n = 62$  students.

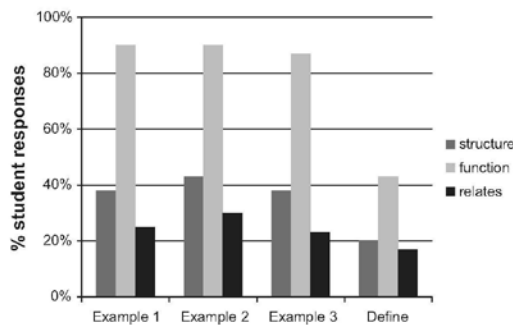


Fig. 2. Human scoring of student responses to format XD (give an example followed by define).  $n = 69$  students.

most commonly used categories in student responses. These seven categories were found in >10% of student responses. For the question, "Give an example of the principle: form reflects function," both the format DX and format XD responses contained 20 categories, although only 11 categories are shown. These 11 categories were found in >10% of student responses (Fig. 4).  $\chi^2$  analysis of the lexical categories for the "define" question between the two question formats demonstrated no significant difference between the question formats [ $\chi^2$  Define [degrees of freedom (df) = 14,  $N = 273$ ] = 13.61,  $P = 0.479$ ; Fig. 3]. Similarly,  $\chi^2$  analysis of the lexical categories for the "give example" question also demonstrated no significant difference [ $\chi^2$  Give Example (df = 19,  $N = 953$ ) = 28.89,  $P = 0.068$ ; Fig. 4].

Fisher's exact test analysis of format DX revealed a significant difference in the number of structure and function lexical categories used in student responses when comparing "define" and "give example 1," "define" and "give example 2," and "define" and "give example 3" ( $P < 0.05$ ) (Tables 6 and 7). However, there was not a significant difference in the structure and function lexical categories between the examples.

Fisher's exact test analysis of format XD revealed a significant difference in the structure and function lexical categories

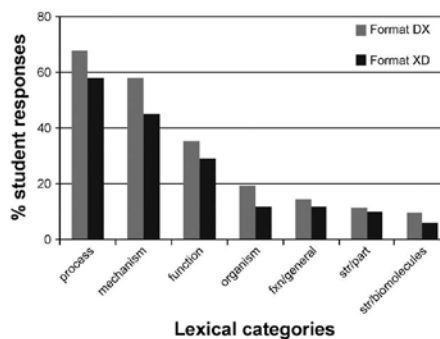


Fig. 3. Lexical categories contained in student responses to, "Define the principle: form reflects function." Only categories found in >10% of the student responses are shown. fxn, Function; str, structure. Format DX (define followed by give an example):  $n = 62$  students; format XD (give an example followed by define):  $n = 69$  students.

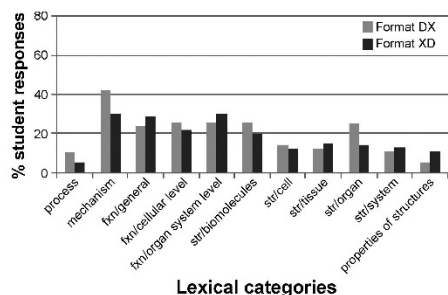


Fig. 4. Lexical categories contained in student responses to, "Give an example of the principle: form reflects function." Only categories found in >10% of the student responses are shown. fxn, function; str, structure. Format DX (define followed by give an example): n = 62 students; format XD (give an example followed by define): n = 69 students.

between "define" and "give example 1," "define" and "give example 2," and "define" and "give example 3" (P < 0.05) (Tables 8 and 9). However, there was not a significant difference in the structure and function lexical categories between the examples.

**Student interviews.** The four students interviewed interpreted the questions, appropriately giving a definition of structure and function along with examples. Student responses were coded in the same manner as written responses for structure, function, and structure relates to function. One student related structure and function in her definition and examples for both verbal and written responses. Another student identified structures only when prompted to define the core concept in both written and verbal responses, but related structure and function in examples of the core concept in both written and verbal responses. The other two students identified only structures in their written responses to the definition question, yet during the

Table 6. Format DX structure lexical categories with frequency by question prompt and Fisher's exact test results comparing structure lexical categories by question prompt

