CAT DISSECTION AND HUMAN CADAVER PROSECTION
VERSUS SCULPTING HUMAN STRUCTURES FROM CLAY:
A COMPARISON OF ALTERNATE APPROACHES TO
HUMAN ANATOMY LABORATORY EDUCATION

A Dissertation in

Biology

By

John R. Waters

© 2008 John R. Waters

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2008
The dissertation of John R. Waters was reviewed and approved* by the following:

Robert B. Mitchell  
Professor of Biology  
Chair of Committee

Richard J. Cyr  
Professor of Biology  
Dissertation Adviser

Simon Gilroy  
Professor of Biology

Peggy N. Van Meter  
Associate Professor of Education

Douglas R. Cavener  
Professor of Biology  
Head of the Department of Biology

*Signatures are on file in the Graduate School.
ABSTRACT

Dissection and vivisection are traditional approaches to biology laboratory education. In the case of human anatomy teaching laboratories, there is a long tradition of using human and animal cadaver specimens in the classroom. In a review of the literature comparing traditional dissection and vivisection lessons to alternative lessons designed to reduce the time spent dissecting or the numbers of animals used, we conclude that it is difficult to come to any conclusion regarding the efficacy of different approaches. An analysis of the literature is confounded because many studies have very low statistical power or other methodological weaknesses, and investigators rely on a wide variety of testing instruments to measure an equally varied number of course objectives. Additional well designed studies are necessary before educators can reach any informed conclusions about the efficacy of traditional versus alternative approaches to laboratory education. In our experiments, we compared a traditional cat dissection based undergraduate human anatomy lesson to an alternative where students sculpted human muscles onto plastic human skeletons. Students in the alternative treatment performed significantly better than their peers in the traditional treatment when answering both lower and higher order human anatomy questions. In a subsequent experiment with a similar design, we concluded that the superior performance of the students in the alternative treatment on anatomy exams was likely due to the similarity between the human anatomy representation studied in lab, and the human anatomy questions asked on the exams. When the anatomy questions were presented in the context of a cat specimen, students in the traditional cat dissection treatment outperformed their peers in the alternative treatment. In a final experiment where student performance on a human anatomy exam was compared between a traditional prosected human cadaver treatment and the alternative clay sculpting treatment, no significant difference were detected, suggesting that the complexity or simplicity of the anatomy representation is less important than the similarity between the learning experience and the testing experience.
# TABLE OF CONTENTS

List of Tables......................................................................................................................vi
List of Figures......................................................................................................................viii
Preface....................................................................................................................................xii
Acknowledgements............................................................................................................xiv

Chapter 1. THE ROLE OF DISSECTION, VIVISECTION, MODELS, AND TECHNOLOGY IN LIFE SCIENCE CLASSROOMS.............................................................................. 1
  Introduction......................................................................................................................... 1
  Method................................................................................................................................. 2
  Results and Discussion....................................................................................................... 5
  Conclusion........................................................................................................................... 24
  Recommendations............................................................................................................ 26
  References.......................................................................................................................... 30

Chapter 2. CAT DISSECTION VS. SCULPTING HUMAN STRUCTURES IN CLAY: AN ANALYSIS OF TWO APPROACHES TO UNDERGRADUATE HUMAN ANATOMY LABORATORY EDUCATION................................................................. Appendix
  Introduction........................................................................................................................ 86
  Method............................................................................................................................... 87
  Results............................................................................................................................... 89
  Discussion........................................................................................................................ 91
  References........................................................................................................................ 93

Chapter 3. HUMAN CLAY MODELS VERSUS CAT DISSECTION: HOW THE SIMILARITY BETWEEN THE CLASSROOM AND THE EXAM AFFECTS STUDENT PERFORMANCE..................................................................................................................34
  Introduction....................................................................................................................... 34
  Method............................................................................................................................... 37
  Results............................................................................................................................... 44
  Discussion........................................................................................................................ 53
  Conclusion......................................................................................................................... 56
  References......................................................................................................................... 57

Chapter 4. MODELING HUMAN ANATOMY FROM CLAY VERSUS STUDYING PROSECTED HUMAN CADAVERS: AN ANALYSIS OF TWO APPROACHES TO INTRODUCTORY HUMAN ANATOMY..................................................................................59
  Introduction....................................................................................................................... 59
  Method............................................................................................................................... 62
  Results............................................................................................................................... 68
  Discussion........................................................................................................................ 78
  Conclusion......................................................................................................................... 81
  References......................................................................................................................... 83
Appendix. CAT DISSECTION VS. SCULPTING HUMAN STRUCTURES IN CLAY: AN ANALYSIS OF TWO APPROACHES TO UNDERGRADUATE HUMAN ANATOMY LABORATORY EDUCATION………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………
LIST OF TABLES

Chapter 1

Table 1. General description of studies reporting no significant effects on student performance between traditional animal or dissection based approaches and alternative lessons that decrease vivisection or dissection. 7-8

Table 2. Summary statistics, power, and 95% confidence intervals of studies detecting no significant treatment effects between traditional and alternative approaches to instruction. 9

Table 3. Sample sizes of studies that reported detecting no significant treatment effects between traditional and alternative approaches to instruction, but did not report means and/or standard deviations. 10

Table 4. General description of studies reporting that student performance increased in an alternative non-animal or non-dissection based lesson compared to a traditional animal or dissection based lesson. 12-13

Table 5. General description of studies reporting that student performance increased in a traditional animal or dissection based lesson compared to an alternative lessons that decrease animal use or dissection time 19

Chapter 2 (see Appendix)

Table 1. Experimental design 88

Table 2. Attitude toward using preserved animals in human anatomy teaching laboratories 89

Table 3. Attitude toward the importance of dissection when learning functions of a structure 89

Table 4. Attitude toward taking a future human anatomy course 90

Chapter 3

Table 1. Subsets of higher order anatomy questions used on the muscular system laboratory exam. 41

Table 2. Attitude toward using human cadavers vs. anatomy models in human anatomy teaching laboratories 52
Table 3. Attitude toward the importance of dissection in human anatomy teaching laboratories when learning the functions of a structure

Table 4. Attitude toward human anatomy courses

Chapter 4

Table 1. Actual α reliability coefficients ($\rho_{xx}$) based on the number of questions in each subset, and estimated Spearman-Brown α reliability coefficients ($\rho_{yy}$) if there were 10 questions in each subset.

Table 2. Attitude toward using human cadavers vs. anatomy models in human anatomy teaching laboratories

Table 3. Attitude toward the importance of dissection in human anatomy teaching laboratories when learning the functions of a structure

Table 4. Attitude toward human anatomy courses
Chapter 2 (see Appendix)

Fig. 1. High-quality plastic human model (~48 in. tall) used in all human anatomy laboratory sections. 87

Fig. 2. Maniken figures used by students in the human-clay sculpting treatment group. Groups of 2–3 students built the muscular system onto a half skeleton model. 88

Fig. 3. A portion of the muscular system sculpted onto a Maniken figure. 88

Fig. 4. A student sculpture of the gastrointestinal system and hepatic portal veins. 88

Fig. 5. Clay sculpting tools used by anatomy laboratory students. 89

Fig. 6. Students in the human-clay sculpting treatment group worked together in groups of 2–3 during the laboratory sessions. 89

Fig. 7. Laboratory exam scores (means ± SE) before and after the experimental portion of the course. For these topics, students had identical laboratory experiences and took identical laboratory exams. There were no significant differences in student performance (exam 1: P = 0.11, exam 4: P = 0.23). 90

Fig. 8. Laboratory exam scores (means ± SE) for lower-order questions (i.e., “Identify the pinned structure.”) during the experimental portion of the course. On the muscular system laboratory exam, students in the human-clay sculpture treatment group performed significantly better than students in the cat dissection control group when answering identical questions using high-quality plastic human models (P < 0.000001) and when answering similar questions pinned on human-clay sculptures vs. cats (P = 0.012). On the gastrointestinal/cardiovascular system (GI & CV sys) exam, the scores of students in the human-clay sculpting treatment group were again significantly higher than those of students in the cat dissection control group when answering similar questions pinned on human-clay sculptures vs. dissected cats (P < 0.000001). 90

Fig. 9. Laboratory exam scores (means ± SE) for higher-order questions (e.g., identify an injured muscle based on a painful limb movement) during the experimental portion of the course. All students took identical written exams on applied human anatomy. On both exams, students in the human-clay sculpting treatment group performed significantly better than the students in the cat dissection control group (exam 2: P < 0.000001, exam 3: P = 0.007). 91
Chapter 3

Fig. 1. Three combined lecture exams (mostly higher order questions) and two combined laboratory exams (mostly lower order questions) taken by all students served as a control. There were no significant differences between the control exam scores (mean ± SE) of students assigned to any of the three treatment groups in the lecture (P=0.85), or in the laboratory (P=0.41).

Fig. 2. Laboratory exam scores (mean ± SE) for three types of higher order human anatomy questions. 1) When asked to identify structures on a human cadaver diagram, students in the human clay sculpting treatment performed significantly better than those in both the cat dissection treatment without a handout (P < 0.000001), and with a handout (P < 0.000001). There was no difference between the two cat dissection treatments (P = 0.79). 2) When asked to identify human muscles seen in a transverse section diagram, students in the human clay sculpting treatment again performed significantly better than those in both the cat dissection treatment without a handout (P = 0.009), and with a handout (P = 0.01). There was no difference between the two cat dissection treatments (P = 0.97). 3) When asked to analyze a functional anatomy question on muscle action of a human, there were no significant differences between any of the three comparisons: clay sculpting to cat dissection-no handout (P = 0.25), clay sculpting to cat dissection w/handout (P = 0.35), nor between the two cat treatments (P = 0.88).

Fig. 3. Laboratory exam scores (mean ± SE) for two types of higher order cat anatomy questions. 1) When asked to identify structures on a cat external anatomy diagram, students in the human clay sculpting treatment performed significantly worse than those in both the cat dissection treatment without a handout (P < 0.000001), and with a handout (P < 0.000001). There was no difference between the two cat dissection treatments (P = 0.68). 2) When asked to analyze a functional anatomy question on muscle action of a cat, students in the cat dissection treatment that did not use a handout performed significantly better than their classmates assigned to the human clay sculpting treatment (P = 0.02), but no other significant treatment effects were detected between either the two cat dissection treatments (P = 0.34), or cat dissection using a handout versus the human clay sculpting treatment (P = 0.21).
Fig. 1. A single digestive system laboratory exam taken by all students served as a control. There were no significant differences between the control exam scores (mean ± SE) of students assigned to the two treatment groups (P = 0.72) or among the sections taught by the three instructors involved in the experiment (P = 0.41).

Fig. 2. Laboratory exam scores (mean ± SE) for two types of lower order questions. Students in both treatments were required to identify the names and actions of the same muscles pinned on identical high quality plastic models. There were no significant treatment effects on identical muscle identification questions (P = 0.33) or on identical muscle action questions (P = 0.86).

Fig. 3. Laboratory exam scores (mean ± SE) for two types of low order sister questions. For these sister questions, students assigned to the human cadaver group identified a muscle name or muscle action covered in the classroom exercise that was pinned on a human cadaver, and students assigned to the human clay sculpture group identified the same muscle or action pinned on a human clay sculpture. There were no significant treatment effects on students’ ability to identify the names of muscles on these sister questions (P = 0.92), however students in the human clay sculpting treatment were able to identify significantly more muscle actions pinned as sister questions than students from the human cadaver group (P = 0.04).

Fig. 4. Laboratory exam scores (mean ± SE) for two types of cross-over low order questions. For these cross-over questions, students assigned to the human cadaver group identified a muscle or action covered in the classroom exercise that was pinned on a human clay sculpture, while students assigned to the human clay sculpture group identified the same muscle or action pinned on a human cadaver. For the cross-over questions, there were no significant treatment effects on students’ ability to identify the names of muscles (P = 0.67), or muscle actions (P = 0.87).
Fig. 5. Laboratory exam scores (mean ± SE) for two types of higher order questions. Students were asked to identify novel anatomic features on models or drawings (origins and insertions of muscles not studied in class, or the names of known muscles presented from a previously unseen point of view, such as a cross section through a limb), and also asked to deduce the actions of muscles they had never encountered in class. Students in the prosected human cadaver treatment were able to identify more novel structures than students in the human clay sculpting treatment, however the difference was not statistically significant (P = 0.20). There were also no significant treatment effects on students’ ability to deduce the actions of novel muscles (P = 0.60).
PREFACE

This dissertation began with a simple question, is a cat dissection an effective way to teach an undergraduate human anatomy laboratory course? Dissecting preserved cat specimens is a common activity at institutions that do not have the resources or facilities to maintain a human cadaver teaching laboratory. Dissecting preserved cat specimens is an economical alternative for the institution, and provides students the opportunity to explore and observe the anatomy of a biologic specimen. Many cat body systems, such as the circulatory, digestive, nervous, and urinary systems are structurally very similar to those found in humans. Disadvantages of a cat dissection experience in a human anatomy laboratory course include the fact that other body systems, such as the skeletal and muscular systems, are structurally and often functionally quite different from those of humans. This presents a challenge to students who wish to learn human anatomy from non-human specimens.

Chapter one of the dissertation reviews the available literature where dissection, vivisection, or some other “traditional” animal or tissue based laboratory experience was compared to a non-animal or non-dissection oriented approach, by measuring student performance or student attitudes. There is no consistent methodology to these experiments, and the reported results are varied.

Chapter two describes an experiment where a traditional cat dissection lesson was compared to an alternative lesson. In the alternative lesson, students sculpted anatomic structures from clay and “built” body systems one organ at a time rather than performing a cat dissection. Students in the alternative treatment performed better than students in the dissection treatment on tests which included both higher and lower order human anatomy questions, but there were some confounding variables that complicated interpretation of the data. While the success of the
students in the clay sculpting treatment may have been due to an inherent superiority of a clay sculpting approach, it may also have been due to the fact that the clay sculptures looked more human than a preserved cat, and this similarity may have helped the students in the clay sculpting treatment answer human anatomy questions more successfully than their classmates in the cat dissection treatment. The students in the clay sculpting treatment also used a supplemental handout not available to students in the dissection treatment. Finally, the clay sculptures were by their nature less complex than a biologic organism, and this simplicity of representation may have helped the students in the clay sculpting treatment to learn anatomic structures more easily. The experiments described in chapters three and four were designed to address these issues more fully. The research described in this dissertation will help anatomy instructors make informed decisions about their curricula.
Chapter 2:


used with permission from the American Physiological Society.

See Appendix.

The author is most grateful for the opportunity to work with an outstanding group of faculty at The Pennsylvania State University. Thank you Professors Richard Cyr, Peggy Van Meter, Simon Gilroy, and Robert Mitchell for your mentorship and unfailing encouragement; I cannot imagine a better dissertation committee. Thanks also to all of the outstanding anatomy teaching assistants and students at Penn State, as well as my family, Lori, Joel, and Alex. I am also indebted to the life science faculty and students at Mohawk Valley Community College in Utica, New York for their generosity and assistance. Thank you Professors Robert Jubenville, Salvatore Drogo, and Donald Kelly. Finally, I wish to especially thank Professor William Perrotti of Mohawk Valley Community College, who also served on my dissertation committee: without your friendship, enthusiasm, and support, this dissertation would never have been possible.
CHAPTER 1

The role of dissection, vivisection, models, and technology in life science classrooms.

INTRODUCTION

When an educator considers using an animal based lesson in a classroom, the teacher is faced with both ethical and pedagogic decisions. Is it moral to sacrifice an animal to meet an educational goal? Is the animal based lesson the most effective way to teach the course’s concepts? Two extremes in opinion exists; one posits that in no circumstance is the use of animals justified to achieve a pedagogical goal, while the other extreme posits no ethical consideration whatsoever for animal use in the classroom. The majority of educators fall somewhere in the middle of these extremes, and many educators would like to minimize classroom use of animals, but only if alternatives are equally effective in meeting instructional goals. Unfortunately, the available research data on this topic is limited and sometimes conflicting. In this review, we examine the design, results, and conclusions of studies comparing traditional animal or dissection based lessons to alternative lessons that decrease animal use or dissection time, with the goal of providing guidance to educators who wish to evaluate the use of animals in their classrooms.

Some reviewers suggest that most of the alternative approaches that eliminate or reduce the need for animals have been shown to be equally, or even more, effective than animal based laboratory exercises (26, 38). The Humane Society of the United States publishes a compelling list of Comparative Studies of Dissection and Other Animal Uses (2) which includes a total of thirty-three studies under the headings “demonstrate equal or comparable student performance between dissection and alternative methods” and “alternatives were more effective instructional aids than dissection”. Only two studies suggesting an animal based approach may be more
Accordingly, it is important to review the design, the statistical analysis, and the results of each study before drawing conclusions from the body of work. Furthermore, the question: “can a non-animal alternative replace an animal based classroom exercise” creates a false dichotomy. The answer will likely depend on the objectives of the lesson and the nature of the specific exercises. One must also be careful not to over generalize the results from a relatively small number of studies using very different student populations, experimental designs, and educational approaches. In this review, we focus on research papers that attempted to measure the effectiveness of two or more approaches to science or clinical education, in an effort to help educators make informed decisions about their curricula.

METHOD

The studies included in the Results and Discussion section of this review all attempt to compare the efficacy of a “traditional” animal or dissection-oriented classroom lesson with an “alternative” approach utilizing computer simulations, models, or some other method that reduces vivisection or dissection. An electronic search of the ISI Web of Science data base (http://apps.isiknowledge.com/) using the search terms, “animal”, “alternative”, “dissection”, and “education” was conducted. The Web of Science database includes the following resources: Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index. The search often led to additional research and review papers that are discussed here. We exclude some studies occasionally cited by other reviewers that simply describe a non-
animal approach without making any comparisons (6, 30), that do not explicitly report student
performance data when comparing treatments (22, 35), or that replace a lecture with educational
technology or a dissection with a prosection (24).

The studies that are presented below can be divided into two groups: those detecting no
statistically significant treatment effects, and those that detected one or more significant
treatment effects. When reporting their results, authors almost always report the likelihood that
they may incorrectly conclude that there are significant treatment effects (reject a true null
hypothesis, probability = α, also referred to as a type I statistical error). However, authors often
do not consider the likelihood of incorrectly concluding that there are no statistically significant
treatment effects (accepting a false null hypothesis, probability = β, also referred to as a type II
statistical error). When one detects no treatment effect, it is tempting to conclude that the two
teaching methods must be equally effective. However, if the statistical power, or the ability to
detect a significant treatment effect, of the study is low, then one risks concluding that two
treatments are the same when in fact, they are not (committing a type II error). The most
common reason for a lack of statistical power is that too few participants were included in the
experiment; a problem which confounds many of the studies reviewed here.

