EFFECT OF ADMINISTERING STUDENT EVALUATIONS OF TEACHING AFTER FINAL EXAMS IN A VETERINARY MEDICAL SCHOOL

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Misty R. Bailey
May 2021

ACKNOWLEDGEMENTS

The author would like to thank the members of the doctoral committee, Dr. Patrick Biddix (chair), Dr. Jimmy Cheek, Dr. Katie High, and Dr. India Lane for their valuable contributions to improve the study and reporting of the results; Sandra Harbison for multiple brainstorm sessions; Cary Springer and Dr. Jennifer Morrow for statistical consulting; Dr. Dianne Mawby for getting the dissertation project started; and a multitude of professional colleagues for their encouragement and support during the PhD journey.

ABSTRACT

In higher education, instructors and administrators use student evaluations of teaching (SETs) as formative and summative assessments of instruction; thus, they need adequate response rates. Veterinary students are often requested to complete over 30 SETs each semester, and response rate is shown to decline as the number of SETs increases. Nonresponse threatens the validity and reliability of results. Allowing students to complete SETs after final exams was postulated to help increase response; however, students' knowledge of their final course grade has been previously shown to negatively influence SET scores. A possible rationale for this influence is attributional bias, in which people attribute success to themselves and failure to someone else.

This mixed-methods case study explored how administering SETs to veterinary students after final exams (versus before) affected numerical scores on SET items, completion rate, and volume and substance of comments. Participants (n = 262) were randomly assigned to before or after finals groups, and 171 students completed 2,926 SETs and 1,149 open-ended comments. Attributional theory guided the analyses.

Students were more likely to complete evaluations before finals (versus after), and first-year students completed more SETs than third-year students. Compared to the prior year, in which SETs were administered before finals, students in the study year completed 31% fewer SETs. Timing of SET delivery did not significantly affect the number of students providing comments or comment length. Timing also did not affect SET item scores, but third-year students rated instructors higher than first-year students

on five of 10 items. Students' expected grade was positively correlated with all 10 SET items, suggesting attributional bias might be occurring, but not due to timing.

Three themes emerged from student comments, regardless of timing: classroom climate, achievement striving and goal attainment, and operational deliverables. Students provided insights into the essence of their learning, illustrating how instructors facilitated or impeded achievement of student goals and providing their perceptions of the curriculum. Regardless of the focus of student comments, they were more positive than critical before and after final exams.

Alternatives to increase SET response rates and to refine and improve programs to assess teaching are provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

The importance of assessment and measurement at all levels of higher education has increased over the last 25 years, partly because of higher stakeholder expectations of accountability (Chapman & Joines, 2017; Estelami, 2015; Leveille, 2006; Lombardi, 2013; Warman, 2015). It is important to assure higher education stakeholders–students and their families, administrators, accrediting bodies, governing boards, and legislators–that colleges and universities are fulfilling their duty to educate students and prepare graduates (Chapman & Joines, 2017; Estelami, 2015; Lombardi, 2013; McCarthy, 2012). Student evaluations of teaching (SETs) are a form of assessment to measure the success of students' education, and interest in SETs has made them one of the most researched and debated topics in academia (Beran, Donnon, & Hecker, 2012; Clayson, 2009). Illustrating this interest, in September 2020, the ERIC database returned 4,029 results using the search phrase "student evaluation of teacher performance," which is the designated ERIC descriptor for SETs (Benton & Cashin, 2010).

Although SETs were originally intended to provide formative feedback that instructors could use to identify strong and weak points in their teaching methods and courses, SETs quickly evolved to include a summative assessment of teaching quality (Algozzine et al., 2004; Schiekirka et al., 2012). This summative function is the source of the most controversy surrounding SETs because they are now used by administrators in making decisions about faculty promotion, tenure, and salary (Algozzine et al., 2004). In 1998, leading SET researcher Peter Seldin (1999a) surveyed 598 deans throughout the United States and found that 97.5% of them considered didactic teaching to be a major factor in overall faculty performance. In that

same study, 88.1% of deans believed that the best information source to evaluate teaching performance was systematic student ratings (Seldin, 1999a).