Format DX Structure Lexical Categories	Prompt/Frequency			
	Define	Give example 1	Give example 2	Give example 3
Structure	0	0	0	0
Biomolecules	6	15	15	18
Cell	1	6	7	13
Cell components	0	1	2	1
Complex structures	0	4	2	0
Organ	0	14	16	17
Organ components	0	0	1	6
Organ system	0	8	6	6
Tissue	0	12	4	6
Tissue components	0	1	1	1
Part	7	0	0	0
	Fisher's exact test		Significance (P < 0.05)	
Define: give example 1	31.394		0.000	
Define: give example 2	28.499		0.000	
Define: give example 3	31.126		0.000	
Example 1: example 2	6.367		NS	
Example 2: example 3	6.744		NS	

NS, nonsignificant. DX, define followed by give an example.

Table 7. Format DX function lexical categories with frequency by question prompt and Fisher's exact test results comparing function lexical categories by question prompt

Format DX Function Lexical Categories	Prompt/Frequency			
	Define	Give example 1	Give example 2	Give example 3
Function	22	1	4	2
Cellular level	3	15	16	17
Disorders	0	1	3	5
General	9	14	11	19
Organ level	0	3	7	5
Organ system level	0	27	12	9
Organism level	1	1	0	1
	Fisher's exact test		Significance (P < 0.05)	
Define: give example 1	61.710		0.000	
Define: give example 2	43.650		0.000	
Define: give example 3	46.637		0.000	
Example 1: example 2	10.579		NS	
Example 2: example 3	4.818		NS	

NS, nonsignificant. DX, define followed by give an example.

interview related structure and function in their responses to both the definition and give examples questions. The students also used similar lexical categories in their verbal and written responses. For the definition question, student responses were assigned to the same structure, structure/organ, and function categories in both verbal and written responses. For the example question, three of the four student responses were assigned to the same categories: structure/organ, function/organ, structure/tissue, function/tissue.

During interviews, we examined students use of structures and functions within the levels of organizations. When asked which level of organization they typically thought about as

Table 8. Format XD structure lexical categories with frequency by question prompt and Fisher's exact test results comparing structure lexical categories by question prompt

Format XD Structure Lexical Categories	Prompt/Frequency			
	Define	Give example 1	Give example 2	Give example 3
Properties of structures	0	9	8	6
Structure	2	0	0	0
Biomolecules	4	16	9	17
Cell	2	6	8	11
Cell components	0	2	4	4
Complex structures	0	2	2	2
Organ	0	9	12	9
Organ components	0	3	3	1
Organ system	0	5	13	9
Tissue	1	6	8	9
Tissue components	0	2	2	1
Part	7	0	0	0
	Fisher's exact test		Significance (P < 0.05)	
Define: give example 1	31.394		0.000	
Define: give example 2	28.499		0.000	
Define: give example 3	31.126		0.000	
Example 1: example 2	6.367		NS	
Example 2: example 3	6.744		NS	

NS, nonsignificant. XD, give an example followed by define.

Table 9. *Format XD function lexical categories with frequency by question prompt and Fisher's exact test results comparing function lexical categories by question prompt*

Format XD Function Lexical Categories	Prompt/Frequency			
	Define	Give example 1	Give example 2	Give example 3
Function	20	0	0	1
Cellular level	0	19	14	12
Disorders	1	2	1	5
General	8	18	19	22
Organ level	0	4	6	4
Organ system level	3	28	18	17
Organism level	0	0	3	0
	Fisher's exact test		Significance ( $P < 0.05$ )	
Define: give example 1	63.200		0.000	
Define: give example 2	56.722		0.000	
Define: give example 3	47.608		0.000	
Example 1: example 2	5.654		NS	
Example 2: example 3	6.881		NS	

NS, nonsignificant. XD, give an example followed by define.

examples of structures and functions, three out of four students interviewed identified the organ level. The interviewer then asked why they thought of organs as examples of structures and functions. Students described the tangible nature of organs and being able to identify a clear purpose as reasons for focusing at the organ level, as exemplified in the following student responses:

Because, I feel like they [organs] incorporate enough of tissues and cells. I feel like that's the level that you get to where you're actually, you have something that has a purpose. Obviously tissues have a purpose too, but I feel like tissues just build up organs, that's what their main thing, whereas like your organ does something.—*Student 1*

Because they [organs] are the ones I remember the most, are bigger things, and not the smaller things, 'cause I can't wrap my mind around the smaller things. I can see the bigger things clearly, and I can dissect it better than I could do with the smaller things.—*Student 2*

I think it's just easier for me to find an example when it's something that I can visibly see, or I've seen before, rather than going to the atomic level, or the molecular level.—*Student 3*

The fourth student thought of cells when thinking of examples and mentioned red blood cells as an example. The student then described her thinking about cells.