For this reason, when the data were available, we performed a power analysis on all
studies that did not detect a treatment effect. A power analysis is a statistical technique that
determines whether or not a study can confidently accept a null hypothesis. If β is the probability
of overlooking a statistically significant difference between two treatments, then 1 - β, or power,
is the probability of accurately detecting a significant treatment effect. Without adequate power,
an investigator is more likely to detect no treatment effects, and thus reach inaccurate
conclusions. The statistical power of a study is determined by the sample size, the levels of α and
\( \beta \) set by the investigator, and the treatment effect size. Treatment effect size is simply the difference between the mean values of two groups (ie. the control and experimental treatment groups) expressed as a proportion of the overall standard deviation. When estimating treatment effect size using Cohen’s d, the following formula was used.

\[
\text{Cohen’s } d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}
\]

where \( \hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \)

The statistical power of individual studies was estimated using the reported means and standard deviations to estimate treatment effect size (Cohen’s d), and then with information on sample size, referring to the power tables available in Cohen (5), for pair-wise comparisons using a t-test. No power analysis was performed on studies detecting significant treatment effects, since those studies obviously had adequate power to reject a null hypothesis.

It has been suggested that confidence intervals are preferable to estimated power when considering studies that did not reject the null hypothesis (31), so ninety-five percent confidence intervals for the difference between the treatment means \( \bar{X}_1 - \bar{X}_2 \) were also calculated. If one theoretically repeats an experiment many times, and calculates the 95% confidence interval of the difference between two means each time, then 95% of the calculated intervals should contain the true difference between the two means. By calculating one such interval based on experimental data, one can say that there is a 0.95 probability that the true mean is contained somewhere within the calculated interval. When such an interval includes the value zero, which indicates that there is no difference between the two means, an investigator will usually accept a null hypothesis.
Given that sample sizes were often very small, 95% confidence intervals based on a t-distribution were calculated using the following equation.

\[
95\% \text{ confidence interval} = \bar{X}_1 - \bar{X}_2 \pm t_{df,0.05/2} \hat{\sigma}_{\bar{X}_1 - \bar{X}_2}
\]

The value \(\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}\) is an estimate of the overall standard deviation of a study. By multiplying this by the critical t value for a two tailed t-test (\(\alpha = 0.05, \text{df} = n_1 + n_2 - 2\)), one can calculate the ninety-five percent confidence interval of \(\bar{X}_1 - \bar{X}_2\).

RESULTS AND DISCUSSION

*Studies reporting no significant effects on student performance between traditional animal or dissection based approaches and alternative lessons that decrease animal use or dissection time*

The eleven studies summarized in Table 1 all reported detecting no significant treatment effects between animal or dissection oriented laboratory exercises and alternative approaches that decreased animal use. The studies were conducted in veterinary school, graduate, undergraduate, and high school classrooms with learning objectives ranging from surgical training, to organ physiology, to medical anatomy using human cadavers, to very introductory anatomic explorations using frog dissection. With such a wide range of objectives, instructors used many different types of assessment tools, such as videotaping and evaluating live surgeries, written examinations, laboratory practical exams, and multiple choice questions to quantify student performance.

Seven of the studies (noted in Table 1) included some form of mean test scores or grades, standard deviation, and sample size in their results, making a power analysis possible. One of these seven, Fawver et. al. (14), did not detect a significant treatment effect in one comparison described within the study, but did detect a significant treatment effect in a separate comparison.
within the same study. In this case, the non-significant treatment effect is discussed here, and the significant treatment effect will be discussed later (see Table 4).
<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Population</th>
<th>Comparison Made (treatments)</th>
<th>Measurement Instruments of Student Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bauer et. al. (3)†</td>
<td>veterinary medical students</td>
<td>surgical training during a small animal surgery rotation based on clinical caseload experience only vs. clinical caseload + practice surgeries on dog cadavers</td>
<td>course grades for animal surgery rotation based on subjective scores of students’ skills and objective written tests</td>
</tr>
<tr>
<td>Carpenter et. al. (4)*</td>
<td>veterinary medical students</td>
<td>surgical training on cadaver vs. anesthetized dogs</td>
<td>time to complete live animal surgery, videotaped review of live animal surgery, strength of sutured incisions</td>
</tr>
<tr>
<td>Dewhurst et. al. (11)*</td>
<td>undergraduate physiology students</td>
<td>intestinal physiology exercise using rat intestine vs. computer program</td>
<td>identical pre-test and post-tests of 50 questions (mostly short-answer)</td>
</tr>
<tr>
<td>Downie and Meadows (12)†</td>
<td>undergraduate biology students</td>
<td>anatomy exercise comparing rat dissection to alternative exercise based on models and diagrams</td>
<td>written examinations mostly testing factual recall</td>
</tr>
<tr>
<td>Fawver et. al. (14)*</td>
<td>veterinary medical students</td>
<td>cardiovascular physiology exercise using anesthetized dogs vs. a videodisc program</td>
<td>13 item multiple choice/short answer test</td>
</tr>
<tr>
<td>(also listed in Table 4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greenfield et. al. (18)*</td>
<td>veterinary medical students</td>
<td>surgical training on soft tissue models vs. anesthetized dogs</td>
<td>videotaped review of live animal surgery (no mean or sd given), surgical skills grades and overall grades of subsequent small animal surgery rotation</td>
</tr>
</tbody>
</table>

*Table continued on next page*
<table>
<thead>
<tr>
<th>Study</th>
<th>Participants</th>
<th>Exercise Description</th>
<th>Assessment Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guy and Frisby (20)†</td>
<td>undergraduate pre-nursing and allied health students</td>
<td>human anatomy laboratory exercise based on prossected human cadavers vs. videotape exercise</td>
<td>10 point weekly quizzes, 4 mid-term exams, comprehensive final (no one tested on cadavers)</td>
</tr>
<tr>
<td>Kinzie et al. (27)*</td>
<td>high school biology students</td>
<td>anatomy exercise comparing interactive videodisc (IVD) prelab + frog dissection vs. videotape prelab + frog dissection vs. frog dissection only vs. IVD pre-lab only</td>
<td>10 frog anatomy id questions from diagrams, 11 anatomy questions pinned on frogs, 9 multiple choice questions about dissection procedures</td>
</tr>
<tr>
<td>Pavletic et al. (39)†</td>
<td>veterinary medical students</td>
<td>surgical training on cadaver vs. anesthetized dogs</td>
<td>post-graduation survey of subjects’ employers regarding subjects’ competency as veterinarians</td>
</tr>
<tr>
<td>Prentice et al. (41)*</td>
<td>professional and graduate students</td>
<td>human anatomy laboratory exercise based on human cadaver dissection vs. stereoscopic slides</td>
<td>5 laboratory examinations, each comprised of 40 identification questions :10 human cadaver, 24 stereoslides, 6 bones and x-rays</td>
</tr>
<tr>
<td>Strauss and Kinzie (43)*</td>
<td>high school biology students</td>
<td>anatomy exercise comparing frog dissection to interactive videodisc simulation</td>
<td>10 frog anatomy id questions from diagrams, 10 anatomy questions pinned on frogs, 5 multiple choice questions about dissection procedures</td>
</tr>
</tbody>
</table>

*Studies included mean test scores or grades, standard deviation, and sample size in their results, making a post-hoc power analysis possible. †Studies that did not report standard deviations or other summary statistics.
A statistical power index of 0.8 is both defensible and widely accepted as the criterion to confidently conclude that there are no measurable treatment effects (accept a null hypothesis) (36). When statistical power was calculated, all five of the seven studies have low power (Table 2) based on performing pair-wise comparisons via t-tests. When one calculates 95% confidence intervals for the treatment differences ($\bar{X}_1 - \bar{X}_2$), the value zero falls within just one standard deviation of the treatment difference for all seven studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size ($n_1, n_2$)</th>
<th>Effect size (Cohen’s d)</th>
<th>power</th>
<th>$\beta$</th>
<th>95% confidence interval of $\bar{X}_1 - \bar{X}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carpenter et. al. (4)</td>
<td>12, 12</td>
<td>σ not reported</td>
<td>0.6</td>
<td>0.4</td>
<td>could not be calculated</td>
</tr>
<tr>
<td>Dewhurst et. al. (11)</td>
<td>6, 8</td>
<td>0.6</td>
<td>0.2</td>
<td>0.8</td>
<td>-9.3 &lt; $\mu_1 - \mu_2$ &lt; 15.7</td>
</tr>
<tr>
<td>Fawver et. al. (14)</td>
<td>42.5, 42.5</td>
<td>0.6</td>
<td>0.8</td>
<td>0.2</td>
<td>-2.1 &lt; $\mu_1 - \mu_2$ &lt; 3.9</td>
</tr>
<tr>
<td>Greenfield et. al. (18)</td>
<td>a 18, 18</td>
<td>0.3</td>
<td>0.1</td>
<td>0.9</td>
<td>-0.5 &lt; $\mu_1 - \mu_2$ &lt; 0.6</td>
</tr>
<tr>
<td></td>
<td>b 18, 18</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>-1.4 &lt; $\mu_1 - \mu_2$ &lt; 3.1</td>
</tr>
<tr>
<td>Kinzie et. al. (27)</td>
<td>15, 15</td>
<td>1.0</td>
<td>0.8</td>
<td>0.2</td>
<td>-1.9 &lt; $\mu_1 - \mu_2$ &lt; 5.5</td>
</tr>
<tr>
<td>Prentice et. al. (41)</td>
<td>23, 16</td>
<td>0.2</td>
<td>0.1</td>
<td>0.9</td>
<td>-4.2 &lt; $\mu_1 - \mu_2$ &lt; 5.2</td>
</tr>
<tr>
<td>Strauss and Kinzie (43)</td>
<td>a 8, 9</td>
<td>0.2</td>
<td>0.1</td>
<td>0.9</td>
<td>-3.6 &lt; $\mu_1 - \mu_2$ &lt; 4.3</td>
</tr>
<tr>
<td></td>
<td>b 8, 9</td>
<td>0.8</td>
<td>0.3</td>
<td>0.7</td>
<td>-2.2 &lt; $\mu_1 - \mu_2$ &lt; 4.7</td>
</tr>
</tbody>
</table>

Different comparisons within one study noted as $a$ and $b$.

*Power reported by authors.

It should be noted that the studies by both Fawver et. al. (14) and Kinzie et. al. (27) had adequate statistical power to confidently conclude that there was no statistically significant difference in at least one comparison. It is difficult to generalize Fawver et. al.’s (14) results without a detailed description of their testing instrument (Table 1). In the study by Kinzie et. al. (27), no significant effect on student performance was reported when directly comparing a frog dissection experience to a dissection simulation program using an exam comprised primarily of lower order frog anatomy identification question (Table 1). Kinzie et. al. (27) did report significant treatment effects when a dissection experience and a simulation experience
supplement one another. By performing an analysis of covariance followed by planned comparisons, Kinzie et. al. did have enough statistical power to find that students who used the simulation program and then dissected a frog achieved higher test scores than students that either performed a dissection only or used the simulation program only. While not addressing the question of which approach, dissection alone or a simulation alone, results in greater student performance, this result does demonstrate the effectiveness of supplementing one experience with another.

The four remaining studies described in Table 1 also reported detecting no significant differences in student performance between some type of traditional animal or dissection oriented approach versus a non-animal/non-dissection alternative (3, 12, 20, 39). Unfortunately, none of these four studies reported standard deviations, and three studies did not report mean values for all comparisons (3, 20, 39), making a detailed power analysis impossible. Nonetheless, one can attempt to make some general estimates about power based on the sample sizes of each study (Table 3).

Table 3. Sample sizes of studies that reported detecting no significant treatment effects between traditional and alternative approaches to instruction, but did not report means and/or standard deviations.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size (traditional, alternative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bauer et. al. (3)</td>
<td>28, 29</td>
</tr>
<tr>
<td>Downie and Meadows (12)</td>
<td>2656, 56 *</td>
</tr>
<tr>
<td>Guy and Frisby (20)</td>
<td>141.5, 141.5 *</td>
</tr>
<tr>
<td>Pavletic et. al. (39)</td>
<td>36, 12 *</td>
</tr>
</tbody>
</table>

* Estimated sample size based on authors’ descriptions.

Depending on treatment effect size, at least two of the studies, Bauer et. al. (3) and Pavletic et. al. (39), likely have an insufficient sample size for the difference they attempted to
measure. If the treatment effects are small and a two tailed t-test is conducted with alpha set to 0.05, then approximately 400 students are needed in each treatment group to obtain a power value of 0.8 (5). With a medium effect size, sixty-four students in each treatment group are necessary to achieve the same statistical power. In addition to the issues described above, two of the studies, Downie and Meadows (12) and Pavletic et. al. (39), allowed students to choose their treatment group, thus violating the random assignment of subjects assumption of most inferential statistical tests. Generally small sample sizes, methodological weaknesses, and a lack of summary statistics preclude an educator from confidently drawing any conclusions from most of the studies described in Table 1.

Studies reporting significant effects on student performance in favor of an alternative non-animal or non-dissection based lesson

The nine studies listed in Table 4 all report some type of significant treatment effect in favor of a non-animal/non-dissection oriented approach. These studies were also conducted in veterinary, medical, undergraduate, and high school classrooms with a range of learning objectives and assessment tools similar to those described earlier. Six of these nine compared some type of technology based alternative to an animal based lesson. Fowler and Brosius (15) and Henman and Leach (21) both replaced an animal laboratory with a film or video. Erickson and Clegg (13), Lilienfield and Broering (32), Predavec (40), and Fawver et. al. (14) all replaced an animal laboratory with an interactive computer program. All four computer programs included animations and three included a quizzing feature (13, 32, 40).

While all of the alternative approaches appear to be thoughtful and well-designed exercises, three of the six studies investigating the efficacy of technology based alternatives were particularly vague about their data analysis. Erickson and Clegg (13) focus on students’ learning preferences, and only state in the second to last sentence that computer based learning “improves
Table 4. General description of studies reporting that student performance increased in an alternative non-animal or non-dissection based lesson compared to a traditional animal or dissection based lesson

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Population</th>
<th>Comparison Made (treatments)</th>
<th>Measurement Instruments of Student Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erickson and Clegg (13)</td>
<td>veterinary medical students</td>
<td>cardiovascular physiology exercise comparing animal and computer based modules</td>
<td>not reported</td>
</tr>
<tr>
<td>Fawver et. al. (14) (also listed in Table 1)</td>
<td>veterinary medical students</td>
<td>cardiac fibrillation physiology exercise using anesthetized dogs vs. a videodisc program</td>
<td>9 item multiple choice/short answer test</td>
</tr>
<tr>
<td>Fowler and Brosius (15)</td>
<td>high school biology students</td>
<td>anatomy exercise based on performing animal dissection vs. watching dissection films</td>
<td>“factual knowledge” test, motor skills test, plus three standardized exams: Sequential Tests of Educational Progress (problem solving ability), Test on Understanding Science (methods and aims of science), Facts About Science (student attitude)</td>
</tr>
<tr>
<td>Griffon et. al. (19)</td>
<td>veterinary medical students</td>
<td>surgical training on dog cadavers vs. anatomic and hemodynamic plastic models</td>
<td>evaluation of surgical skills using models, written anatomy exams, and scoring surgical skills during a ovariohysterectomy on an anesthetized dog</td>
</tr>
<tr>
<td>Henman and Leach (21)</td>
<td>undergraduate pharmacy students</td>
<td>pharmacology experiments using videotapes vs. animal organ experiments</td>
<td>short answer written pre- and post-tests</td>
</tr>
</tbody>
</table>

*table continued on next page*
<table>
<thead>
<tr>
<th>Lilienfield and Broering (32)</th>
<th>medical and graduate students</th>
<th>electrocardiography and cardiac cycle physiology exercise based on an anesthetized dog laboratory vs. a computer program</th>
<th>141 question written final examination, which included 12 five-part multiple choice questions on cardiovascular physiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olsen et. al. (37)</td>
<td>veterinary medical students</td>
<td>basic surgical skills training on anesthetized dogs vs. a hemostasis model</td>
<td>evaluation of surgical skills performed using models, and an anesthetized dog</td>
</tr>
<tr>
<td>Predevac (40)</td>
<td>undergraduate biology students</td>
<td>anatomy and physiology exercise comparing rat dissection to a computer program</td>
<td>33 question multiple choice quiz comprise of 11 text based questions, 12 diagram questions, and 10 questions pinned on dissected rats or microscope slides</td>
</tr>
<tr>
<td>Waters et. al. (45)</td>
<td>undergraduate anatomy students</td>
<td>anatomy exercises comparing cat dissection to sculpting human anatomic structures from clay</td>
<td>two 50 question laboratory exams stressing structure identification (low order questions), and two smaller exams stressing higher order questions</td>
</tr>
</tbody>
</table>
student skills in problem solving and information handling”, but offer no data nor cite any
references other than their personal experience. A 1968 study by Fowler and Brosius (15) also
lacks detailed summary statistics. They used a series of standardized tests to measure the factual
knowledge, problem solving ability, understanding of science as a field, manual laboratory skills,
and attitudes toward science of high school students performing invertebrate and vertebrate
dissections versus watching a series of anatomy films. The authors reported measuring
significant differences favoring the film treatment in students’ factual knowledge and problem
solving ability, and detected no treatment effects for the other categories. Unfortunately, Fowler
and Brosius (15) omitted any reference to average test scores, standard deviations, or even
significance criteria, making a critique of their analysis problematic. A third commonly cited
study is a meeting abstract by Henman and Leach (21), which briefly describes an experiment
comparing a pharmacology video series to animal based laboratory experiments in a course taken
by pharmacy undergraduates. They report that the students in the video treatment earned
consistently higher scores on a post-test composed of short answer questions, but there is
insufficient information available to judge if this data can be generalized to other life science
courses. The information missing from all three studies limits their utility as instructors try to
make informed curricular decisions.

The other three studies to compare a technology based to an animal based learning
experience were more descriptive. Lilienfield and Broering (32), in one of the few studies to
include exam questions in their report, showed that medical students using a computer program
answered a mix of higher and lower order questions on the cardiovascular system more
successfully than did students assigned to a laboratory utilizing an anesthetized dog. Predevac
(40) reported that undergraduate biology students using a computer program designed to teach
the reproductive anatomy of a rat, outperformed students assigned to a rat dissection experience, on a test consisting of thirty-three multiple choice questions in text, diagram, and specimen formats.

An animal versus non-animal learning experience was not the only difference between the two treatment groups in this pair of studies however. In both studies, the students assigned to the computer based lessons were able to work individually, while the other students worked in some sort of lab group, therefore, the pace the material was presented may have been have been more personalized for the students assigned to those treatments. More importantly, only the students in the computer based treatments had the benefit of extensive quizzing before, during, or after the lesson. Predevac (40) notes that the rat anatomy program includes “a bank of interactive quiz questions”. Lilienfield and Broering (32) describe that students in the computer based treatment answered pre-lesson and post-lesson quiz questions the day of the lesson. Lilienfield and Broering (32) do note that these students did not receive their scores until after the lesson had been completed, and that the quiz questions were different from the multiple choice questions used in their analysis. Regardless, the quizzing features of the software programs used in both studies provide feedback that is not available to the students assigned to the animal treatments. This feedback, independent of animal versus non-animal treatment effects, may result in improved exam scores for the students assigned to the computer program treatment. If both of these studies had incorporated some type of quizzing experience into both treatment groups, the question of an animal versus a non-animal based lesson could be more clearly addressed.