In the last 15 years, online completion of SETs has become the norm, replacing the prior practice of using paper SETs administered in the classroom (Algozzine et al., 2004; Berk, 2005; Hativa, 2013). When SETs began to be completed online, a reduction in response rate was seen across higher education (Adams, 2012; Chapman & Joines, 2017; McAlpin et al., 2014; McCarthy, 2012; Porter & Whitcomb, 2005; Reisenwitz, 2016; Wilson & Ryan, 2012; Young, Joines, Standish, & Gallagher, 2019). The reason for the reduction in response rate for online SETs is largely unknown, but the decrease is speculated to be partially due to no longer having a captive audience, possible technical difficulties, the time needed to respond outside of class, and students' belief that their feedback is inconsequential (Adams, 2012; Benton & Cashin, 2010; Estelami, 2015; Young et al., 2019).

Timing of SET Solicitation

Online SETs allow for more flexibility for students to complete SETs at their own convenience. Students are given several days or weeks during which they may complete their SETs at any time and any location; those who administer SETs must decide when the window for student completion will open and close (Estelami, 2015). Most often, SETs are due before the final exam period begins (Arnold, 2009; Estelami, 2015; Hativa, 2013; Hoefer, Yurkiewicz, & Byrne, 2012; Young et al., 2019). A reason for this practice likely stems from a time in which students completed SETs on paper in the classroom–instructors did not have time to administer an SET during a final exam period, and students were likely too stressed to focus on anything except their exam (Arnold, 2009; Estelami, 2015). Another reason for this practice, though, is to

avoid any potential negative effect based on the difficulty of the final examination or students' knowledge of their final grade (Arnold, 2009; Schiekirka et al., 2012). Results from studies on the relationship between grades and SETs are conflicting, but several researchers have shown an association between expected grade and SET scores (Aleamoni, 1999; Clayson, 2009). A positive correlation between expected and/or actual grades and SET scores has been found to range between .14 to .47 (Aleamoni, 1999; Hativa, 2013; Marsh & Roche, 1997). In those studies, as expected or actual final grades increased, so did SET scores, and vice versa. Although these correlations were statistically significant, in most of these studies, the correlations were interpreted as practically small, according to standards for behavioral science research (Hativa, 2013; Hinkle, Wiersma, & Jurs, 2003).

Many researchers have investigated the relationship between SET and the grades students *expect* in a course, possibly due to the standard practice of having students complete SETs before final exams when they do not yet know their actual grades (Hativa, 2013). Despite the mass of research on students' expected grade correlation with SETs, relatively few researchers have sought to understand the effect of students knowing their *actual* grades (Cho & Cho, 2017; Hativa, 2013). In studies of undergraduate students who knew their actual grade, timing of evaluations (after final examinations) was shown to influence SET scores–some items negatively and other items positively (Arnold, 2009; Hoefer et al., 2012). Researchers in Korea also showed that students who knew their final course grade before completing an SET tended to give lower scores than students who did not know their grade (Cho & Cho, 2017). In contrast, in two separate studies conducted in medical schools, researchers found that timing did not affect SET scores (Canaday, Mendelson, & Hardin, 1978; McNulty et al., 2010). In one of those studies,

students tended to give fewer and less substantial open-ended comments after final exams (McNulty et al., 2010). Additionally, as Estelami (2015) pointed out, studies using actual student grades as a variable in analysis are difficult because researchers are rarely able to connect actual student grades to anonymous SET responses. Therefore, the research on how timing, and thus student knowledge of grade, affect SETs is inconclusive.

Statement of the Problem

In veterinary medical schools, due to team teaching in many pre-clinical courses, students have more instructors to evaluate than typical undergraduate and graduate students (Malone, Root Kustritz, & Molgaard, 2018). Team teaching occurs when one course is taught by more than one instructor. This practice is common across veterinary and human medical schools because of the need for multiple content experts within a given course (McNulty et al., 2010). Researchers at one veterinary school reported requesting over 50 SETs from each student, each semester (Malone et al., 2018). A decrease in response rate has been shown to occur when 11 or more SETs are administered (Adams & Umbach, 2010), and so the team-teaching nature of pre-clinical veterinary courses naturally lowers response rates.

Nonresponse for SETs threatens the reliability and validity of results (Benton & Cashin, 2010; Hativa, 2013; Malone et al., 2018; Richardson, 2005). Therefore, nonresponse bias in SET results is a concern; it has been postulated that providing students the opportunity to complete SETs after final exams might help increase response rates (Bacon, Johnson, & Stewart, 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019). Additionally, students at a German medical school preferred to be given time to complete SETs after final exams (Schiekirka et al., 2012);

however, it is still standard practice at universities in the United States to close SETs before final exams (Arnold, 2009; Hativa, 2013; Young et al., 2019).