Because a lot of cells in our bodies have organelles and they perform a lot of different functions. They provide a broad spec of functions and they.... A cell is like, reproduce, divide, busy, busy bodies. I feel like my head's, a million thoughts is always going through it. So I guess I would consider myself.... A cell is equipped to do certain functions. So I guess I make sure I'm equipped with certain information or I have to know a certain amount of information on a topic. Honestly, there's a picture of a little cell in my head that's running around and it's talking to little organs and organ system.—*Student 4*

## DISCUSSION

In our study, students answered a total of two question types. To address our research question of whether or not question

order matters for formative assessment, one-half of the students answered a "remember" (define) question first followed by three "give example" (understand) questions (format DX). Only 2% of the students in format DX related structure and function in their definition, whereas 48% related structure and function in their third example. The other one-half of the students answered three "give example" (understand) questions followed by a "remember" (define) question (format XD). For this group who defined the core concept last, 17% of students related structure and function in their definition, and 23% related structure and function in their third example. Students performed better (related structure to function) on the definition question when it followed the example questions and on the example questions when they followed the definition question. The performance of students in this population may be related to the implicit nature of the core concept within this General Physiology course. The concept is explicitly taught in the prerequisite anatomy and physiology courses, but our results suggest that few students transfer their understanding of the concept in this study. Transference is learning a concept in one context and applying it in another (8). The minimal transference observed in this study suggests the need for a curriculum that explicitly helps students recognize the importance of this concept (20), and use the concept as a way to build connections between prior knowledge and newly introduced ideas.

**Cognitive level.** The cognitive level of a question, as well as the question order, may affect student performance. In this study, the definition question is at Bloom's lowest level, remember, and the give-example question is at the second Bloom's level, understand (2). In both groups of students (format DX and XD), fewer students were able to relate structure and function in their definitions than with examples. In their definitions, many students reiterated the question prompt without further explanation. We observed that some student repeated the words "form" and/or "function." For example, students wrote: "I am not exactly sure, but I would define it as the shape or form of an object is specific to its function." "It is the principle that describes how the form of a structure correlates with its function."

We also observed that students used vague terms like "something" in place of "form." For example, students answered: "It means that the general shape of something is related to its general function." "The way something exists or is made should allow it to function sufficiently."

It is plausible that as students are exposed to this definition during their learning process, they have difficulty with constructing meaning or connecting it to their existing knowledge. Learning this definition requires unpacking each component. The new information (the definition) is compared with existing knowledge (terms that make up the definition). If the student is unclear about the components of the definition (i.e., the term "form" or the term "function"), then the new information will fail to connect to their existing knowledge.

Our interviews probed students to define the individual words "structure" and "function" in further detail. Students were able to define function, but often had difficulty with structure. Their responses indicate it may be easier to give an example of structure than to produce an abstract definition:

Oh man. They're such simple words, but they're hard to define. Structure is just something that, kind of like a building, but not necessarily like the actual physical building. It's just like your organs are like a structure, like structures. Your kidney is a structure. It's just like a mass of matter that is specific for something.—*Student 1*

Structure. It is . . . I don't even know how to explain it. How is it that you know what these things are, but you don't know how to explain what they are? A structure is, I guess . . . something physical in the body. Oh my goodness. I don't know how to explain it.—*Student 4*

Students were more comfortable giving a definition of function:

A task or a duty or an action. That's my, an action. I would say and the function would be an action, it's like proactive, it's not like a structure which doesn't perform the action itself, it's just good for the function.—*Student 3*  
Not the goal, but what something is made to do, what its purpose is.—*Student 4*

This may also explain why more students were coded as a 1 for function than for structure in both their definitions and examples (Figs. 1 and 2). One implication of this result is that instructors need to make sure students know how to define the terms “form” or “function” in their own words to ensure students have a mental framework for the core concept.