In the third technology versus animal laboratory comparison, Fawver et. al. (14), reported finding no significant differences between animal and non-animal based laboratory lessons for two separate laboratory exercises: general cardiac physiology and cardiac fibrillation. The
general cardiac laboratory exercise comparison (see Table 1) was discussed earlier. The second comparison described in the same paper, a cardiac fibrillation laboratory exercise, had high statistical power (0.9), high enough to conclude that the students in the non-animal treatment had outperformed the students in the animal based treatment if the authors had used the 95% confidence interval to test their null hypothesis ($2.4 < \mu_1 - \mu_2 < 9.6$). Unfortunately, the authors did not include a detailed description of their testing instrument, so it is difficult to know if their results can be generalized to other courses. The authors also note that students assigned to the technology treatment were permitted to schedule additional time to work on the tutorial if desired. This extra time for study and practice may explain the superior performance demonstrated by the students assigned to the non-animal treatment. If both treatment groups had equal time in the teaching laboratory, then the effectiveness of the two approaches could have been compared more directly.

The final three studies summarized in Table 4 each investigated using a model of some type to replace an animal based laboratory experience, with the goal that the model might provide a more effective learning experience with regard to the course goals. Given that two of the courses were veterinary surgery courses [Griffon et al. (19) and Olsen et al. (37)], and one was a human anatomy laboratory [Waters et al. (45)], this may seem counter-intuitive at first; however the model treatment in all three studies likely reduced the transfer distance between the learning and testing environments. When a student uses concepts learned in one setting, and applies those concepts to a different setting, transfer of learning has occurred. The term “near-transfer” describes one end of a continuum when the two experiences are similar. The term “far-transfer” describes the other end of the continuum when the two experiences are very different (16).
Transfer distance was likely manipulated when Griffon et. al. (19) used a hemostasis model of a dog’s reproductive system to replace a dog cadaver exercise. When students practiced performing an ovariohysterectomy on the model, the model’s “blood vessels” pulsed with red fluid, thus giving the students a more realistic experience than could be gained practicing on the dog cadaver. All of the students were evaluated as they performed an ovariohysterectomy on a live dog, and the students in the model treatment were ranked significantly more proficient. In another study, Waters et. al. (45), reported that students who sculpted clay models of human anatomic structures were significantly more successful answering higher order human anatomy questions than students performing a cat dissection. Neither study proposed that models are a universally superior approach, but in both cases, the models likely decreased the transfer distance between the laboratory exercise and the examination.

This conclusion is supported by Olsen et. al. (37), who compared the surgical skills of veterinary students practicing surgical techniques on a hemostasis model versus a live dog. This study compared the students in the two treatment groups on thirteen different surgical skill outcomes, and in eleven of the thirteen comparisons, detected no significant treatment effects. The two significant treatment effects demonstrated that students in the model treatment were more successful tying vascular sutures and holding surgical instruments. If models are generally superior to live animals, then one would expect the students in the model treatment to consistently perform better than students in the live animal treatment. On the other hand, if the previous two studies were measuring transfer distance effects, then one would expect there to be little or no difference between hemostasis models and live dog surgeries, since both would be a case of near transfer. While models may not be generally superior to all animal or dissection approaches, a model that increases the similarity between the learning activity and the learning
objective is likely a viable alternative because it mimics the “real world” situation more accurately than a cadaveric specimen.

Student performance increased in a traditional animal/dissection based lesson

Table 5 summarizes seven studies that reported learning outcomes demonstrating increased student performance in traditional animal or dissection based lessons. As in the previously discussed studies, there are a wide range of sample populations: from medical residents, to undergraduates, to high school students. The measurement instruments also vary with one study evaluating emergency room surgical technique, and others using essay or objective tests. Five of the studies described in Table 5 included comparisons where an animal or dissection based laboratory was replaced by a technology based animal/dissection simulation. In three of these studies [Dewhurst and Meehan (10), Leathard and Dewhurst (29), Marszalek and Lockard (33)] the authors concluded that a simulation was just as effective as an approach using animals, but a closer look at their data reveals some interesting findings.

Dewhurst and Meehan (10), combined the laboratory grades of 65 students across two different universities who used a variety of animal physiology laboratory simulation software. The authors did not detect a significant treatment effect on undergraduate student performance comparing an animal based exercise and a computer simulation using a 95% confidence interval, but they did note a treatment effect favoring the animal approach using a 90% confidence interval. By changing the confidence interval from 95% to 90%, the authors increased the $\alpha$ probability of incorrectly rejecting a true null hypothesis (committing a type I statistical error) to 0.1, but they also decreased the $\beta$ probability of incorrectly accepting a false null hypothesis (committing a type II statistical error). One confidence interval is not necessarily better than the other, but the statistical power ($1 - \beta$) of the study is greater using the 90% confidence interval.
Table 5. General description of studies reporting that student performance increased in a traditional animal or dissection based lesson compared to an alternative lessons that decrease animal use or dissection time

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Population</th>
<th>Comparison Made (treatments)</th>
<th>Measurement Instruments of Student Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custalow et. al., (9)</td>
<td>emergency medicine residents</td>
<td>surgical training comparing training video + practices on anesthetized pigs vs. video alone</td>
<td>evaluation of surgical technique on anesthetized pigs 6 months after training</td>
</tr>
<tr>
<td>Dewhurst and Meehan (10)</td>
<td>undergraduate students</td>
<td>physiology exercises using “traditional” (animal) protocols vs. computer programs</td>
<td>grades on laboratory reports</td>
</tr>
<tr>
<td>Granger and Calleson (17)</td>
<td>medical students</td>
<td>anatomy exercises comparing dissection only exercises to dissection + prossection + other classroom resources</td>
<td>grades on written (higher order concepts) and practical (structure identification) exams</td>
</tr>
<tr>
<td>Kariuki P and Paulson R (25)</td>
<td>high school biology students</td>
<td>anatomy exercise comparing earthworm and frog dissections to computer programs</td>
<td>Multiple choice structure identification test with structures pinned on animal specimens and shown in diagrams</td>
</tr>
<tr>
<td>Leathard and Dewhurst (29)</td>
<td>medical students</td>
<td>intestinal physiology exercise comparing computer simulation to a demonstration using rat intestine</td>
<td>5 assigned essay questions</td>
</tr>
<tr>
<td>Marszalek and Lockard (33)</td>
<td>middle school students</td>
<td>anatomy exercise comparing preserved frog dissection vs. computer tutorial program vs. computer simulation program + additional digital media</td>
<td>identical pre-test, immediate post-test, and delayed (3 month) post-test consisting of true-false, matching, and multiple choice questions</td>
</tr>
<tr>
<td>Matthews (34)</td>
<td>undergraduate biology students</td>
<td>anatomy exercise comparing fetal pig dissection vs. dissection simulation program</td>
<td>25 question oral exam, plus a 25 question computer based exam, where students identified structures</td>
</tr>
</tbody>
</table>
Given that changing the approach of a teaching laboratory to or from a technology alternative will likely require a significant investment of resources, it is advisable to make such a decision based on studies with greater statistical power.

In a second study, Leathard and Dewhurst (29) concluded that performance was comparable between medical students using an intestinal motility computer simulation and those students attending a tutored demonstration utilizing intestine isolated from an animal. However, when one calculates an overall mean for each treatment to compare the percentage of students giving satisfactory answers to the first four of the five questions used to assess student performance, the students assigned to the animal based exercise seem to have performed significantly better at the 95% confidence interval ($4.3 < \mu_1 - \mu_2 < 17.1$). Students in the animal based treatment also answered the fifth assessment question more successfully than their classmates in the alternative treatment, but the authors noted that the students in the animal based treatment did not have time to cover that material as part of the planned animal based laboratory exercise, so that part of the animal based lesson was simply summarized by an instructor. Since the concepts covered by the fifth question were no longer part of an animal versus non-animal comparison, it was not included in the calculation of the confidence interval described above.

Even though the authors conclude that student performance “is comparable for both groups”, the data presented suggests that the students assigned to the animal based exercise outperformed the students in the computer simulation treatment.

In the third study, Marszalek and Lockard (33) compared a middle school frog dissection laboratory exercise to a dissection tutorial program and a dissection simulation program. Both alternatives were interactive, with the tutorial program built around videos of actual dissections, and the simulation built around an animated dissection exercise. Students were administered a
pre-test, an immediate post-test, and a delayed post-test (three months after the exercise), which consisted of true-false, matching, and multiple choice questions. Students assigned to the dissection treatment demonstrated a significantly higher gain from pre-test to immediate post-test than students assigned to the simulation treatment. When pre-test to delayed post-test scores were examined, the students in the dissection group performed significantly better than their classmates in both the simulation and tutorial treatments. However in their discussion, Marszalek and Lockard focus on changes from immediate post-test to delayed post-test, and detecting no significant treatment effects there, conclude that: “when looking at knowledge retained over the long term, this study’s results support the view that alternatives to dissection are as effective as the actual practice of physical dissection.” If an instructor is concerned with both how much of the lesson is learned and retained over time though (pre-test to delayed post-test gain), the dissection treatment was clearly the superior approach.

The results reported by the four remaining studies summarized in Table 5 are more straight forward. Kariuki and Paulson (25) compared an earthworm and frog dissection to computer-animated earthworm and frog dissections with high school biology students. When the performance of each treatment group was compared using a test comprised of multiple choice questions, pinned specimens, and diagrams, the students who performed an actual dissection earned significantly higher test scores than the students who performed a simulated dissection. Matthews (34) performed a similar experiment comparing a fetal pig dissection to a computer simulation in an undergraduate biology course. While there were no statistically significant treatment effects when students took a computer based exam (the author does not describe this exam), students in the dissection treatment performed significantly better than their classmates in
the simulation treatment on an oral exam which emphasized the name and function of anatomic structures.

Without being able to review the specific lessons and examinations investigators used in their research, it is difficult to offer a simple explanation as to why the studies listed in table 5 report that an animal or dissection based experience is more effective than a non-animal or non-dissection lesson. It is clear though that animal and dissection experiences often help students answer both lower and higher order anatomy and physiology questions more successfully than alternative lessons.

In a different type of study, Custalow et. al., (9) reported that a surgical procedure video supplemented with an animal based surgery experience is more effective than video alone when training physicians to perform resuscitative surgical procedures (as measured by an evaluation of their surgical technique six months post training). However, as demonstrated by Kinzie (27) earlier in this review, the finding of Custalow et. al. (9) could simply be demonstrating an advantage of including any sort of supplemental experience in a classroom, rather than an effect directly attributable to an animal based experience.

The final study summarized in Table 5 by Granger and Calleson (17) compare two groups of medical students enrolled in an anatomy course during different years. The students enrolled in the course during the first year primarily performed cadaver dissections in their anatomy laboratory. The students enrolled in the same anatomy course the following year performed a limited amount of cadaver dissection supplemented with observing prosected cadavers plus completing multimedia web tutorials. Each year, the curriculum was divided into four curricular blocks, with each block focusing on different regions of the body. Student achievement was measured using grades on both a written and a practical exam given during
each curricular block. Written exams were composed of mostly higher order questions requiring students to apply what they had learned in the anatomy laboratory. Practical exams were mostly lower order questions that asked students to identify anatomic structures. All exams were non-cumulative and administered at the end of each curricular block.

Examining student performance on the first two block exams, students from the first year who focused on dissections earned significantly higher written exam scores than did the class the second year who performed a limited number of dissections. Scores on practical exams were mixed. It is difficult to interpret the data from the last two curricular blocks, since non-identical exams were given across the two years of the study. However, the positive effect of increased dissection time on students’ ability to answer higher order questions suggests that a dissection experience cannot necessarily be replaced by other activities.

*Studies measuring the effect of a traditional animal or dissection based experience and an alternative non-animal or non-dissection experience on student attitudes*

When investigators survey students about their attitudes toward animal use and dissection in experiments where traditional and alternative approaches were compared, there is evidence that students prefer the instructional approach they experienced. Olsen et. al. (37) reported statistically significant treatment effects between veterinary students practicing surgical techniques using an anesthetized animal versus a hemostasis model. Strauss and Kinzie (43) demonstrated that over time, high school students who performed a frog dissection and those who worked through a simulated dissection program both preferred the treatment they experienced. In each of these studies, the differences were statistically significant. Studies comparing surgical training on cadaver versus live dogs [Carpenter et. al. (4)], and studying anatomy via cat dissection versus building clay models [Waters et. al. (45)] show similar but statistically non-significant trends in student preference. Only Kinzie et. al. (27), in their...
studying comparing combinations of frog dissection and a frog simulation program, reported no treatment effect on student preference toward animal use.

When students were allowed to choose or were assigned to an animal or non-animal oriented lesson consistent with their preferences, it is not surprising that the lesson reinforced their preferences. In Downie and Meadow’s (12) study were students could opt-out of a rat dissection experience and study a model instead, students preferences mirrored their treatment choice. In Dewhurst et. al.’s (11) study comparing an intestinal physiology demonstration using rat intestine versus a computer simulation, it is not clear if students were permitted to choose their treatment group or not, but the authors report clear differences in student attitudes toward animal use and non-animal alternatives that are consistent with their treatment group prior to the beginning of the experiment. After the experiment, students’ preferences were also consistent with the approach used in their treatment group.

Other studies measured student attitudes toward an animal or dissection based lesson versus a non-animal or non-dissection experience by exposing a single group of students to both approaches, and then surveying the group about their attitudes. Results are mixed with some authors reporting trends or comments in favor of an animal or dissection based approach (7, 8, 17, 23), and others in favor of a non-animal or non-dissection alternative approach (13, 42).

CONCLUSION

Are non-animal or non-dissection based alternatives superior, or at least just as effective, as more traditional animal or dissection based approaches to education? Issues such as a lack of statistical power, failure to present summary statistics, as well as a wide range of course goals, measurement instruments, and sample populations (see Tables 1, 4, and 5) make the studies
discussed here difficult to interpret, and lend weight to the call by the National Science Teachers Association for additional research (1).

Most of the studies detecting no treatment effect had insufficient statistical power. Statistical power can be improved by increasing sample size and decreasing within treatment variance (discussed below). When authors detect no treatment effect in an experiment with too little power, they risk committing a type II statistical error, in which the null hypothesis is falsely accepted. That is, researchers conclude the experimental treatments were equally effective when the design of the study limits the possibility of detecting a true difference (11, 14, 18, 27, 39, 41, 43). Given the high probability of accepting a false null hypothesis in most of the studies that did report summary statistics (Table 2), it seems unwise to assume that there is not a similar risk in some or all of the studies that omitted summary statistics from their results (Table 3).

Little or no opportunity to evaluate test questions or review course goals further complicates evaluating this literature. The test questions likely reflect the authors’ goals for their courses, but one cannot expect all course goals to be the same. The studies reviewed here were conducted in high school, undergraduate and graduate/professional school classrooms. The characteristics of the students, and the learning objectives of the courses in these settings are likely very different. Some introductory courses may emphasize memorizing the names and functions of structures. In this case, the opportunity for quizzing and drilling provided by instructional technology alternatives may be very helpful. However, other courses may take a more functional approach, requiring students to use information from the classroom to solve novel problems they did not have a chance to practice, and still other courses, such as those at a veterinary or medical school, may be almost entirely focused on students mastering a technique. These courses may require the students to have the opportunity to explore something more
realistic than a computer image. If virtual reality programs are found to be effective alternatives in medical school surgical courses, one cannot generalize such a finding to conclude that a web based dissection simulation is going to be equally effective in a high school or undergraduate classroom, especially when the goals of these courses may range from the identification of structures, to teaching biomechanics and physiology. When authors neglect to report their learning objectives and the testing instruments they use to evaluate whether or not the students achieve those objectives, it becomes difficult to draw conclusions about specific animal/dissection vs. non-animal/non-dissection approaches.

Given the generally low statistical power of the studies detecting no significant treatment effects, and the difficulty to generalize from the studies that do detect significant treatment effects, it seems unwise to conclude that non-animal/non-dissection alternatives are a superior, or even an equally effective approach to instruction. Likewise, it is not possible to conclude that dissection experiences are similar to alternatives. More statistically valid and detailed studies are necessary before one can reach any such conclusion.

RECOMMENDATIONS

Educators and administrators will have a much easier time determining how a study may inform their teaching if authors address statistical power and present more information in their studies. Since it is quite possible that an alternative may be just as effective as a more traditional approach, statistical power must be addressed a priori. If there have been similar studies, treatment effect sizes can be estimated if the authors of those studies included summary statistics (means, standard deviations, and sample sizes). An investigator can then consult power tables such as those published by Cohen (5) to estimate the necessary sample size. If there have not
been similar studies, investigators can make a more qualitative estimate of anticipated treatment effects using Cohen’s guidelines to estimate sample sizes.

As part of an a priori power analysis, investigators should consider appropriate levels for \( \alpha \) and \( \beta \). By convention, incorrectly rejecting a true null hypothesis (probability = \( \alpha \)) is considered a more serious problem than incorrectly accepting a false null hypothesis (probability = \( \beta \)). This is likely based on an assumption that if no significant treatment effect is detected, a person is unlikely to change what he or she is doing. However, when investigating the efficacy of animal versus non-animal (or dissection versus non-dissection) approaches to education, some people interpret no detectable treatment effect (accepting the null hypothesis) as evidence that one should change what is being done, and stop using animals in favor of an alternative approach (26, 38). Such a decision will most likely require the revision of course material, the purchase of new equipment or models, and retraining the faculty offering the course. To make such a large curricular change while not considering the probability of incorrectly accepting a false null hypothesis (\( \beta \)) seems unwise. It becomes very important then to make certain that the research used to make such decisions has enough statistical power (1 – \( \beta \)) to detect a treatment effect, if an effect does indeed exist. A power level of 0.8 is generally considered necessary to confidently accept the null hypothesis. It is up to the investigators to justify the appropriate significance criteria and power levels before their experiments begin, and then include this information in their reports.

Investigators can maximize statistical power by setting appropriate levels of \( \alpha \) and \( \beta \), insuring that the number of participants in the study is sufficiently large, and minimizing the amount of random error within each treatment group. Random error, or within group variance, can be minimized by using exam questions that will generate both reliable and valid test scores.
Reliability is a measure of the stability, or reproducibility, of a set of scores, and can be calculated in a variety of ways (44). If students’ performance is based more on luck, or some other parameter that varies over time, then the scores cannot be considered reliable. It is therefore important for investigators to report the reliability coefficient of every set of questions used to test a parameter.

While reliability is a measure of the stability of a group of test scores, validity is a measure of whether or not an exam actually tests what we wish to know about a group of students. A valid biology exam for an undergraduate biology class will test what we expect those students to know about biology. Validity is not a characteristic of the test itself, but like reliability, validity is a characteristic of the data generated by the exam. That is, an exam which may be highly valid when used in a college biology classroom, would not be valid in a fourth grade classroom. Validity is more difficult to quantify than reliability, and for teacher-generated tests, is often established by having a group of experienced educators review an exam designed for a specific population of students, and then decide whether or not the scores the exam will generate are likely to accurately reflect what those students know about a topic (28). If a set of exam scores lack statistical reliability or validity, then one cannot confidently accept or reject a null hypothesis based on those scores.

Ideally investigators will help educators evaluate research studies by explaining the goals for the course described in the study, publishing a copy (or at least samples) of the exam questions as supplemental material, including reliability coefficients and an explanation of how the exam was validated, and of course, reporting basic summary statistics. In addition, with the exception of Marszalek and Lockard (33), who administered both immediate and delayed post-tests, all of the reviewed studies only used immediate post-tests. It would be interesting if more
studies examine how the affective experience of participating in a dissection or animal based laboratory affects long term retention of information.

No one study will be able to answer the question of which approach, animals or models, dissection or simulation, is best in the classroom. The answer will almost certainly vary from classroom to classroom, and objective to objective. If educators are going to make informed decisions however, those decisions must be based upon well designed, statistically valid studies that include detailed descriptions of course goals, student populations, and testing instruments.
REFERENCES


29. **Leathard HL and Dewhurst DG.** Comparison of the cost effectiveness of a computer-assisted learning program with a tutored demonstration to teach intestinal motility to medical students. *ALT-J* 3: 118-125, 1995.


CHAPTER 2

Cat dissection vs. sculpting human structures in clay: an analysis of two approaches to undergraduate human anatomy laboratory education.

In Appendix
CHAPTER 3

Human clay models versus cat dissection: how the similarity between the classroom and the exam affects student performance.

INTRODUCTION

A dissection experience has been part of the study of human anatomy for centuries, and has traditionally focused on the dissection of a human cadaver (10). However, many institutions do not have the resources to support a human cadaver teaching laboratory, and rely instead on having students dissect animals, such as cats, when studying human anatomy. While there is limited research comparing invertebrate (6, 8), frog (9, 11, 17), rat (4, 16), and fetal pig (12) dissections to other alternatives in high school and undergraduate introductory biology classrooms, there is only one published study that investigated the efficacy of a cat dissection in an introductory human anatomy course at the undergraduate level (19).

In 2005, Waters et. al. (19) reported that students performing cat dissections in an introductory human anatomy laboratory earned significantly poorer test scores than students who sculpted human anatomic structures from clay. The reason for the reported performance difference was unclear, and there are a variety of possible causes. For one, it is possible that the clay sculptures are a much simpler representation of anatomy than a dissected cat specimen. For example, when using a cat specimen to study the anatomy of the muscular system, a muscle may be surrounded by connective tissue, blood vessels, nerves, or other organs, and together, these structures form a much more complex (albeit realistic) anatomic representation. In contrast, even though the clay sculpture is a much simpler representation, this simplicity may promote learning. Butcher (1) demonstrated that students studying a simple representation of cardiac blood flow earned higher test scores than students who studied a more realistic, but more complicated
representation. However, in a separate study comparing a human clay sculpting experience to a prospected cadaver lesson (see Chapter 4), there were no treatment effects on student performance as measured by a muscular system exam, suggesting that the increased complexity of a biologic specimen does not interfere with learning muscular system anatomy.

Waters et. al. (19) also noted that the students in the cat dissection treatment used a laboratory manual to guide their dissections, while the students in the human clay sculpting treatment used the same laboratory manual, but supplemented with a handout describing the order to sculpt each muscle. The supplemental handout also included specific questions asking the students to consider the action of each muscle. It is possible that prompts within the supplemental handout gave the students in the human clay sculpting treatment an advantage over their classmates in the dissection treatment when answering functional anatomy questions on the muscular system examination. To test this hypothesis, this study included two cat dissection treatment groups: one that used the same laboratory manual to guide their dissection, and a second that used the laboratory manual supplemented with a handout that listed specific muscles and included questions about the action of each muscle. If the presence of a supplemental handout has no effect on student performance, then we expect there to be no detectable differences between the cat dissection and no handout treatment versus the cat dissection plus a handout treatment.

Finally, Waters et. al. (19) discussed the possibility that the reason the students assigned to the human clay sculpting treatment performed better than their classmates in the cat dissection treatment was due to differences in how human anatomy concepts were represented in the classroom versus the exams. Human cadavers, anatomy software, anatomic models, and cat dissections are all representations of anatomy. The questions on an anatomy exam may use
anatomic representations identical to those used in the classroom, such as pinned specimens or diagrams the students have studied, or very different representations. When the anatomy representations on an exam are very different from those used in the classroom, as occurs when students study cat dissections but are asked human anatomy questions, then the students must transform and apply their knowledge of cat anatomy to a human representation. The terms near and far transfer describe the extreme ends of a spectrum measuring the amount of similarity or dissimilarity between the studied and tested representations (7). All of the higher order muscular system questions used in Waters et. al.’s 2005 study tested human anatomy concepts. Therefore, when students studied a cat dissection and were later tested on human anatomy; this may have been a case of far transfer when compared to students studying a human model (even a human clay sculpture) and then being tested on a human anatomy exam. If the human clay sculpting treatment represented a case of near transfer, then this could account for the higher test scores. To test this hypothesis, we included higher order questions presented in a human context, a cat context, and also horse and frog contexts. We expect the human anatomy questions to be a case of near transfer for students in the human clay sculpting treatment, and far transfer for the students in the two cat dissection treatments. Conversely, the cat anatomy questions are expected to be a case of near transfer for students in the two cat dissection treatments, and far transfer for the students in the human clay sculpting treatment. No treatment groups studied the anatomy of the horse muscular system, but since horses and cats are both mammalian quadrupeds, we expect the horse anatomy questions to be a case of near transfer for the two cat dissection treatments, and far transfer for the human clay sculpting treatment. If the similarity between the studied and tested representations has a significant effect on student performance, then we expect each treatment group to perform better on near transfer exam questions, and to perform worse on far
transfer exam questions. We expect no treatment effects on student performance for frog anatomy questions, since the dissimilarity between amphibians and mammals are likely to make this a case of far transfer for both treatment groups.

To test if there is a treatment effect on student attitudes, students were also surveyed about their preferences to study anatomy via dissection versus models, about the value of a dissection experience when one has to learn both the name and function of a structure, and about their general attitude toward anatomy courses. We expect there to be no differences prior to the beginning of the experiment, but student attitudes may change as they have the chance to work with different types of specimens.

METHOD

Course Description

This analysis is based on the performance of 222 students enrolled in ten laboratory sections of an introductory human anatomy course taught at The Pennsylvania State University in University Park, Pennsylvania, during the spring 2005 semester. The course has no prerequisites, and most of the students were majoring in nursing, kinesiology, or another allied health field. Almost all of the students were typical college age (18-25). Students enrolled in this course all attended a single combined lecture section which met twice per week for fifty minutes each meeting, and also attended one laboratory section, which met twice per week for 115 minutes each meeting. Individual laboratory sections enrolled between twenty and twenty-four students per section, and were taught by an undergraduate or a graduate teaching assistant. All teaching assistants were supervised by, and met weekly with, the course coordinator (the PI).
Experimental Design

The lecture portion of the course was not manipulated and remained the same for all students during the entire semester. This study used a quasi-experimental design to compare the effectiveness of three types of instructional materials used during a unit on the muscular system. The first two conditions are a replication of the clay sculpting and cat dissection experiences that were tested in Waters et. al. 2005 (19). In the third condition, the students performed a cat dissection, but also received a handout that was parallel to the handout used in the clay sculpting condition. The effects of these instructional materials were evaluated by performance on a laboratory exam. Although this exam was part of the students’ coursework, items were designed to inform the questions of this study. To avoid an instructional bias, the laboratory teaching assistants who led the laboratory sections were not shown the specific exam questions until after each instructional unit was completed.

In the periods of the semester before and after the intervention phase, all students experienced identical laboratory materials and activities. Before the experimental period, students studied the skeletal system; the nervous and urogenital system came after the experimental phase. Students’ performance on these exams was used to ensure that differences across conditions could not be attributed to any factor other than experimental differences.

Instructional Conditions

When studying the anatomy of the muscular system, laboratory sections were assigned to one of the three treatment groups. In the traditional group, students performed a cat dissection that was primarily guided by photographs and diagrams of both cat and human structures in their laboratory manual. In the supplemented dissection group, students also performed a cat dissection guided by the same laboratory manual, but were also provided a nineteen page
supplemental handout which emphasized a hands-on investigation of the structure and function of the muscles. The supplemental handout guided the order of the students’ dissection, and for each muscle, required students to first observe its size and shape, then palpitate the attachment points, and finally to pull on the muscle, and deduce its action.

In the human clay sculpting treatment, students studied the structure and function of the same muscles using the same laboratory manual, but instead of a cat dissection experience, students in this condition sculpted human muscles from clay onto an eighteen inch tall plastic human skeleton model purchased from Zahourek Systems (Loveland, CO). This condition also included a supplemental handout. The handout used in this condition was parallel to the one used by the supplemented cat dissection group, and emphasized an investigative approach to the functional anatomy of the muscular system. The handout listed the order to build the muscles onto the skeleton model, asked the students to record the attachment points, and then directed the students to deduce the muscle’s actions by attaching a piece of string to the muscle’s attachment points on a skeleton, and pull on the string to observe the movement produced.

Three laboratory sections were assigned to each of the cat dissection conditions; the clay sculpting condition included four laboratory sections. In each condition, three to four students worked together on a single representation. The students in the human clay sculpting treatment had no access to preserved cat specimens, and students in the two cat dissection treatments (no handout and handout) had no access to the human clay sculptures. All students had access to high quality muscular and skeletal models (ie. SOMSO: Coburg, Germany), anatomy atlases, and textbooks in their laboratory sections, and were encouraged by their laboratory instructors to use these as part of their studies. Laboratory instructors were assigned to treatment groups so that
there was a mix of graduate and undergraduate instructors with varying amounts of teaching experience.

**Evaluation of Student Performance**

*Control exam scores.* Student performance on three lecture exams (mostly higher order questions, fifty questions per exam) and two laboratory exams (skeletal system and nervous/urogenital systems: mostly lower order questions, 50 questions per exam) served as controls. For these units, all students had the same classroom experiences and took identical exams. The average of the three lecture exam scores and the average of the two laboratory exam scores were calculated for each treatment group.

*Experimental exam scores.* The goal of this anatomy course, like many others, is to not only teach students the names and functions of the structures covered in class, but to also equip students to use what they have learned in new settings that they may encounter in the future. With this in mind, this analysis focuses on how well students were able to answer seven different types of higher order anatomy questions (Table 1). The higher order questions were not designed to reflect specific future scenarios for any one type of student, but rather represented examples of scenarios that were not specifically covered during any of the classroom exercises, and would therefore require students to apply the information they acquired in the classroom to a new situation. The higher order questions were identical for all treatment groups. These questions required students to identify muscles presented from a novel point of view, or to analyze functional anatomy problems and then describe the action of a muscle. These questions were either text or diagram based and administered during one of the fifty minute lecture sections. All exam questions were reviewed by the course instructors.
Table 1. Subsets of higher order anatomy questions used on the muscular system laboratory exam.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Description</th>
<th>Expected transfer distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human cadaver picture questions</td>
<td>Photograph of posterior shoulder and back with arrows pointing to specific muscles</td>
<td>Cat dissection treatments: far Human clay sculpting treatment: near</td>
</tr>
<tr>
<td>Human cadaver cross-section picture questions</td>
<td>Photograph of transverse section through thigh with arrows indicating specific muscles</td>
<td>Cat dissection treatments: far Human clay sculpting treatment: near</td>
</tr>
<tr>
<td>Analysis of human limb movement questions</td>
<td>Diagram and text illustrating abduction and adduction of upper limb</td>
<td>Cat dissection treatments: far Human clay sculpting treatment: near</td>
</tr>
<tr>
<td>Cat picture questions</td>
<td>Diagram of undissected cat asking students to identify muscles with only external anatomy cues</td>
<td>Cat dissection treatments: near Human clay sculpting treatment: far</td>
</tr>
<tr>
<td>Analysis of cat limb movement questions</td>
<td>Diagram and text illustrating flexion and extension of upper limb</td>
<td>Cat dissection treatments: near Human clay sculpting treatment: far</td>
</tr>
<tr>
<td>Dissected horse diagram questions</td>
<td>Diagram showing lateral view of superficial horse muscles</td>
<td>Cat dissection treatments: near Human clay sculpting treatment: far</td>
</tr>
<tr>
<td>Dissected frog diagram questions</td>
<td>Diagram showing ventral view of superficial frog muscles</td>
<td>Cat dissection treatments: far Human clay sculpting treatment: far</td>
</tr>
</tbody>
</table>

Evaluation of Student Attitudes

Immediately before the muscular system unit, and before they knew which approach would be used in their laboratory section, students were asked two Likert scale questions (as a 1-5 rating) for each of the following categories: the usefulness of performing cat dissections versus studying anatomy models in a teaching laboratory, and more specifically, the importance of a dissection experience when studying anatomy. The students’ responses for each pair of questions were averaged. Lower numbers on the Likert scale represented a preference for dissection, and higher numbers a preference for using models. The students were also asked a pair yes/no questions about their general attitude toward anatomy courses. For these binomial questions, the value 1 was assigned to “yes” and 2 to “no”. Responses for the pair of questions were averaged.
The students were asked the same questions immediately after the muscular system unit, and at the end of the course after everyone had had a chance to work with both models and dissected specimens.

Data Analysis

Reliability. A statistical reliability coefficient ($\alpha$) was calculated for the scores from each exam, and when appropriate, from scores generated by subsets of exam questions. A reliability coefficient is an estimate of the stability, or reproducibility, of a set of scores, and can range from 0.0 to 1.0. A reliability coefficient of 0.5 or greater is generally considered adequate (15). When calculating the $\alpha$ reliability coefficient for a subset exam scores, it is likely that the calculated reliability for the subset will be lower than that of the overall exam. This is because the subset of questions has fewer questions than the overall exam. If the $\alpha$ reliability coefficient of a subset of questions is low, but the questions themselves are not problematic, then the Spearman-Brown prophecy formula can be used to estimate what the $\alpha$ reliability coefficient would have been if there had been a greater number of questions (18).

Power analysis. It is important for any experiment to have adequate statistical power to detect a treatment effect. Experiments with low statistical power are more likely to mistakenly accept a false null hypothesis (detect no treatment effect, when in fact there was a treatment effect). By convention, a statistical power of at least 0.8 is considered necessary to confidently accept the null hypothesis that the difference between treatment groups is statistically insignificant (13). The statistical power of an experiment is determined by the number of subjects, the probability of rejecting a true null hypothesis ($\alpha$), and treatment effect size. Treatment effect size can be estimated by calculating Cohen’s $d$, which estimates the observed difference between treatment groups as a proportion of the overall standard deviation.
Cohen’s d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}}

where \( \hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \)

While there are no set rules for interpreting Cohen’s d, Cohen considered a d value of 0.5 to be indicative of a medium sized treatment effect, which he described as a difference visible to someone that is familiar with the parameter (3). Very large sample sizes can make it possible to detect very small treatment effects, but such differences may not be of practical significance to a classroom instructor. For this experiment, we defined a medium treatment effect to be of practical significance. Therefore, with an accessible sample population of the 222 student enrolled in ten anatomy sections, and seventy to eighty students assigned to each of the three treatment groups, a statistical power of 0.84 could be attained by setting the probability of rejecting a true null hypothesis \( (\alpha) \) to 0.05 (5).

**Control anatomy exams.** A single factor analysis of variance was performed on the lecture and laboratory control anatomy exam scores for each of the three treatment groups to test for a significant treatment effect. If any effects were identified, pair-wise comparisons using t-tests would be conducted as follow up.

**Experimental anatomy exams.** For the scores on the muscular system exam, student performance was evaluated for each of the seven subsets of question (Table 1) by calculating the mean exam score for each of the three treatment groups, and then performing three pair-wise comparisons: 1) between the traditional cat dissection (no supplemental handout) and the human clay sculpting treatment; 2) between the supplemented cat dissection (who used supplemental handout) and the human clay sculpting treatment; and 3), between the two cat dissection treatments (without and with a supplemental handout). A t-test statistic was calculated for all
comparisons using an unbiased estimate of standard deviation. To test the null hypothesis that the difference between each pair of average scores equals zero, a P value was calculated.

*Student attitude surveys.* For the student attitude surveys, the average response for each question was calculated within each treatment group, and then the null hypothesis that the difference between the treatment means equals zero was tested as described above.

**RESULTS**

**Control Exam Scores**

To ensure that there was no bias in the assignment of students to the three treatment groups, nor to the individual sections, the averaged scores of three lecture exams (α reliability coefficients = 0.7 to 0.8) and two laboratory exams (α reliability coefficients = 0.9 each) that were not included in the experimental portion of the course were used as controls (Fig. 1). When the averaged scores of the three lecture exams for each treatment group were compared to one another using a single factor analysis of variance, no significant difference was detected among the treatment groups ($\bar{X}_{\text{cat: no HO}} = 39.1\%$, $\sigma_{\text{cat: no HO}} = 6.2\%$, $n_{\text{cat: no HO}} = 58$; $\bar{X}_{\text{cat: w/HO}} = 39.6\%$, $\sigma_{\text{cat: w/HO}} = 5.1\%$, $n_{\text{cat: w/HO}} = 61$; $\bar{X}_{\text{clay}} = 39.1\%$, $\sigma_{\text{clay}} = 4.9\%$, $n_{\text{clay}} = 76$). Similarly, when the averaged scores of the two laboratory exams for each treatment group were compared to one another using a single factor analysis of variance, no significant difference was detected among the treatment groups ($\bar{X}_{\text{cat: no HO}} = 41.8\%$, $\sigma_{\text{cat: no HO}} = 5.7\%$, $n_{\text{cat: no HO}} = 58$; $\bar{X}_{\text{cat: w/HO}} = 41.8\%$, $\sigma_{\text{cat: w/HO}} = 4.4\%$, $n_{\text{cat: w/HO}} = 61$; $\bar{X}_{\text{clay}} = 40.6\%$, $\sigma_{\text{clay}} = 7.2\%$, $n_{\text{clay}} = 76$). The students in each of the treatment groups seem to have been of similar ability.
Fig. 1. Three combined lecture exams (mostly higher order questions) and two combined laboratory exams (mostly lower order questions) taken by all students served as a control. There were no significant differences between the control exam scores (mean ± SE) of students assigned to any of the three treatment groups in the lecture ($P=0.85$), or in the laboratory ($P=0.41$).