**Conceptual priming.** Our results show that students performed better on the last questions in the sequence. Specifically, students in the format DX group performed better on the examples than students in the XD group, whereas students in the XD group performed better on the definition than did students in the DX group. This suggests that, for the DX group, the definition acted as a prime for the example question, and for the XD group, the examples acted as a prime for the definition question. In each case students had greater success retrieving or applying the core concept after priming. Interestingly, the three “give example” questions in each format do not seem to act as a prime for each other (Figs. 1 and 2). However, previous studies have shown that, when the cognitive task is similar, there is less likely to be priming (31).

**Lexical analysis.** Lexical analysis facilitates assessment in large-enrollment classes where the logistics of written responses can be challenging. Additionally, lexical analysis is valuable for an instructor to visualize the heterogeneous ideas students have in their written responses and assess class understanding. For the 2 questions used in this study, we identified 13 lexical categories for the “define” question and 20 lexical categories for the “give example” question. The “give example” question allows students the opportunity to demonstrate the versatility of their knowledge, as there are fewer constraints on the ideas that students can use compared with the “define” question. The use of lexical analysis to extract and categorize a wide range of student ideas from written responses may help instructors observe the heterogeneity of student thinking.

In addition, categories can be used to build predictive models that mimic human scoring. In this paper, we focused on comparing the two question formats, using human scoring and lexical analysis (extraction and categorization of phrases). The lexical analysis allowed us to look at the content of student responses without regard for accuracy when answers related structure and function. However, human scoring evaluates the accuracy of student responses. An additional step, model build-

ing, is needed to replicate the accuracy determined by human scoring during automated computer scoring. Predictive models that can automatically score student responses with high agreement to human scoring will be presented in a subsequent manuscript. Future research will focus on the development of predictive models to aid instructors in assessing the completeness of student responses in a short amount of time.

Categorization using lexical analysis in this study allows instructors to quickly assess the levels of organization, which is unique to anatomy and physiology education. For example, instructors can determine whether students only draw examples from a specific level of organization taught in class, such as at the organ level, or whether they show a preference for certain levels of organization. Our results suggest students in this study drew on only a few levels of organization. The structures identified in student written responses were primarily at the molecular, organ, and organ system levels, whereas the functions included in student responses were at the cellular and organ system levels. This corresponds to our interview data in which three of the four students identified organs as the level of organization from which they draw.

Additionally, lexical analysis allows instructors to distinguish student usage of mechanism and process. Sometimes students use mechanism and process interchangeably, but at times they do not. With our coding, we were attempting to identify the use of each word separately, as this may be important for future model building. Hill et al. (12) defined mechanism as follows: “Mechanisms refers to the components of actual, living animals and the interactions among those components that enables the animals to perform as they do.”

Lira and Gardner (16) also focus on behaviors when defining mechanism. They found that students may use these two words to represent different levels of complexity: students describe mechanisms as being more detailed, whereas processes are mechanisms with “less details” and “less steps.”

**Implications for teaching.** Lexical analysis revealed that students used only a few levels of organization in their responses. This was also reflected in student interviews. We recommend that instructors discuss more examples of the core concept structure and function within a variety of levels of organization. Students can develop a broader application of the core concept and reason across multiple levels of organization to help them become more familiar with “less visible” examples. The majority of students enrolled in anatomy and physiology and general physiology courses intend to work in healthcare fields, where comprehensive knowledge of the human body is necessary. Therefore, students should discern the relationship between structure and function at every level of organization. Formative assessment questions should be designed to incorporate multiple levels of organization to assess student understanding. For example, if formative assessment questions are targeted at the organ level, they may not accurately assess student conceptions, or their incorrect ideas, at the molecular or cellular levels.

**Limitations.** One of the limitations of this study is the sample size. A relatively large junior level course was randomly split in half, and each half received the questions in a different order. This was necessary to alleviate extraneous variables, such as instructor, in the study. Prior research has shown that ~350 or more student responses are necessary to build statistically significant predictive models (11, 25). This

study was done at the beginning of our data collection, and we will continue to collect data to build statistical models that predict human scoring of these two questions. These statistical models will also allow instructors to promptly predict the completeness and correctness of student responses, in addition to viewing the richness of student ideas through extraction and categorization demonstrated in the lexical analysis presented here.

Another possible limitation in this study is the method of question administration. The questions were administered one at a time, and the student could not go back to an earlier question. This method of administration was necessary to ensure question order was maintained to address our research question. Students may prefer to answer easier questions first and more difficult questions later, but, in this study, students did not have the option of changing the order of the questions. Although this may have caused some anxiety for the students, specifically format XD, which had the more difficult questions first, this is highly unlikely as the questions were administered online outside of class as low-stakes homework with no time limit. Students were aware that points were awarded for completion rather than correctness and were encouraged to give their best effort. This form of low-stakes formative assessment can help increase student confidence through positive feedback, encourage ideas to develop, and assist with identifying any areas of student confusion (30).

The structure function concept is presented in various ways in the literature, as shown in Table 1. For this study, we selected the wording “structure reflects function,” which students in this institution encounter in their foundational anatomy and physiology course, and students were familiar with this representation of the concept. To determine whether different wordings of the question prompt might affect student responses for this study, we compared student responses to two versions of our question prompt. In one version, we used the word “reflects” (the textbook wording; Ref. 17) and in the other we used the word “determines” (physiology education literature; Ref. 21). We administered the question versions to students in a junior level General Physiology course, randomly assigning one-half of the course to the “reflects” prompt and the other one-half to the “determines” prompt. We conducted human coding of the responses using the methods described in this study. We then compared the frequency of human coding for structure, function, and related structure and function and found no significant differences between the groups [ $\chi^2$  Structure ( $df = 1, N = 110$ ) = 0.370,  $P = 0.543$ ;  $\chi^2$  Function ( $df = 1, N = 110$ ) = 0.771,  $P = 0.380$ ;  $\chi^2$  Relates Structure Function ( $df = 1, N = 110$ ) = 0.270,  $P = 0.604$ ]. These results suggest that, at least in this population, the variation in wording did not affect how students responded to the question prompt.

**Conclusion.** Research in the use of lexical analysis for anatomy and physiology education is novel. The content of anatomy and physiology courses tends to be laden with medical terminology and levels of organization. Lexical analysis is suitable for identifying specific terms and phrases in student writing and can be used to automatically score large numbers of student-written responses in a relatively short amount of time. Written assessment coupled with lexical analysis has the potential to reveal student understanding of core concepts in anatomy and physiology education. During assessment design, instructors should con-

sider question order, as it may lead to conceptual priming and affect student explanations.

#### ACKNOWLEDGMENTS

We thank the students who participated in this study and the biology faculty who administered these questions in their class. We also thank Sally Treat for thoughtful comments regarding this manuscript.

#### DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

#### AUTHOR CONTRIBUTIONS

K.P.C. and L.B.P. conceived and designed research; K.P.C. and L.B.P. performed experiments; K.P.C. and L.B.P. analyzed data; K.P.C. and L.B.P. interpreted results of experiments; K.P.C. and L.B.P. prepared figures; K.P.C. and L.B.P. drafted manuscript; K.P.C. and L.B.P. edited and revised manuscript; K.P.C. and L.B.P. approved final version of manuscript.