**Experiment Exam Scores**

The muscular system exam was comprised of higher order questions consisting of novel concepts and relationships that the students had not previously encountered in the classroom. The exam was divided into seven subsets of questions (Table 1). Even though the overall $\alpha$ reliability coefficient for the entire exam was strong ($0.9$), $\alpha$ reliability coefficients were also calculated for each of the seven subsections, and with one exception, ranged from $0.6$ to $0.9$. The $\alpha$ reliability coefficient for the questions asking students to identify muscles on a horse anatomy diagram was
extremely low (0.3). Since there were a limited number of questions in each subset, the Spearman-Brown prophecy formula was used to estimate what the α reliability coefficient would have been if there were ten question in each subset. After a Spearman-Brown adjustment for ten questions per subset, the estimated reliability coefficients for all seven question subsets ranged from 0.6 to 0.9

*Human higher order anatomy questions.* Figure 2 summarizes the scores for the three types of human higher order questions. When asked to identify muscles on a photograph of a dissected human cadaver, the students in the human clay sculpting treatment earned significantly higher scores (mean = 61%, SD = 34%, n = 84) than both the students that performed cat dissections without a supplemental handout (mean = 31%, SD = 25%, n = 69) and with the supplemental handout (mean = 32%, SD = 27%, n = 69). When asked a similar set of identification questions using a picture of a transverse section through the middle of the thigh, students in the human clay sculpting treatment again earned significantly higher scores (mean = 54%, SD = 41%, n = 84) than the students who performed cat dissections without a supplemental handout (mean = 37%, SD = 36%, n = 69) or with the supplemental handout (mean = 37%, SD = 38%, n = 69). However, when students were asked to analyze functional anatomy questions about human muscle actions there were no significant differences in test scores between the cat dissection with no handout group (mean = 24%, SD = 22%, n = 69), the cat dissection with a supplemental handout group (mean = 24%, SD = 24%, n = 69), or the human clay sculpting group (mean = 28%, SD = 22%, n = 84). In two out of three measures, a human clay sculpting experience had a significantly positive effect on students’ ability to ask higher order human anatomy questions.
Fig. 2. Laboratory exam scores (mean ± SE) for three types of higher order human anatomy questions. 1) When asked to identify structures on a human cadaver diagram, students in the human clay sculpting treatment performed significantly better than those in both the cat dissection treatment without a handout ($P < 0.000001$), and with a handout ($P < 0.000001$). There was no difference between the two cat dissection treatments ($P = 0.79$). 2) When asked to identify human muscles seen in a transverse section diagram, students in the human clay sculpting treatment again performed significantly better than those in both the cat dissection treatment without a handout ($P = 0.009$), and with a handout ($P = 0.01$). There was no difference between the two cat dissection treatments ($P = 0.97$). 3) When asked to analyze a functional anatomy question on muscle action of a human, there were no significant differences between any of the three comparisons: clay sculpting to cat dissection-no handout ($P = 0.25$), clay sculpting to cat dissection w/handout ($P = 0.35$), nor between the two cat treatments ($P = 0.88$).

Cat higher order anatomy questions. Figure 3 summarizes the scores for two types of higher order cat anatomy questions. When asked to identify muscles from an external anatomy diagram of a cat, the students in the human clay sculpting treatment earned significantly poorer scores (mean = 18%, SD = 17%, $n = 84$) than the students that performed cat dissections without
a supplemental handout (mean = 52%, SD = 35%, n = 69) or with the supplemental handout (mean = 50%, SD = 36%, n = 69). When the students were asked to analyze functional anatomy questions about cat muscle actions, the students assigned to the human clay sculpting treatment again earned poorer scores (mean = 32%, SD = 28%, n = 84) than the students that performed cat dissections without a supplemental handout (mean = 44%, SD = 33%, n = 69) or with the supplemental handout (mean = 38%, SD = 32%, n = 69), but only the comparison between the human clay sculpting group and the cat dissection with no supplemental handout group was significant. Performing a cat dissection clearly helps students identify muscles on a cat diagram, and seems to have some positive effect when analyzing functional anatomy questions about cat muscle actions.
Fig. 3. Laboratory exam scores (mean ± SE) for two types of higher order cat anatomy questions. 1) When asked to identify structures on a cat external anatomy diagram, students in the human clay sculpting treatment performed significantly worse than those in both the cat dissection treatment without a handout ($P < 0.000001$), and with a handout ($P < 0.000001$). There was no difference between the two cat dissection treatments ($P = 0.68$). 2) When asked to analyze a functional anatomy question on muscle action of a cat, students in the cat dissection treatment that did not use a handout performed significantly better than their classmates assigned to the human clay sculpting treatment ($P = 0.02$), but no other significant treatment effects were detected between either the two cat dissection treatments ($P = 0.34$), or cat dissection using a handout versus the human clay sculpting treatment ($P = 0.21$).

_Horse and frog higher order anatomy questions._ In addition to the human and cat anatomy questions, students were asked to identify muscles on organisms that neither group had studied in the laboratory (Fig. 4). On questions requiring students to identify muscles on a horse cadaver diagram, the students in the human clay sculpting treatment performed significantly worse (mean = 10%, SD = 13%, n = 84) than both the students who performed cat dissections without a supplemental handout (mean = 24%, SD = 22%, n = 69) or with the supplemental
handout (mean = 25%, SD = 22%, n = 69). Once again, there was no significant difference between the two cat dissection treatments.

On questions requiring students to identify muscles on a frog cadaver diagram, there were no significant differences in test scores between the human clay sculpting group (mean = 52%, SD = 34%, n = 84), the cat dissection with no handout group (mean = 46%, SD = 33%, n = 69), or the cat dissection with a supplemental handout group (mean = 46%, SD = 32%, n = 69). Performing a cat dissection helped students answer questions on horse anatomy, but all three treatment groups performed similarly when attempting to answer questions on frog anatomy.

![Fig. 4. Laboratory exam scores (mean ± SE) for two types of higher order non-human/non-cat anatomy questions.](image)

1) When asked to identify structures on a horse cadaver diagram, students in the human clay sculpting treatment performed significantly worse than those in both the cat dissection treatment without a handout ($P = 0.000003$), and with a handout ($P = 0.000001$). There was no difference between the two cat dissection treatments ($P = 0.92$). 3) When asked to identify structures on a frog cadaver diagram, there were no significant differences between any of the three comparisons: clay sculpting to cat dissection-no handout ($P = 0.30$), clay sculpting to cat dissection w/handout ($P = 0.23$), nor between the two cat treatments ($P = 0.89$).
Student Attitudes

Immediately prior to the muscular system unit, immediately after the muscular system unit, and then at the conclusion of the course, students were asked three types of questions relating to their anatomy experiences: to indicate their general preference for studying anatomy by performing a cat dissection versus using anatomic models in the classroom (Table 2), to rate the importance of a cat dissection experience versus anatomy models when the course goal is to learn both the names and functions of anatomic structures (Table 3), and finally, to share their general attitude toward anatomy as a field of study (Table 4). When asked their opinions before the experimental portion of the course, there were no differences detected between any of the treatment groups to any of the questions, indicating that the assignment to each of the treatment groups was unbiased with regard to student attitude. By the mid-point of the course, when asked about their preferences for performing dissections versus using models (Table 2), and the value of dissection versus studying models when learning anatomic names and functions (Table 3), both groups performing dissections preferred the dissection approach significantly more than the group building human clay sculptures. This response pattern continued even to the conclusion of the course, after all students had a chance to work with both models and dissected cats. There were no differences detected in student attitude between the two cat treatments for any question. When students were asked about their general attitude toward anatomy (Table 4), there were no treatment effects detected. Students show clear preferences toward the specific approach used in their anatomy class when asked to rate one approach versus the other, but a cat dissection versus a human clay sculpting experience does not seem to affect how they feel about anatomy as a field of study.
Table 2. Attitude toward using human cadavers vs. anatomy models in human anatomy teaching laboratories

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Cat dissection no handout</th>
<th>Cat dissection with handout</th>
<th>Human clay sculpting</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>2.94 ± 0.14 (62)</td>
<td>2.93 ± 0.12 (59)</td>
<td>2.92 ± 0.10 (74)</td>
<td>$P_{1,3}=0.92$, $P_{2,3}=0.93$, $P_{1,2}=0.99$</td>
</tr>
<tr>
<td>Mid-Experiment</td>
<td>2.78 ± 0.16 (47)</td>
<td>2.51 ± 0.15 (51)</td>
<td>3.22 ± 0.13 (58)</td>
<td>$P_{1,3}=0.04$, $P_{2,3}=0.0006$, $P_{1,2}=0.23$</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>2.38 ± 0.13 (51)</td>
<td>2.53 ± 0.16 (49)</td>
<td>3.14 ± 0.13 (69)</td>
<td>$P_{1,3}=0.0001$, $P_{2,3}=0.003$, $P_{1,2}=0.48$</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1-5 where 1 = mostly cadaver dissection/cadaver dissections are most important, and 5 = mostly anatomy models/anatomy models are most important. Number in parentheses is number of students. *Scores of two Likert scale questions were averaged: “If you were in charge of this course and could use cadaver dissection and/or anatomy models to teach your students, what would you use to teach your students anatomy?” and “As an anatomy student, is it more important to study anatomy with a cadaver dissection, or is using anatomy models more important?” † P value for difference between treatment groups within each survey. See METHODS section for calculations. Number in parentheses is number of students.

Table 3. Attitude toward the importance of dissection in human anatomy teaching laboratories when learning the functions of a structure

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Cat dissection no handout</th>
<th>Cat dissection with handout</th>
<th>Human clay sculpting</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>3.15 ± 0.14 (61)</td>
<td>3.25 ± 0.11 (59)</td>
<td>3.20 ± 0.11 (74)</td>
<td>$P_{1,3}=0.75$, $P_{2,3}=0.74$, $P_{1,2}=0.55$</td>
</tr>
<tr>
<td>Mid-Experiment</td>
<td>2.65 ± 0.15 (47)</td>
<td>2.78 ± 0.15 (51)</td>
<td>3.32 ± 0.13 (58)</td>
<td>$P_{1,3}=0.001$, $P_{2,3}=0.009$, $P_{1,2}=0.53$</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>2.70 ± 0.14 (51)</td>
<td>2.71 ± 0.15 (49)</td>
<td>3.18 ± 0.13 (69)</td>
<td>$P_{1,3}=0.02$, $P_{2,3}=0.03$, $P_{1,2}=0.93$</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1-5 where 1 = cadaver dissections are best/most helpful, and 5 = anatomy models are best/most helpful. Number in parentheses is number of students. *Scores of two Likert scale questions were averaged: “In lab, we spend a lot of time learning the names of structures. What is the best way to learn both the names and how structures work, studying a cadaver dissection or studying anatomy models?” and “In your opinion, what is the most helpful way to learn both the name and the function of a structure?”. † P value for difference between treatment groups within each survey. See METHODS section for calculations.
Table 4. Attitude toward human anatomy courses

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Cat dissection no handout1</th>
<th>Cat dissection with handout2</th>
<th>Human clay sculpting3</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>0.92 ± 0.03 (62)</td>
<td>0.86 ± 0.03 (59)</td>
<td>0.89 ± 0.03 (75)</td>
<td>( P_{1,3}=0.51, P_{2,3}=0.37, \ P_{1,2}=0.17 )</td>
</tr>
<tr>
<td>Mid-Experiment</td>
<td>0.79 ± 0.05 (47)</td>
<td>0.92 ± 0.05 (51)</td>
<td>0.86 ± 0.03 (59)</td>
<td>( P_{1,3}=0.21, P_{2,3}=0.32, \ P_{1,2}=0.06 )</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>0.90 ± 0.04 (51)</td>
<td>0.90 ± 0.04 (49)</td>
<td>0.90 ± 0.03 (69)</td>
<td>( P_{1,3}=0.94, P_{2,3}=0.99, \ P_{1,2}=0.94 )</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores where 1 = yes, 2 = no. Number in parentheses is number of students. *Scores of two binomial questions were averaged: “If there was time in your schedule, would you consider taking another anatomy course after this one?” and “Do you enjoy studying anatomy?”. † P value for differences between treatment groups within each survey. See METHODS section for calculations.

DISCUSSION

The control exam scores indicate that there was no detectable bias in the assignment of the ten different sections to the three treatment groups. However, significant treatment effects were detected when students were asked to identify cat or human muscles from pictures. Students assigned to the human clay sculpting treatment earned significantly higher scores than their classmates on higher order questions requiring them to identify human muscles from a picture, and earned significantly lower scores than their classmates on questions requiring them to identify cat muscles from a picture. This suggests that when students are required to identify a muscle in a novel context, there is not a consistent advantage or disadvantage to a cat dissection versus a human clay sculpting experience, but rather, whichever approach minimizes the transfer distance between the laboratory specimen and the specific type of exam question seems to work best. We expected to see a similar treatment effect on the functional anatomy questions requiring students to analyze the action of a muscle, but the results were not as clear.

When students were presented with a novel functional anatomy question, and then asked to identify a muscle’s action, there was a statistically non-significant trend favoring the students
assigned to the human clay sculpting treatment when the questions were presented in a human anatomy context, and a similar trend favoring the students assigned to the cat dissection treatments when the questions were presented in a cat anatomy context. This trend is consistent with our hypothesis that student performance will improve when the studied representation is similar to the form of the tested representation. Only one comparison however detected a significant treatment effect: when answering a cat muscle action question, the cat dissection treatment with no supplemental handout performed better than the human clay sculpting treatment. More data is necessary to make any conclusive statement about the effect of a cat dissection versus human clay sculpting experience on students’ ability to answer muscle action questions.

When students were asked comparative anatomy questions requiring them to identify muscles on a horse diagram, there was a significant treatment effect in favor of both cat dissection treatments compared to the human clay sculpting treatment. The horse muscle diagram presented muscles from a lateral view, and given that the cats and horses are both mammalian quadrupeds, the shape and organization of the horse muscles were likely more similar to cat specimens than to human specimens. We expect that the cat-horse representations where more similar than human-horse, and result in the observed increase in exam scores.

No differences in exam scores between any of the treatment groups were detected when students were asked to identify muscles on a frog diagram. For students enrolled in an introductory anatomy course, amphibian anatomy may be perceived as so different from that of both cat and human, that the transfer distances for each treatment group were similar. If transfer distances were similar, then we expect no detectable treatment effects.
There was also no treatment effect detected when a traditional cat dissection experience built around a laboratory manual was compared to a cat dissection using the same manual, but supplemented with a handout emphasizing functional anatomy. This demonstrates that the student performance differences between a traditional cat dissection experience and a human clay sculpting experience reported by Waters et al. (19) are not attributable to the human clay sculpting treatment using a supplemental handout versus the cat dissection having no handout.

When students were asked about their general preferences regarding studying from models versus performing cat dissections (Table 2) and for their opinion about the value of performing dissections when they need to learn both the name and function of a structure (Table 3), there were no significant treatment effects prior to the experimental portion of the course, demonstrating again that there was no bias in the assignment of the individual sections to the three treatment groups. However, when students were asked the same questions immediately after completing the muscular system unit, and then a third time at the end of the semester (after everyone had a chance to work with both models and dissected cats), the students demonstrated a significant preference toward the type of specimen used in their treatment group. This result is consistent with most other studies that have asked similar questions (2, 14, 17, 19), and suggests that student attitude toward dissection versus models alone should not determine the specific approach an instructor chooses for the classroom, since on average, students will likely come to value whichever approach they experience.

When asked about their general attitude toward anatomy courses (Table 4), there were no detectable treatment effects at any point during the semester. This result is consistent with that reported by Waters et al (19) when a human clay sculpting lesson was compared to a cat dissection experience, but differs from a separate comparison between a human clay sculpting
experience and a prospected human cadaver lesson (see Chapter 4). In the latter experiment, students that studied a human cadaver in class were significantly more positive about anatomy courses than their classmates who sculpted human structures from clay. The affective experience of working with a human cadaver may be distinct from working with models or even dissecting other organisms. Additional research is required to demonstrate whether or not such an experience may improve student learning.

CONCLUSION

This study cannot be interpreted to suggest that either cat dissection or human clay sculpting are universally equivalent approaches, nor that one is always superior to the other. If one wishes to teach human anatomy, then a specimen that looks like a human, even a human anatomy clay sculpture is preferable to a non-human specimen. However, if one wishes to teach the anatomy of a cat or any other organism (such as in a comparative vertebrate anatomy course), then dissecting an animal specimen is likely the more effective approach.
REFERENCES

CHAPTER 4


INTRODUCTION

Human cadaver dissection has been part of anatomy since at least 300 BCE (9), and anatomic models have been part of human anatomy laboratories since the fifteenth century (11); but as teaching tools, are anatomic models as effective as cadaver dissection? Some argue that a cadaver experience should forever be part of a human anatomy laboratory (10, 14, 16), while others suggest that for at least some types of learning, an anatomic model may be an effective alternative to either dissection or vivisection. There are only a few studies that attempted to compare student performance in classrooms using dissection or vivisection to classrooms using anatomic models. In an introductory undergraduate anatomy course, Downie and Meadows (4) compared a rat dissection lesson to lessons using models and found no significant differences in student performance as measured by scores on a written exam, however a lack of summary statistics and the students choosing their treatment group makes interpretation of this study problematic. Three other studies compared the efficacy of an animal specimen to a model when learning surgical skills.

Greenfield et. al. (7) reported no significant treatment effects when comparing the surgical skills of veterinary students that practiced on an anesthetized dog versus surgical models, however this is likely due to the study having extremely low statistical power (see Chapter 1). Griffon et. al. (8) compared surgical skills of veterinary students that practiced ovariohysterectomies on either dog cadavers or reproductive system models. All students were evaluated during surgeries performed on anesthetized dogs. The students who practiced on
models were ranked more proficient than their classmates in the dog cadaver treatment, possibly because the models were dynamic, incorporating model blood vessels that pulsed with red fluid, and thus provided a more realistic surgical experience than a dog cadaver. In a similar study measuring basic surgical skills of veterinary students, Olsen et. al. (15) compared students that practiced surgery using a dynamic blood vessel model to students that practiced using an anesthetized dog. Again, both treatment groups were evaluated using anesthetized dogs. In this case, both treatment groups had similarly realistic experiences, and in eleven of thirteen measures of surgical skill, both groups performed equally well. Student performance improves when the material studied in the classroom and the material presented on an exam is represented in similar ways. The similarity or dissimilarity between representations in the classroom and representations on a test are described by the terms near transfer and far transfer (6). The surgical models used by Griffon et. al. (8) and Olsen et. al. (15) likely represented a live animal surgery realistically enough to be a case of near transfer.

Near transfer versus far transfer may also have been a factor in a study comparing cat dissection to sculpting human structures from clay in an introductory undergraduate human anatomy course. In a 2005 study by Waters et. al. (19), undergraduate students assigned to a human clay sculpting treatment were more successful than their classmates who performed a cat dissection when answering both higher and lower order human anatomy questions. The anatomy representations between the classroom and the test were therefore more similar (near transfer) for the students in the human clay sculpting treatment, and less similar (far transfer) for the students in the cat dissection treatment. Alternatively, these results could also be explained by the human clay sculpture being a simpler representation of muscular system anatomy.
Butcher (1) demonstrated that learning is improved when students study a simple diagram of blood flow through the heart as opposed to studying a more realistic, but more complicated diagram. If students in the cat dissection treatment described by Waters et. al. (19) were required to interpret a more complicated representation of muscular system anatomy, with blood vessels, nerves, other organs and additional muscles all visible, then the relative simplicity of the human clay sculptures could also account for the improved exam scores. The dissection-clay model muscle anatomy comparison was therefore confounded with the cat representation being a case far transfer and more complicated, versus the human representation being a case of near transfer and less complicated.

This experiment was designed to address the confound in the Waters et. al. (19) study by comparing the learning outcome of students studying a prosected human cadaver to those sculpting human structures from clay. In both treatments, the students studied a human representation of anatomy and were then tested on human anatomy, so the representational transformations, or transfer distance, required of both groups were the same between the studied and tested representations. If the group differences reported by Waters et. al. (19) were due to the human clay sculpting treatment representing anatomy in a simpler and easier to learn context, then the students in the human clay sculpting condition of this study should obtain higher scores on the posttest than the student in the prosected human cadaver experience. On the other hand, if students who sculpted human clay structures in the study by Waters et. al. (19) obtained their superior posttest scores because of the similarity of the studied and tested representations (near transfer), the students in this study should perform equally well regardless of the study condition. Finally, if a human cadaver lesson is an inherently superior learning experience compared to
sculpting anatomic structures from clay, then the students in the prosected human cadaver
treatment should earn higher posttest scores than the students in the human clay sculpting group.

METHOD

Course Description

This analysis is based on the performance of 108 students enrolled in ten laboratory
sections of an introductory human anatomy and physiology course taught at Mohawk Valley
Community College in Utica, New York, during the fall 2005 semester. The only course pre-
requisite was high school chemistry, and most of the students were working toward an RN
degree in nursing (80%) or other allied health fields. The class was composed of approximately
30% international students and 70% U.S.-born students. Half of the students were typical college
age (18-25), and the other half were returning adult students (over 25). Students enrolled in this
course all attended a lecture section which met for 150 total minutes each week, plus one
laboratory section, which met once per week for three hours. Individual laboratory sections were
taught by one of three experienced full-time faculty members, and enrolled between ten and
sixteen students per section.

Experimental Design

During the first ten weeks of the course, all students studied the same lessons covering
the human integumentary system, digestive system, skeletal system, and histology. During the
eleventh and twelfth week, students studied the anatomy of the muscular system. When teaching
the anatomy of the muscular system, instructors used one of two different approaches in their
laboratory sections. In the control treatment (five laboratory sections), the anatomy lesson was a
guided hands-on investigation of the structure and function of twenty-one trunk, arm, and neck
muscles using a prosected human cadaver (three to four students per cadaver). Students were
required to not only observe the size and shape of each muscle, but to pull on the muscles, noting the attachment points, and deduce each muscle’s action. High quality muscular and skeletal models, anatomy atlases, posters, and textbooks were also available for the students to study. In the experimental treatment (five laboratory sections), students studied the structure and function of the same muscles, but used a modified lesson where they sculpted the same twenty-one muscles from clay onto an eighteen inch tall plastic human skeleton model (three to four students per model) purchased from Zahourek Systems (Loveland, CO). The students in the human clay sculpting treatment had no access to the human cadavers, but were provided with the same supplemental materials as the students in the control treatment. The students in the prosected human cadaver group had no access to the human clay sculptures. Each of the three instructors taught a combination of control and experimental sections during the two week muscle unit. The PI was present in all sections to ensure the course material was presented in a consistent manner.

The objectives for each lesson included not only learning the names, insertions, and actions of the twenty-seven trunk, neck, and arm muscles, but to also develop investigative skills to allow the students to characterize muscles not covered in this course. For instance, by studying the joints crossed by a skeletal muscle, the shape of the skeletal muscle, where that muscle attaches to bones, and the direction that the skeletal muscle fibers run, students should be able to describe the action of any muscle, even if they do not know its name. This functional approach requires students to do more than simply memorize a table of facts, and should prepare them to understand novel situations in future academic and clinical settings. The course objectives and lessons were reviewed and evaluated by four anatomy and physiology educators, each with between twelve and twenty-five years experience.
Evaluation of Student Performance

Prior to the muscular system unit, a thirty-two item exam covering the material from the digestive system, comprised of both lower and higher order questions, was used to measure if there were any differences in ability level of students enrolled in the ten different laboratory sections. At the conclusion of the muscular system unit, student performance was evaluated using a forty-two question exam composed of three different types of lower order questions, which emphasized information covered in the laboratory that could be memorized, as well as higher order questions that required the students to apply their knowledge to analyze a novel functional or clinical anatomy problem. In effort to avoid instructional bias toward the exam items, the test was not written until after the students completed the muscle unit. The exam for the muscular system was composed of the four general types of questions described below.

1) Lower order questions which were identical for both treatment groups and required students to make an identification (a muscle’s name or attachment - 4 questions) or describe a muscle’s action (6 questions). These questions were either text/diagram based or pinned on one of the high quality plastic models (ie. SOMSO: Coburg, Germany) used in the classroom.

2) Higher order questions which were presented in a context that was not covered by any of the anatomy exercises, and were also identical for both treatment groups. These questions required students to identify structures presented from a new point of view (such as a transverse limb section - 4 questions), or analyze functional anatomy problems to describe the action of a muscle (5 questions). Like the lower order questions described above, these questions were either text/diagram based or were a muscle pinned on one of the high quality models used in the classroom.
3) Lower order “sister questions” that tested the same muscle for each treatment group, where the prosected human cadaver group answered a question pinned on a human cadaver, and the human clay sculpture group answered the same question pinned on a human-clay sculpture. These questions required students to identify a muscle’s name (8 questions) or action (11 questions).

4) Lower order “cross-over questions” that also tested the same muscle for each group, but the prosected human cadaver group answered a question pinned on a human-clay sculpture, while the human clay sculpture group answered the same question pinned on a human cadaver. These questions also required students identify a muscle’s name (2 questions) or action (2 questions).

Evaluation of Student Attitudes

Immediately before the muscular system unit, and before they knew which approach would be used in their laboratory section, students answered general questions regarding their opinions on the usefulness of using human cadavers versus anatomy models in a teaching laboratory, and specifically the importance of a dissection experience when studying anatomy. There were two questions in each of the two categories. The questions were framed as five point Likert scales, where lower values indicated a preference for cadavers and dissection and higher numbers indicate a preference for using models in the laboratory. Participants also answered a pair of yes/no questions about their general attitude toward anatomy courses. These questions were scored as “yes”=1 and “no”=2. The students answered the same questions immediately after the muscular system unit.
Data Analysis

Reliability. Reliability is a measure of the stability, or reproducibility, of a set of test scores. Reliability coefficients can range from 0.0 to 1.0, with higher coefficients the most desirable. In this experiment, α reliability coefficients were calculated for both the control (digestive system) exam scores and experimental (muscular system) exam scores, as well as the scores of the four subsets of questions described above. Because the reliability coefficient may decrease as the number of items on a given scale decreases, reliability estimates for subscales may be lower than the reliability estimate for the overall test. If the scores from a small subset of questions have a low reliability coefficient due to the limited number of questions asked, but the questions themselves were otherwise effective, one can use the Spearman-Brown prophesy formula to estimate what the reliability coefficient would have been if there had been a greater number of questions (18). It should be noted that if the test questions were ineffective to begin with, thus producing unstable scores, the Spearman-Brown prophesy formula will probably not yield an acceptable reliability coefficient.

Power analysis. Since accepting the null hypothesis is potentially as important as rejecting it, we attempted to attain a statistical power of at least 0.8, which conventionally, is the desired minimum to confidently accept the null hypothesis (13). Statistical power is affected by sample size, the probability of rejecting a true null hypothesis (α), and treatment effect size. Effect size is the difference between the mean values of the treatments (ie. control minus experimental) as a proportion of standard error (standard deviation of the overall mean), and is often reported as Cohen’s d.
Cohen’s d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}

where \( \hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma^2_1 + \sigma^2_2}{n_1 + n_2}} \)

Effect sizes indicate if the differences between experimental conditions are meaningful. Cohen suggested that a d value of approximately 0.5 is indicative of a medium effect size and would be plainly visible to someone familiar with the parameter (3). While it is possible that there may be small treatment effects in this experiment, given the investment of time and money an institution must make to offer an anatomy laboratory course, we defined a medium effect size or larger to be of practical significance.

For this experiment, our accessible sample population was the 108 students enrolled in ten anatomy sections. With approximately 54 students in each of the two treatment groups, a statistical power of 0.8 could be attained by adjusting the probability of rejecting a true null hypothesis (\( \alpha \)) to 0.1 (5). While higher than the conventional value of \( \alpha = 0.05 \), we decided that the resulting increase in power, and the concurrent decreased probability of incorrectly accepting a false null hypothesis (\( \beta \)), made this a reasonable trade off.

Anatomy exams. Student performance was evaluated for each type of question described earlier by calculating the mean exam scores, and performing a pair-wise comparison between the prospected cadaver and human clay sculpting treatments. A t-test statistic was calculated for all comparisons using an unbiased estimate of standard deviation. To test the null hypothesis that the difference between each pair of mean scores equals zero, a p value was calculated.

Student attitude surveys. For the student attitude surveys, the average response for each question was calculated within each treatment group, and then the null hypothesis that the difference between the treatment means equals zero was tested as described above.
RESULTS

Control Exam Scores

We compared scores on the digestive system laboratory exam to ensure that there was no bias in the assignment of students to the two treatment groups. On the digestive system laboratory exam, there were no significant differences between the scores of students assigned to the prosected human cadaver group (mean = 54%, SD = 18%, n = 55) and the scores of students assigned to human-clay sculpting group (mean = 53%, SD = 15%, n = 51). When the digestive system laboratory exam scores of the three individual laboratory instructors’ sections involved in this experiment were compared to one another using a single factor analysis of variance, no significant difference was detected among the instructors’ sections (mean$_1$ = 56%, SD$_1$ = 17%, n$_1$ = 41; mean$_2$ = 51%, SD$_2$ = 16%, n$_2$ = 32; mean$_3$ = 52%, SD$_3$ = 17%, n$_3$ = 33). These analyses demonstrate that the students in each of the sections and treatment groups were of similar ability.
There were no significant differences between the control exam scores (mean ± SE) of students assigned to the two treatment groups ($P = 0.72$) or among the sections taught by the three instructors involved in the experiment ($P = 0.41$).

**Experiment Exam Scores**

Four types of questions were used on the muscular system laboratory exam: lower-order identical questions, lower-order sister questions, lower-order cross-over questions, and higher-order identical questions. Within the four types of questions, students were required to answer identification questions (name a structure) and more functional questions (identify a muscle action). Lower order questions were classified as those concepts and anatomic relationships that the students had an opportunity to study during the laboratory session. Higher order questions were classified as novel concepts and relationships that the students had not previously encountered in the classroom. Even though the overall $\alpha$ reliability coefficient for the entire exam was strong (0.8), actual $\alpha$ reliability coefficients were also calculated for each of the
subsections described below, and with one exception, ranged from 0.4 to 0.5 (Table 1). Since there were a limited number of questions in each subset, the Spearman-Brown prophecy formula was used to estimate what the α reliability coefficient would have been if there were ten question in each subset. With one exception, Spearman-Brown estimated reliability coefficients ranged from 0.5 to 0.6. The α reliability coefficient for the lower order cross-over questions asking students to identify a structure was extremely low.

Table 1. Actual α reliability coefficients ($\rho_{xx}$) based on the number of questions in each subset, and estimated Spearman-Brown α reliability coefficients ($\rho_{yy}$) if there were 10 questions in each subset.

<table>
<thead>
<tr>
<th>Question Subset</th>
<th>Question focus</th>
<th>Actual test length (number of questions)</th>
<th>Actual $\alpha$ reliability coefficient ($\rho_{xx}$)*</th>
<th>Factor by which actual test length must be changed to equal 10 questions*</th>
<th>Spearman-Brown estimated α reliability coefficient ($\rho_{yy}$) * if there were 10 questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower order identical questions</td>
<td>identification action</td>
<td>4</td>
<td>0.4</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0.4</td>
<td>1.7</td>
<td>0.5</td>
</tr>
<tr>
<td>Lower order sister questions</td>
<td>identification action</td>
<td>8</td>
<td>0.5</td>
<td>1.3</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>0.5</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>Lower order cross-over questions</td>
<td>identification action</td>
<td>2</td>
<td>0.05</td>
<td>5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.5</td>
<td>5</td>
<td>0.8</td>
</tr>
<tr>
<td>Higher order questions</td>
<td>identification action</td>
<td>4</td>
<td>0.4</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.4</td>
<td>2.0</td>
<td>0.6</td>
</tr>
</tbody>
</table>

*Spearman-Brown estimated α reliability coefficients ($\rho_{yy}$) were calculated using the equation:

$$\rho_{yy} = k(\rho_{xx})/1+(k-1)(\rho_{xx})$$

where $\rho_{xx}$ = the actual reliability, † and $k$ = the factor by which test length has changed.

Lower-order identical questions. For identical lower order questions (Fig. 2) where students had to identify the name of a muscle pinned on a high quality plastic model, there were no significant differences between the human cadaver group (mean = 63%, SD = 25%, n = 56) and the human-clay sculpting group (mean = 58%, SD = 25%, n = 52). When students were asked to identify the action of the pinned muscle, there was also no significant difference between the human cadaver (mean = 40%, SD = 23%, n = 56) and human-clay sculpting (mean =
39%, SD = 23%, n = 52) treatments. When students were only required to recognize and identify the name or action of a muscle they had studied in class, we detected no measurable differences between the prosected human cadaver and human-clay sculpting treatments.

**Fig. 2.** Laboratory exam scores (mean ± SE) for two types of lower order questions. Students in both treatments were required to identify the names and actions of the same muscles pinned on identical high quality plastic models. There were no significant treatment effects on identical muscle identification questions ($P = 0.33$) or on identical muscle action questions ($P = 0.86$).

*Lower-order, non-identical questions.* The muscular system laboratory exam also included two types of related, but non-identical lower order questions. In the first type, “sister questions” were used that tested the same muscle for each treatment group, where students in the prosected human cadaver group answered a question pinned on a cadaver, and students in the human-clay sculpting group answered the same question pinned on a human-clay muscle sculpture. In the second type of non-identical lower-order question, “cross-over questions” were
used to again test the same muscle for each treatment group, but in this case, students in the prosected human cadaver group answered a question pinned on a human-clay muscle sculpture, while students in the human-clay sculpting group answered the same question pinned on a cadaver.

Figure 3 summarizes the results of the lower-order sister questions. When students were asked sister questions requiring them to name a muscle, there were no significant differences between the human cadaver group (mean = 54%, SD = 21%, n = 56) and the human-clay sculpting group (mean = 53%, SD = 25%, n = 52). However, when students were asked sister questions requiring them to identify the action of a muscle, the students assigned to the human-clay sculpting treatment performed significantly better (mean = 52%, SD = 18%, n = 52) than their classmates in the prosected human cadaver treatment (mean = 44%, SD = 20%, n = 56). A clay sculpting experience helped students to answer lower-order muscle action questions.
Fig. 3. Laboratory exam scores (mean ± SE) for two types of low order sister questions. For these sister questions, students assigned to the human cadaver group identified a muscle name or muscle action covered in the classroom exercise that was pinned on a human cadaver, and students assigned to the human clay sculpture group identified the same muscle or action pinned on a human clay sculpture. There were no significant treatment effects on students’ ability to identify the names of muscles on these sister questions ($P = 0.92$), however students in the human clay sculpting treatment were able to identify significantly more muscle actions pinned as sister questions than students from the human cadaver group ($P = 0.04$).

When students were asked lower-order cross-over questions (Fig. 4) that focused on identifying the name of a muscle, no significant differences were detected between the human cadaver group (mean = 49%, SD = 38%, n = 56) and the human-clay sculpting group (mean = 46%, SD = 32%, n = 52), however, the statistical reliability of these questions was extremely low (Table 1), so no conclusion can be drawn based on this data. When asked to identify the action of
a muscle on a lower-order cross-over question, there were no differences in the performance of students in the human cadaver group (mean = 38\%, SD = 27\%, n = 56) and those in the human-clay sculpting group (mean = 38\%, SD = 30\%, n = 52). Answering lower-order muscle action questions in a format other than the one emphasized by the assigned treatment does not seem to favor one group of students over the other.

![Experiment Exam Scores](image)

**Fig. 4.** Laboratory exam scores (mean ± SE) for two types of cross-over low order questions. For these cross-over questions, students assigned to the human cadaver group identified a muscle or action covered in the classroom exercise that was pinned on a *human clay sculpture*, while students assigned to the human clay sculpture group identified the same muscle or action pinned on a *human cadaver*. For the cross-over questions, there were no significant treatment effects on students’ ability to identify the names of muscles (*P* = 0.67), or muscle actions (*P* = 0.87).
**Higher-order identical questions.** Students in both conditions were also asked identical higher-order muscle anatomy questions (Fig. 5). These questions asked students to name a characteristic or specific anatomic structure that was presented in a context they had never previously encountered in the classroom. Students in the prossected human cadaver group (mean = 58%, SD = 28%, n = 56) out performed their peers in the human-clay sculpting group (mean = 51%, SD = 24%, n = 52) but the difference was not statistically significant. When students were asked to deduce the action of muscles in a novel situation, there were also no performance differences between the students in the prossected human cadaver group (mean = 35%, SD = 22%, n = 56) compared to the human-clay sculpting group (mean = 33%, SD = 23%, n = 52). We were unable to detect any significant performance differences on identical higher order questions.
Fig. 5. Laboratory exam scores (mean ± SE) for two types of higher order questions. Students were asked to identify novel anatomic features on models or drawings (origins and insertions of muscles not studied in class, or the names of known muscles presented from a previously unseen point of view, such as a cross section through a limb), and also asked to deduce the actions of muscles they had never encountered in class. Students in the prosected human cadaver treatment were able to identify more novel structures than students in the human clay sculpting treatment, however the difference was not statistically significant ($P = 0.20$). There were also no significant treatment effects on students’ ability to deduce the actions of novel muscles ($P = 0.60$).

**Student Attitudes**

Immediately prior to and after the experimental portion of the course, students were surveyed and asked three types of questions relating to their anatomy experiences: to indicate their general preference for studying from a prosected human cadaver versus an anatomic model in the anatomy classroom (Table 2), to rate the effectiveness of human cadavers versus anatomy models when the course goal is to learn both the names and functions of anatomic structures
(Table 3), and finally, to share their general attitude toward anatomy as a field of study (Table 4). In all three cases, there were no differences detected between the prosected human cadaver group and the human-clay sculpting group on any of the pre-experiment survey questions, indicating that the assignment to each of the two treatment groups was unbiased with regard to student attitude. However, at the conclusion of the experiment, students recorded significantly different opinions about their preferences for studying with cadavers versus models (Table 2) and the value of cadavers versus models when learning anatomic names and functions (Table 3), with each group of students clearly preferring the approach used in their assigned treatment.

Interestingly, there was also a difference on the post-experiment survey in students’ attitudes toward anatomy as a field of study (Table 4), with the students assigned to the prosected human cadaver treatment recording a small, but significantly more positive attitude than their peers assigned to the human-clay sculpting treatment.