#### REFERENCES

1. American Association for the Advancement of Science (AAAS). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: AAAS, 2011.
2. Anderson LW, Krathwohl DR, Bloom BS, Bloom BS. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.
3. Bell B, Cowie B. The characteristics of formative assessment in science education. *Sci Educ* 85: 536–553, 2001. doi:10.1002/sce.1022.
4. Birenbaum M, Tatsuoka KK. Open-ended versus multiple choice response formats—it does make a difference for diagnostic purposes. *Appl Psychol Meas* 11: 385–395, 1987. doi:10.1177/014662168701100404.
5. Bloom BS. *Taxonomy of Educational Objectives: The Classification of Education Goals. Cognitive Domain. Handbook 1*. New York: Longman, 1956.
6. Chi MT, Feltovich PJ, Glaser R. Categorization and representation of physics problems by experts and novices. *Cogn Sci* 5: 121–152, 1981. doi:10.1207/s15516709cog0502\_2.
7. Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum, 1988.
8. Duit R. On the role of analogies and metaphors in learning science. *Sci Educ* 75: 649–672, 1991. doi:10.1002/sce.3730750606.
9. Federer MR, Nehm RH, Opfer JE, Pearl D. Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Res Sci Educ* 45: 527–553, 2015. doi:10.1007/s11165-014-9435-9.
10. Ha M, Nehm RH, Urban-Lurain M, Merrill JE. Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE Life Sci Educ* 10: 379–393, 2011. doi:10.1187/cbe.11-08-0081.
11. Haudek KC, Prevost LB, Mosearella RA, Merrill J, Urban-Lurain M. What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE Life Sci Educ* 11: 283–293, 2012. doi:10.1187/cbe.11-08-0084.
12. Hill RW, Wyse GA, Anderson M. *Animal Physiology* (4th Ed.). Sunderland, MA: Sinauer, 2016.
13. Huck SW, Bowers ND. Item difficulty and sequence effects in multiple choice achievement tests. *J Educ Meas* 9: 105–111, 1972. doi:10.1111/j.1745-3984.1972.tb00765.x.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174, 1977. doi:10.2307/2529310.
15. Leary LF, Dorans NJ. Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Rev Educ Res* 55: 387–413, 1985. doi:10.3102/00346543055003387.
16. Lira ME, Gardner SM. Structure-function relations in physiology education: where's the mechanism? *Adv Physiol Educ* 41: 270–278, 2017. doi:10.1152/advan.00175.2016.
17. Marieb EN, Hoehn K. *Human Anatomy & Physiology* (10th Ed.). San Francisco, CA: Benjamin Cummings, 2016.
18. Marks AM, Cronje JC. Randomized items in computer-based tests: Russian roulette in assessment? *J Educ Technol Soc* 11: 41–50, 2008.
19. Martinez M. Cognition and the question of test item format. *Educ Psychol* 34: 207–218, 1999. doi:10.1207/s15326985Sep3404\_2.

20. Michael J, Modell H, McFarland J, Cliff W. The "core principles" of physiology: what should students understand? *Adv Physiol Educ* 33: 10–16, 2009. doi:10.1152/advan.90139.2008.
21. Michael J, McFarland J. The core principles ("big ideas") of physiology: results of faculty surveys. *Adv Physiol Educ* 35: 336–341, 2011. doi:10.1152/advan.00004.2011.
22. Nehm RH, Haertig H. Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *J Sci Educ Technol* 21: 56–73, 2012. doi:10.1007/s10956-011-9282-7.
23. Opfer JE, Nehm RH, Ha M. Cognitive foundations for science assessment design: knowing what students know about evolution. *J Res Sci Teach* 49: 744–777, 2012. doi:10.1002/tea.21028.
24. Posner MI. *Chronometric Explorations of Mind*. Hillsdale, NJ: Wiley, 1978.
25. Prevost LB, Smith MK, Knight JK. Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE Life Sci Educ* 15: ar65, 2016. doi:10.1187/cbe.15-12-0267.
26. Schuman H, Presser S. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. Orlando, FL: Academic, 1981.
27. Schwarz N, Sudman S. Introduction. In: *Context Effects in Social and Psychological Research*, edited by Schwarz N, Sudman S. New York: Springer, 1992. doi:10.1007/978-1-4612-2848-6\_1.
28. Schwarz N, Strack F. Context effects in attitude surveys: applying cognitive theory to social research. *Eur Rev Soc Psychol* 2: 31–50, 1991. doi:10.1080/14792779143000015.
29. Singh C. Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer. *Phys Rev Spec Top Phys Educ Res* 4: 1–10, 2008. doi:10.1103/PhysRevSTPER.4.010105.
30. Smyth K. The benefits of students learning about critical evaluations rather than being summatively judged. *Assess Eval High Educ* 29: 370–378, 2004. doi:10.1080/0260293042000197609.
31. Strack F. "Order effects" in survey research: activation and information functions of preceding questions. In: *Context Effects in Social and Psychological Research*, edited by Schwarz N, Sudman S. New York: Springer, 1992. doi:10.1007/978-1-4612-2848-6\_3.
32. Urbain-Lurain M, Moscarella RA, Haudek KC, Giese E, Sibley DF, Merrill JE. Beyond multiple choice exams: using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines. In: *Proceedings of the 39th ASEE/IEEE Frontiers in Education Conference*, San Antonio, TX, October 2009.
33. Wilensky U, Resnick M. Thinking in levels: a dynamic systems approach to making sense of the world. *J Sci Educ Technol* 8: 3–19, 1999. doi:10.1023/A:1009421303064.