Table 2. Attitude toward using human cadavers vs. anatomy models in human anatomy teaching laboratories

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Prosected human cadaver group</th>
<th>Human-clay sculpting group</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>2.56 ± 0.14 (58)</td>
<td>2.29 ± 0.13 (58)</td>
<td>0.16</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>2.52 ± 0.14 (56)</td>
<td>3.34 ± 0.15 (56)</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1-5 where 1 = mostly cadaver dissection/cadaver dissections are most important, and 5 = mostly anatomy models/anatomy models are most important. Number in parentheses is number of students. *Scores of two Likert scale questions were averaged: “If you were in charge of this course and could use cadaver dissection and/or anatomy models to teach your students, what would you use to teach your students anatomy?” and “As an anatomy student, is it more important to study anatomy with a cadaver dissection, or is using anatomy models more important?” . † P value for difference between treatment groups within each survey. See METHODS section for calculations. Number in parentheses is number of students.
Table 3. Attitude toward the importance of dissection in human anatomy teaching laboratories when learning the functions of a structure

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Prosected human cadaver group</th>
<th>Human-clay sculpting group</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>2.73 ± 0.15 (58)</td>
<td>2.72 ± 0.12 (56)</td>
<td>0.96</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>2.75 ± 0.15 (58)</td>
<td>3.66 ± 0.15 (55)</td>
<td>0.00003</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1-5 where 1 = cadaver dissections are best/most helpful, and 5 = anatomy models are best/most helpful. Number in parentheses is number of students. *Scores of two Likert scale questions were averaged: “In lab, we spend a lot of time learning the names of structures. What is the best way to learn both the names and how structures work, studying a cadaver dissection or studying anatomy models?” and “In your opinion, what is the most helpful way to learn both the name and the function of a structure?”. † P value for difference between treatment groups within each survey. See METHODS section for calculations.

Table 4. Attitude toward human anatomy courses

<table>
<thead>
<tr>
<th>Survey*</th>
<th>Prosected human cadaver group</th>
<th>Human-clay sculpting group</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Experiment</td>
<td>1.16 ± 0.05 (58)</td>
<td>1.21 ± 0.05 (56)</td>
<td>0.57</td>
</tr>
<tr>
<td>Post-Experiment</td>
<td>1.10 ± 0.04 (57)</td>
<td>1.25 ± 0.06 (54)</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores where 1 = yes, 2 = no. Number in parentheses is number of students. *Scores of two binomial questions were averaged: “If there was time in your schedule, would you consider taking another anatomy course after this one?” and “Do you enjoy studying anatomy?”. † P value for difference between treatment groups within each survey. See METHODS section for calculations.

DISCUSSION

In this study, we sought to determine if a human clay sculpting experience in an undergraduate introductory anatomy teaching laboratory results in similar student performance as a prosected cadaver experience. With the exception of non-identical lower order sister questions that asked students to identify a muscle action, we detected no statistically significant treatment effects when comparing the performance of anatomy laboratory students who studied skeletal muscles using a prosected cadaver versus those that sculpted the same muscles from clay onto a plastic skeleton. These results are consistent with our hypothesis that when comparing a prosected cadaver learning experience to a human clay sculpting approach to anatomy education,
the most important factor may be transfer distance: the amount of similarity between the learning
and testing representations. In this study, both treatment groups studied material that looked like
human beings, and were then tested on human anatomy. Since both treatments were a case of
near-transfer, we expected there to be little or no difference in student performance on the course
examination. As noted, the one exception was student performance on non-identical lower order
sister questions that tested students’ recognition of a muscle’s action.

On the muscle action sister questions, students in the human clay sculpting treatment out
performed their classmates studying prosected human cadavers. This may seem counterintuitive,
given that the limbs of the human clay sculptures are rigid and cannot be moved. One possible
explanation of this difference was how the students could interact with their specimens during
the laboratory period.

As described in the Methods section, all students had access to high quality muscle and
skeletal models. All students also used a similar laboratory exercise that encouraged them to
interact with the anatomy specimens as much as possible. Because the human skeleton models in
the laboratory were articulated, making it very easy to move the limbs, the laboratory exercise
directed students to attach each end of a string to a human skeleton at the points a skeletal muscle
originates and inserts. By pulling on the string, students could easily observe the muscle’s action
as the skeleton moved. The joints of the skeleton models could usually be articulated more easily
than even the human cadavers, and may have been a superior learning experience when studying
the action of muscles. Differences between the human cadavers and the plastic models used for
sculpting however, may have influenced students’ decisions to engage the skeletal models.
Students studying human cadavers could pull on muscles without tearing them, manipulate
limbs, and observe at least limited flexion, extension, abduction, adduction, etc. Because students
working with human cadavers were able to carry out these manipulations on the cadaver itself, these students may not have perceived the need to work with additional models. The plastic models used for sculpting muscles, by contrast, were rigid and students were forced to use the alternative models to study movement. If students in the human clay sculpting treatment were more likely to spend time on such an exercise because their human clay sculptures were immovable, then they may have learned at least the actions of the twenty one muscles covered in class better than the students in the cadaver treatment. Future research in this area should record the time students spend using each classroom resource, in order to determine the contribution of each to student performance.

In this study, student attitudes regarding human cadavers versus models, the importance of dissected specimens to learning, and their opinion of anatomy courses in general were also recorded. Not surprisingly, there were no statistically significant differences between the prosected cadaver and human clay sculpting groups prior to the experiment, thus providing additional evidence that assignment to the two treatment groups was unbiased. However, after completing the two week muscle laboratory exercises, when asked about their general preference for cadavers versus models (Table 2), and about the value of cadavers versus models to learning the names and functions of structures (Table 3), students showed a significant preference in favor of the approach used in their classroom. This is consistent with other studies that surveyed student attitudes about an animal or dissection based approach to education versus a non-animal or non-dissection based approach (2, 15, 17, 19). It is interesting that in this study, students’ laboratory experience also affected their attitude toward human anatomy courses in general (Table 4). Students in the prosected human cadaver laboratories were significantly more positive about taking an anatomy course than the students in the human clay sculpting treatment. While
the performance of the two groups in the current class is nearly identical, there may be other
treatment effects that do not become apparent until sometime in the future. It would be
interesting to perform a longitudinal study to see if there are treatment effects on the amount of
information retained, or students’ attitudes toward the laboratory, over time. In 1999, Marszalek
and Lockard (12) demonstrated that long term (three month) retention of knowledge and
comprehension of frog internal anatomy was improved for junior high school students that
performed a dissection versus studying anatomy with computer programs. It is possible that the
real benefit of a dissection or cadaver experience is not the short term immediate posttest score,
but perhaps a long term effect on retention or other affective measures such as interest,
motivation, or self efficacy.

CONCLUSION

In 2005, Waters et. al. (19) reported that anatomy students who sculpted human structures
from clay learned more about human anatomy than did students who experienced a cat dissection
laboratory exercise. In that study however, the dissection-no dissection comparison was
confounded by the specific forms of the anatomy representation studied. Whereas the human
clay sculpting exercise was well matched to the human focused questions used in testing,
students who studied a cat specimen had to additional task of transforming that anatomic
representation to match the tested human representation. The purpose of this study was to
address that confound by comparing groups of students who had a human cadaver prosection
experience to a group of students who did not have such an experience. Differences in the
representational form were controlled by comparing students who sculpted human structures
from clay to those who studied human prosected cadavers.
The results of this study suggest that the treatment differences reported by Waters et. al. (19) were caused by the similarity of the studied and tested representations. Students who had a human cadaver laboratory experience scored as well on the knowledge posttest as did students who sculpted human structures from clay. The only exception to this was that students in the human clay sculpting treatment were more successful at answering one type of human muscle action question. Overall, we detected no clear differences in student performance between the prosected cadaver and human clay sculpting treatments.

When asked for the opinions about their laboratory experience, we found that students tend to favor whatever approach they have been exposed to, at least as measured by surveys administered immediately after the lesson. What this study, and most others, do not address, is the long term effect of the classroom lessons. Students in the prosected human cadaver treatment were significantly more positive about studying anatomy than their classmates in the human clay sculpting treatment. If one teaching approach engages or challenges students more than the other, will there be long term differences in student achievement, attitudes, or motivation? Longitudinal studies with large groups of students are necessary to address the longer term effects of different approaches to laboratory education. It is then that we can start to make informed decisions regarding not only cadaver use, but the role of animal vivisection, educational technology, and models as educational tools.
REFERENCES


APPENDIX

Cat dissection vs. sculpting human structures in clay: an analysis of two approaches to undergraduate human anatomy laboratory education.
Cat dissection vs. sculpting human structures in clay: an analysis of two approaches to undergraduate human anatomy laboratory education

John R. Waters,1 Peggy Van Meter,2 William Perrotti,3 Salvatore Drogo,3 and Richard J. Cyr1

1Departments of Biology and 2Educational Psychology, The Pennsylvania State University, University Park, Pennsylvania; and 3Life Science Department, Mohawk Valley Community College, Utica, New York

Received 9 August 2004; accepted in final form 2 December 2004

Waters, John R., Peggy Van Meter, William Perrotti, Salvatore Drogo, and Richard J. Cyr. Cat dissection vs. sculpting human structures in clay: an analysis of two approaches to undergraduate human anatomy laboratory education. Adv Physiol Educ 29: 27–34, 2005; doi:10.1152/advan.00033.2004.—Many human anatomy courses are taught using cat dissection. Alternatives are available, but information regarding learning outcomes is incomplete. In 2003, ~120 undergraduates enrolled in a human anatomy course were assigned to one of two treatment groups. In the control group, students performed cat dissections (emphasizing isolation and identification) of the muscular, digestive, and cardiovascular systems. In the experimental treatment group, students built clay sculptures of each human body system. Student learning was evaluated by using both low- and high-difficulty questions. On pre- and postexperiment control exams, there were no significant differences in student performance. On exams after a cat dissection vs. a human-clay sculpting experience, the students in the human-clay sculpting treatment group scored significantly higher than their classmates in the cat dissection group on both the low- and high-difficulty questions. Student attitudes toward dissection and taking future human anatomy courses were also measured. There were no differences in student attitudes at the beginning of the experiment; afterward, students exposed to a cat dissection experience viewed dissection more favorably than students in the human-clay sculpting treatment group. There were no treatment effects on student willingness to take future human anatomy courses. The experimental design makes it difficult to conclude precisely why students assigned to the human-clay sculpting experience performed better on exams, but as each method was performed in this particular human anatomy course, our data indicate that human-clay sculpting may be a viable alternative to cat dissection in an anatomy course in which the students focus on human anatomy.

Animal; alternative; performance; attitude; Maniken

Human anatomy laboratory instruction can employ many different types of specimens, models, software programs, and web sites to help students learn the material. However, it is difficult to know how these materials affect learning outcomes. This study compares the effect of two laboratory education approaches, a cat dissection vs. a human-clay sculpting experience, on the performance and attitudes of undergraduate students enrolled in a human anatomy course. This study examines the use of sculpting human anatomic structures from clay as an alternative to cat dissection, because like dissection, sculpting is an approach that engages the students in exploration of anatomical structures and relationships. Moreover, sculpting and cat dissection require a three-dimensional approach to human anatomy (as opposed to a 2-dimensional picture or computer simulation), and some variation in the size, shape, and position of structures will be apparent to the students when they compare their sculpted specimens.

A dissection or vivisection experience is a common part of teaching laboratories. Students may participate in dissections or manipulation of human cadavers, preserved cats, or other specimens. Human anatomy instructors often try to evaluate alternative approaches to laboratory education by comparing them to some sort of dissection/vivisection experience. Student performance in laboratories based on stereoscopic slide presentations (11), videodisks (1, 5), high-quality plastic models (2), and computer-based instruction (3, 7, 10) have all been compared with animal- or cadaver-based laboratory experiences with varying results. No differences in student performance were detected in some of the studies comparing nonanimal alternatives to dissection/vivisection experiences (1, 2, 5, 11). Hughes (3) found mixed results when comparing a nonanimal alternative to traditional dissection, with students in the dissection lab performing better on some, but not all aspects of the course. Predavec (10) on the other hand, found that computer instruction was more effective than traditional dissection. Conversely, Matthews (7) found a dissection experience to be more effective than a computer-based experience. Although the differences in these studies may well be attributable to the differences in interventions tested, it is also possible that the differences are due to important methodological variations.

Although there are many methodological differences across studies, the most critical variation concerns the measures used to assess outcomes. Differences in outcome assessments may contribute to different conclusions across studies. In science learning for example, Mayer and Sims (8) found that students in an experimental condition did not score higher on a posttest assessing recognition than control participants scored. These experimental students, however, did score higher than control participants on a higher-order test assessing the application of knowledge. The consistency of these findings across studies has led Kintsch (4) to argue that investigators need to be aware that different types of posttests will likely capture different types of knowledge.

In this study, the performance of students enrolled in a large undergraduate human anatomy course (~120 students), assigned to one of two treatment groups (cat dissection vs. human-clay sculpting) was assessed by using both lower- and higher-order examination questions. In designing the exam questions, consideration was given to the objectives of the human anatomy laboratory as taught at The Pennsylvania State University. In this introductory course, the goal was to provide students with a solid foundation in the principles of human anatomy that would enable them to apply those principles in...
Teaching in the Laboratory

Whatever situations come next in their careers. It is difficult to predict exactly what those future situations will be; the students in this course came from a variety of majors and are on career paths that will require varying degrees of proficiency in human anatomy. Therefore, the goal of the higher-order exam questions was to create new situations for the students to apply their understanding of human anatomy.

Using this range of exam questions, we tested the null hypothesis that there is no significant difference in exam performance between students assigned to a cat dissection laboratory and students assigned to a laboratory in which they sculpt human anatomic structures from clay. We also tested the null hypothesis that the different laboratory experiences would have no effect on student attitude toward the use and value of preserved animals in an educational setting nor on their willingness to take a subsequent human anatomy course. Our results indicate a significant difference between the two treatments with the human-clay sculpting group outperforming the cat dissection group on both lower- and higher-order questions. Importantly, neither treatment adversely affected student attitude. We discuss these findings in terms of how the human-clay sculpting experience might facilitate a deeper understanding of human structure/function relationships and promote a facile transfer of knowledge to human, rather than cat, anatomical problem sets.

METHODS

Course Description

In the fall of 2003, ~120 undergraduate students enrolled in an introductory human anatomy course (Biology 129) offered at The Pennsylvania State University’s main campus. This course has no prerequisites. Students enrolled in Biology 129 that semester consisted of 77% health/science majors (mostly kinesiology, nursing, and other allied health fields) and 23% nonhealth/nonscience majors. Laboratory sections were offered in pairs, with each laboratory room identically equipped. The labs were staffed by a mix of graduate and undergraduate teaching assistants supervised by a laboratory coordinator (J. R. Waters). Individual laboratory sections within each treatment group were taught by a mix of new and returning teaching assistants. Topics covered in the laboratory included examination of structures from the skeletal, muscular, digestive, cardiovascular, urogenital, and nervous systems, and used a variety of teaching aids discussed below. All of the anatomy students attended two 50-min lectures together each week. Laboratory sections met twice weekly for 115 min each meeting.

Experimental Design

The anatomy students were assigned to one of two treatment groups, based on which of the two laboratory rooms their section was assigned. Six of the seven laboratory sections were paired by treatment, with students in one room acting as the control group and students in the other room acting as the treatment group (the 7th section was another control group). The students in the control group studied the anatomy of the muscular system, the digestive system, and the cardiovascular system, primarily by dissecting a preserved cat specimen. Their cat dissection experience was supplemented with high-quality plastic human models (Fig. 1) showing corresponding human structures, a laboratory manual, and an anatomy textbook. The students were instructed by their laboratory teaching assistants to compare the structures isolated during the cat dissection to the human models. The treatment group studied the same topics, but instead of performing a cat dissection, they built clay sculptures of human anatomical structures. The students in this human-clay sculpting treatment group had access to the same supplements as the cat dissection control group and were also instructed to use the supplements as they sculpted anatomical structures. Four laboratory sections were assigned to the cat dissection control group, and three laboratory sections were assigned to the human-clay sculpting treatment group (16–20 students per section). Students were not told of their section assignment until the experimental portion of the course began. The laboratory portion of the course was divided into four units, with each unit culminating in a laboratory exam (Table 1).

The first and fourth units of the course served as pre- and postexperiment controls. For the first unit of the laboratory, all students studied the same human skeleton specimens and took the same laboratory exam. During the fourth unit of the course, all students studied the same preserved sheep brain specimens and exposed the same urogenital structures during a cat dissection, and again took a common laboratory exam. Cat dissections vs. human-clay sculpting experiences were compared during the two middle units of the laboratory. It was only during the second and third units of the laboratory course that the students assigned to the two groups had different laboratory experiences.

In the experimental portion of the course, students assigned to the cat dissection control group dissected large (18- to 20-in. long) cat specimens (2–3 students/cat) purchased from Fisher Educational Materials Division. Students assigned to the human-clay sculpting treatment group built the muscular system onto plastic human skeleton Maniken figures purchased from Zahourek Systems (Figs. 2 and 3), or when studying the digestive and circulatory systems, built clay sculptures of human structures into 9- by 13-in. aluminum baking trays (Fig. 4) using clay sculpting tools provided in the laboratory (Fig. 5). Students doing human-clay sculpting also worked in groups of 2–3 students per sculpture (Fig. 6).

The Pennsylvania State University, like many other institutions, lacks the ability to offer human cadaver dissection as a component of a large-enrollment undergraduate introductory human anatomy course. In such instances, when human cadaver dissection is impractical, it is common practice to use preserved cats as alternative instructional material. It is therefore important to note that this study was not designed to test the teaching efficacy of dissection vs. nondissection per se; rather it was to ask whether a clay-sculpting experience using human anatomical relationship might lead to different learning outcomes than a cat dissection experience when the purpose of the course was to learn human rather than cat anatomy, and students were subsequently exposed to human anatomical problem
sets. We recognize an inherent bias in this study regarding the issue of knowledge transfer between different species (i.e., students learning on human-clay sculptures need only consider human anatomy, whereas students learning on dissected cats need to consider both cat and human anatomical relationships). Nonetheless, because of the widespread use of cat dissections as an instructional method in many human anatomy courses, we felt the experimental design reflected common institutional practice and warranted investigation in a manner consistent with how many human anatomy courses are currently taught.

**Evaluation of Student Performance**

Student performance was evaluated by using four laboratory exams. The first and fourth laboratory exams served as pre- and postexperiment controls, and each consisted of 50 laboratory practical questions (mostly knowledge and comprehension) covering the names of skeletal system structures (exam 1) and the names of nervous and urogenital system structures (exam 4). For exams 1 and 4, students in both treatment groups took identical laboratory exams.

The second and third laboratory exams were used to evaluate the experimental portion of the course. Each exam consisted of two parts: an evening practical exam, similar to the pre- and postexperiment control practical exams; and a short answer exam (taken earlier the same day) consisting of higher-order questions. The evening practical exams included questions that were identical for each treatment group and questions that were similar, but not identical for each treatment group.

Table 1. **Experimental design**

<table>
<thead>
<tr>
<th>Instructional Units</th>
<th>Laboratory Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (4 sections)</td>
</tr>
<tr>
<td>1*</td>
<td>Skeleton</td>
</tr>
<tr>
<td>2†</td>
<td>Cat dissection</td>
</tr>
<tr>
<td>3‡</td>
<td>Cat dissection</td>
</tr>
<tr>
<td>4§</td>
<td>Sheep brain and cat dissection</td>
</tr>
</tbody>
</table>

*Preexperiment control (human skeletal system), †experimental portion (muscular system), ‡experimental portion (digestive and cardiovascular systems), §postexperiment control (nervous and urogenital systems).

During the second evening practical exam on the muscular system, all students were told that the exam would include eight muscles pinned on the high-quality plastic human models (Fig. 1). Thirty other exam questions were similar, but nonidentical for each treatment group. Students in the cat dissection control group were asked to identify anatomical structures pinned on dissected cat specimens. Students in the human-clay sculpting treatment group answered questions on the same anatomical structures, but pinned on clay sculptures. The third evening practical exam was structured in the same manner, but consisted of 39 related cat dissection/human-clay sculpture questions.

The higher-order questions of each short-answer exam (taken earlier the day of the evening practical exam) presented identical questions to students from each treatment group (15 questions on the 2nd exam, 26 questions on the 3rd exam). These questions required analysis or evaluation of novel situations not covered in the lecture or laboratory. By presenting students with questions they never had a chance to practice, we strove to measure whether there was a treatment effect on the students’ ability to draw on their foundation of anatomical knowledge when analyzing a new situation. After the students leave this course, it is our desire that they will be able to apply what they have learned to a variety of situations that are
dissimilar to the teaching laboratory environment. The higher-order questions were not meant to reproduce any “real world” situations, but instead, simply present the students with a novel problem. Examples of the higher-order questions included identifying muscles and organs in cross-sectional views, identifying injured muscles based on a patient reporting painful limb movements, and identifying the location of organs based on human surface anatomy (for a complete list of questions, see supplemental material at http://advan.physiology.org/cgi/content/full/00033.2004/DC1). The situations and examples covered by these questions were not discussed with the students in the lecture or laboratory portions of the course. The higher-order questions tested the student’s ability to transfer and use information learned in the laboratory to a new context.

Evaluation of Student Attitudes

Immediately before and immediately after the experimental portion of the laboratory, students were asked Lichert scale questions (as a 1–5 rating) for their opinions on the importance of using preserved animals in a teaching laboratory, and the value of a preserved animal dissection experience when learning the name and the function of a structure. They were also asked whether they would consider taking a dissection experience when learning the name and the function of a structure? The mean response for each question was calculated within each treatment group, and then the null hypothesis that the difference between the treatment means equals zero was tested as described above.

RESULTS

Control Exam Scores

To ensure a nonbiased distribution of students within this study, we examined student performance on preexperiment and postexperiment exam scores (Fig. 7). Before the experiment, there were no significant differences between the laboratory exam scores of the students assigned to the cat dissection control group (mean = 76%, SD = 15.9%, n = 76) and the human-clay sculpting treatment group (mean = 80%, SD = 19%, n = 60). Similarly, the postexperiment exam scores of the cat dissection group (mean = 70%, SD = 20.3%, n = 76) were not significantly different from those of the human-clay sculpting treatment group (mean = 68%, SD = 19.2%, n = 59). We also examined student performance on four lecture exam questions, in which the students were asked to apply concepts discussed in lecture or to analyze a new situation. There were no significant differences (P = 0.25) in the ability of the students assigned to the cat dissection control group

Table 2. Attitude toward using preserved animals in human anatomy teaching laboratories

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Start of Experiment</th>
<th>End of Experiment</th>
<th>P Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat dissection</td>
<td>2.5±0.02 (67)</td>
<td>2.0±0.02 (64)</td>
<td>0.0030</td>
</tr>
<tr>
<td>Human-clay sculpting</td>
<td>2.6±0.02 (54)</td>
<td>2.8±0.02 (47)</td>
<td>0.1800</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1–5 where 1 = very important/a lot and 5 = not very important/not very much. Number in parentheses is the number of students. Scores of two Lichert scale questions were averaged: “As an anatomy student, how do you feel about the use of preserved animals in teaching laboratories?” and “If you were in charge of this course how much would you use preserved animals in your teaching?”. †P value for difference between beginning and end of experiment values within each treatment group. See METHODS for calculations.

Table 3. Attitude toward the importance of dissection when learning functions of a structure

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Start of Experiment</th>
<th>End of Experiment</th>
<th>P Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat dissection</td>
<td>2.5±0.02 (67)</td>
<td>2.2±0.02 (64)</td>
<td>0.0720</td>
</tr>
<tr>
<td>Human-clay sculpting</td>
<td>2.5±0.02 (54)</td>
<td>3.3±0.02 (47)</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores on a scale of 1–5 where 1 = very important/very helpful and 5 = not very important/not very helpful. Number in parentheses is number of students. Scores of two Lichert scale questions were averaged: “In lab, we spend a lot of time learning the names of structures. How important is a cat dissection for learning how structures work?” and “In your opinion, how much does dissecting an animal in lab help if you need to learn both the name and the function of a structure?” The mean response for these two similar questions (both address student attitude) is shown. †P value for difference between beginning and end of experiment values within each treatment group. See METHODS for calculations.
Students in the cat dissection control group (pinned on dissection vs. cats (P = 0.0000001) and when answering similar questions pinned on human-clay sculptures vs. cats (P = 0.012). On the gastrointestinal/cardiovascular system (GI & CV sys) exam, the scores of students in the human-clay sculpting treatment group were again significantly higher than those of students in the cat dissection control group when answering similar questions pinned on human-clay sculptures vs. dissected cats (P < 0.0000001).

Table 4. Attitude toward taking a future human anatomy course

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Start of Experiment</th>
<th>End of Experiment</th>
<th>P Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat dissection</td>
<td>1.3 ±0.01 (66)</td>
<td>1.2 ±0.01 (62)</td>
<td>0.2709</td>
</tr>
<tr>
<td>Human-clay sculpting</td>
<td>1.1 ±0.01 (50)</td>
<td>1.1 ±0.01 (45)</td>
<td>0.4641</td>
</tr>
</tbody>
</table>

Results are means ± SE of scores of 1 = yes, 2 = no. Binomial question: “If there was time in your schedule, would you consider taking another anatomy course after Biology 129?” †P value for difference between beginning and end of experiment values within each treatment group. See METHODS for calculations.

(mean = 76%, SD = 22%, n = 76) vs. the human-clay sculpting treatment group (mean = 79%, SD = 20%, n = 59) to answer these types of questions. These results indicate that students in all sections demonstrated similar ability in the course.

Experimental Exam Scores

Anatomy students traditionally rely on a good deal of memorization to learn the names and functions of anatomical structures. Most instructors hope that some higher-order learning occurs in the laboratory as well. In this study, both lower-order and higher-order exam questions were used to evaluate student performance.

Lower-order questions. For lower-order questions (knowledge and recall) given during the experimental portion of the laboratory course, there were significant treatment effects (Fig. 8). On questions asking the identity of specific muscles using a plastic human model (identical models were used in each treatment), students in the human-clay sculpting treatment group scored significantly higher (mean = 83.9%, SD = 19.9%, n = 62) than students in the cat dissection control group (mean = 56.3%, SD = 30%, n = 75). When students in the human-clay sculpting treatment group were asked to identify the same structures (pinned on human-clay sculptures) as students in the cat dissection control group (pinned on dissected cats), the students in the human-clay sculpting treatment group again scored significantly higher (mean = 81.1%, SD = 16%, n = 62) than the students in the cat dissection control group (mean = 74.2%, SD = 19.3%, n = 75). On the following laboratory exam using the same question format, students in the human-clay sculpting treatment group again earned significantly higher exam scores (mean = 87.2%, SD = 10.6%, n = 56) than the students in the cat dissection control group (mean = 69.1%, SD = 14.4%, n = 76). These data suggest that under these conditions, a human-clay sculpting laboratory experience is more effective than a cat dissection experience when students are required to recall the name or function of anatomical structures.

Higher-order questions. For higher-order questions (primarily analysis), there were also significant treatment effects during the experimental portion of the course (Fig. 9). For the second laboratory exam, students in the human-clay sculpting treatment group earned significantly higher exam scores (mean = 56.6%; SD = 30%; n = 64) than students in the cat dissection control group (mean = 20.1%; SD = 37.4%; n = 80). These questions required the students to analyze novel situations that had not been presented in the lecture or laboratory. The exam scores of the human-clay sculpting treatment group were also significantly higher (mean = 58.2%; SD = 21.6%; n = 62) than those of the cat dissection control group (mean = 49.0%; SD = 22.9%; n = 80) on the third laboratory exam. In this course, a human-clay sculpting experience seems more beneficial than a cat dissection experience when students were asked to apply their knowledge of human anatomy to a new situation.

Student Attitudes

Immediately before and immediately after the experimental portion of the course, students were asked their opinions about
the use of preserved animals in a teaching laboratory and about their feelings toward taking other human anatomy courses.

Before a cat dissection or human-clay sculpting experience, students assigned to each treatment had similar feelings about the importance of using preserved animals in a human anatomy teaching laboratory. After their respective experiences, students in the cat dissection treatment group considered animal use significantly more important than they indicated earlier, whereas students in the human-clay sculpting experience group did not significantly change their opinion (Table 2).

When asked more specifically for their opinion about the importance of a dissection experience when learning the name and function of structures, students again had similar attitudes before their sculpting/cat dissection experiences (Table 3). After these experiences, the students in the cat dissection group viewed dissection as a more valuable learning tool than they did at the beginning of the study ($P = 0.072$), whereas students in the human-clay sculpting experience group valued dissection even less ($P = 0.0002$).

The data suggest that at the beginning of the experiment, the students participating in this study showed no strong preference regarding the use of animals in a teaching laboratory; however, students participating in a cat dissection came to view dissection as more valuable, and students that sculpted human structures from clay viewed dissection as less valuable.

When asked whether they would consider taking a human anatomy course sometime in the future, students in both treatments indicated, on average, positive responses with no significant differences between treatments or within treatments from the beginning to the end of the experiment (Table 4). Neither laboratory experience had a significant effect on students’ willingness to take a subsequent human anatomy course.

**DISCUSSION**

**Student Performance**

Compared to students in the cat dissection control group, students in the human-clay sculpting treatment group were more successful at identifying anatomical structures and transferring what they had learned about human anatomy to new situations involving human anatomy problem sets. Furthermore, whereas students in the human-clay sculpting treatment group valued a cat dissection experience less than their classmates in the control group by the end of the experiment, there were no significant differences at any point in student attitudes toward taking future human anatomy courses. The results of this experiment suggest that a human-clay sculpting experience may be a more effective laboratory teaching method than a cat dissection experience as offered in The Pennsylvania State University’s undergraduate human anatomy laboratory course.

The performance differences observed between the control and sculpting treatment groups cannot be explained by a biased distribution of student aptitude between the two groups. Students in both groups performed equally well on laboratory exams before and after the experimental portion of the course and on both “dry exams” (the skeletal system) and “wet exams” (the nervous and urogenital systems). Students in both groups also performed equally well on a mix of application and analysis level questions asked during the lecture exams. Moreover, student performance differences between the groups cannot be explained by differences in student attitudes toward dissection. Before a cat dissection or human-clay sculpting experience, there were no differences among the students enrolled in this course in their attitudes about the importance or value of dissection or in their willingness to take future human anatomy courses. The treatment differences presented here appear to be the result of using two different approaches to human anatomy laboratory education. Possible explanations for these results include transfer of learning and how actively the students were engaged in the laboratory activities.

Transfer of learning occurs when students learn information or concepts in one context, such as the laboratory, and then transfer what they have learned to a different situation, such as an exam (9). The amount of similarity (also called distance) between the two contexts is referred to as near and far transfer. Near transfer occurs when the learning and testing situations are very similar, and far transfer occurs when the two contexts are less similar. Reasons that students may fail to transfer learning from one context to another include not recognizing the similarities between the two situations, or if the students do recognize the similarities, they are unable to successfully apply what they learned in the new context (6). The students in the human-clay sculpting treatment group may have found it easier to transfer what they learned in the laboratory context to that of the exam questions pinned on the human models, because the two contexts are so similar (both centered on the human anatomy). Students in the cat dissection group were required to transfer their learning between less similar contexts (cat anatomy vs. human anatomy) and had a more difficult time answering those questions.

Transfer of learning may also explain the performance differences between students in the cat dissection control group and the human-clay sculpting treatment group when answering the higher-order questions. Because the higher-order questions were novel situations to all of the students, they represented far transfer for both groups, but the distance of transfer was likely greatest for the students in the cat dissection control group, because these questions focused on human (as opposed to cat) anatomy. It would be informative to repeat this experiment.
with additional exam questions that focus on aspects of cat anatomy, and then measure whether or not transfer of learning is greater for the students in the cat dissection control group than the human-clay sculpting treatment group. A similar experiment testing distance of transfer could be conducted comparing the performance of students in a cadaver-dissection control group to that of students in a human-clay sculpting treatment group. If the differences in student performance observed here are due to transfer of learning, then one might expect a greater rate of transfer as the laboratory and testing contexts become more similar for the dissection group.

A human Maniken clay-sculpting experience was more effective than a cat dissection when the assessments asked content and higher-order questions about human anatomy, possibly because the human-clay sculpting experience engaged the students more actively. In the human anatomy teaching laboratories at The Pennsylvania State University, a cat dissection experience generally emphasizes the isolation and identification of anatomical structures. Students in the cat dissection control group may have focused on isolating structures only partially, identifying them, and then moving on to the next structure. This type of “hit and run” approach contrasts with that required of the students in the human-clay sculpting treatment group, who had to study the general shape and position of a structure in a text or a model, mold the structure of clay, and then place the structure in the correct position relative to the surrounding structures. The students that built clay sculptures of human anatomic structures may have been more actively engaged when studying anatomical relationships.

A human-clay sculpting experience may also present fewer distracters to human anatomy students. When studying structures in the laboratory, students in the human-clay sculpting treatment group started with nothing more than a human skeleton figure or empty tray as a foundation and then added only the anatomical structures they needed to learn. By contrast, students participating in a cat dissection must isolate the structures of interest from all of the surrounding structures. For example, a muscle may be hidden beneath other muscles, covered by connective tissue, and connected to blood vessels and nerves. Some, but not all of these structures may have been discussed at the time the student attempts the cat dissection, and the extra structures may distract the student from the one structure that needs to be identified. The problem of distracters may have also contributed to treatment differences during the laboratory examinations.

When the students were asked to simply identify structures that had been studied in the anatomy laboratories (lower-order questions) on human-clay sculptures vs. dissected cat specimens, it is possible that the students in the human-clay sculpting treatment group were at an advantage compared with their classmates in the cat dissection control group. In an effort to maximize statistical power, only a single approach (cat dissection or human-clay sculpting) was tested within each treatment group, so the students in the human-clay sculpting treatment group identified structures pinned on clay sculptures of human anatomy, and the students in the cat dissection control group identified the same structures pinned on dissected cat specimens. Students in the human-clay sculpting treatment may have had an easier time identifying pinned structures simply because the clay sculptures of human anatomic structures only displayed structures the students had studied for that unit. The students in the cat dissection control group were also required to identify pinned structures they had studied, but these structures were often surrounded by additional anatomic structures that had not been discussed. These additional structures may have acted as distracters when the students in the cat dissection control group were trying to answer questions pinned on cat specimens. However, the presence of distracters does not explain the differences in student performance on the low-order questions pinned on identical human muscular system models. For those questions, all students had studied the same human models (Fig. 1) before the exam, and the exam questions were pinned identically for both groups. The superior performance of the students in the human-clay sculpting experience group on the human model questions was more likely due to transfer of learning.

### Student Attitudes

At the beginning of the experimental portion of the laboratory, students placed in the two treatment groups did not differ significantly in their opinions regarding the importance of using preserved animals to study human anatomy. After their respective laboratory experiences, the students’ attitudes toward the importance of using preserved animals to study human anatomy changed little for students that had been sculpting human structures in clay, but were significantly more positive for students that participated in a cat dissection (Table 2). When asked a more specific question about the value of cat dissection when one has to learn both the name and the function of a structure, student opinions again did not differ at the beginning of the experiment. However, after the experiment, the students in the cat dissection group saw the dissection experience as more valuable \((P = 0.072)\), whereas the students in the human-clay modeling treatment group saw a dissection experience as less important \((P = 0.0002)\) for learning both the names and functions of anatomical structures (Table 3). With no dissection experience, students seem indifferent, or to prefer nondissection approaches to learning human anatomy. After exposure to a cat dissection experience, students view dissection more positively. Similar results regarding dissection experiences were recently presented at the 2004 Annual Meeting of the Human Anatomy and Physiology Society Poster Session (K. Carlyle and G. Kawchuck: “A Comparative Study of Three Different Learning Mediums in the Anatomy Lab: Student Preferences”). When students are exposed to dissection, they come away with the opinion that it is an important part of the laboratory experience. It would have been informative to survey the students in the human-clay sculpting treatment group after they had been exposed to a cat dissection during the last unit of the laboratory course to see whether their attitudes changed as well.

No matter which approach was used in the laboratory, sculpting human structures from clay or a cat dissection, students’ attitudes toward taking a future human anatomy course did not change significantly. Their attitudes toward human anatomy as a field of study are probably based on the sum of their experiences, which includes their level of success mastering the material, their interactions with instructors and other students, and their general feelings toward the material covered in laboratory.
In conclusion, surveys on student attitudes toward cat dissection and human anatomy laboratories provide interesting feedback to the instructor, but they do not address the central question regarding undergraduate human anatomy laboratories: What are the most effective teaching methods? This experiment does not offer a definitive answer, and as with any study of this type, one must be cautious about extrapolating the results beyond the sample population, but the data do suggest that clay sculpting is an instructional approach worth considering in a teaching laboratory emphasizing the identification of human anatomical structures. Any teaching materials that engage students in activities in which they must interact with the material are worth investigating. Whether or not this level of engagement is possible to attain with computer simulations or other popular alternatives will remain unknown until more research is conducted by using a variety of teaching methods, across different student populations, which examine student learning on multiple levels.

REFERENCES


Curriculum Vitae

John R. Waters
208 Mueller Laboratory
University Park, PA 16802
(jrw8@psu.edu, 863-1154)

Education
-Master of Science  Biology, January 1992
  Illinois State University, Normal, IL
-Bachelor of Science  Biology, December 1986
  University of Illinois, Urbana-Champaign, IL

Teaching Experience
-Lecturer and Laboratory Coordinator: Penn State, University Park  PA.  1993 to present
-Lecturer: Lincoln Community College, Normal  IL  1992-1993

Publications


Presentations
Waters, J.R. and W. Perrotti. 2007. Animals, Models, and the Teaching Laboratory. How does science education research inform us when we consider alternatives to dissection or other traditional approaches to A&P teaching? Annual Meeting of the Human Anatomy and Physiology Society, (J. Waters-presenter) San Diego, California.