As noted above, in some studies, the timing of evaluations and student knowledge of final grade have been shown to influence SET scores in undergraduate and graduate student populations in the United States and abroad (Arnold, 2009; Cho & Cho, 2017; Hoefer et al., 2012). One possible rationale for this phenomenon is attributional bias, which states that people attribute their successes to themselves and their failures to someone else (Arnold, 2009; Korn, Rosenblau, Rodriguez Buritica, & Heekeren, 2016; Robinson, 2017). Attributional bias often is more common in highly successful people (Tetlock & Levi, 1982; Weary, 1979; Weiner, 2010). Veterinary medical students have been described as highly competitive, "elite performers"; therefore, the success-driven characteristics of this student population suggest that they would be likely to display attributional bias (Tetlock & Levi, 1982; Zenner, Burns, Ruby, DeBowes, & Stoll, 2005).

Purpose of the Study

The purpose of this study was to determine how administering SETs to veterinary students after final exams would affect item scores, completion rate, and comments. Knowing how timing and student knowledge of grades affect SET results in a sample of veterinary students could help guide veterinary college administrators, faculty, and assessment staff on best practices for timing of availability of evaluations to students. If response rates significantly increased when students completed SETs after final exams, versus before, but rating scores and themes in student comments did not differ, colleges could consider closing SETs after finals to help reduce nonresponse bias. However, if response rates did not improve and/or scores or

comments differed significantly between the before and after finals groups, the results could help guide colleges toward other solutions to low response rates. Additionally, if attributional bias seemed to be occurring, college administrators could examine ways to better train students on how to give constructive feedback (LaBelle & Martin, 2014).

Research Questions

With attributional bias informing and influencing the research, the following research questions guided this study:

RQ1. (*Quantitative*) How does veterinary student completion of SETs after final exams (versus before final exams) affect numerical scores on SET items?

RQ2. (*Quantitative*) How does administering SETs to veterinary students after final exams (versus before) affect completion rate and volume of comments?

RQ3. (*Qualitative*) How does veterinary student completion of SETs after final exams (versus before) affect the qualitative substance of comments?

RQ4. (*Mixed*) How does veterinary student completion of SETs after final exams (versus before) affect the intensity of comment themes and magnitude of sentiment?

Conceptual Framework

The conceptual framework underlying this study was attributional bias. A conceptual framework is a theoretical rationale that specifies relationships among all components of a study and organizes these components (Anfara & Mertz, 2015a, 2015b; Creswell, 2014). Mixed methods studies, such as this one, rely on conceptual frameworks to guide the research questions and inform the findings (Anfara & Mertz, 2015a; Creswell, 2014).

Researchers have frequently proposed that two theories affect SET results: instructor leniency and instructor-learner reciprocity (Clayson, 2009). The leniency hypothesis indicates that learners give more favorable evaluations to instructors whose grading is more lenient; the reciprocity hypothesis expects students with higher grades to provide more positive evaluations (Clayson, 2009). Although some support for both theories exists, much of the research on leniency theory has shown that instructor leniency does not affect how learners evaluate them (Algozzine et al., 2004; Clayson, 2009; Marsh & Roche, 1997). Research supporting the reciprocity theory is stronger and suggests that students who perform well in a course reward instructors with good SET scores, whereas students who perform poorly punish instructors with poor SET scores (Clayson, Frost, & Sheffet, 2006). Conceptually related to reciprocity is attributional theory and bias.

Attributional theory began with psychologist Fritz Heider in 1958 (Weiner, 2010). The theory originated with the idea that people search for causality and understanding, and the theory was often examined in school settings where success and failure for achievement tasks could be observed easily (Weiner, 1979). Causal attribution has been shown to occur for both successes and failures (Zuckerman, 1979). Weiner and colleagues determined that if someone's success was perceived to be based on assistance from other people, the resulting emotion was most often reported to be gratitude (Weiner, 1979). However, if one's failure was attributed to others, the resulting emotion was anger (Weiner, 1979). Students have been found to search for the causality of failure more so than of success, and this search often takes place at a subconscious level (Weiner, 1979).

With attributional bias, also called self-serving bias, people internalize their own success by attributing it to themselves; they externalize their failure by attributing it to someone else (Arnold, 2009; Korn et al., 2016; Robinson, 2017; Zuckerman, 1979). Jhangiani and Tarry (2011) defined two types of attributional bias: personal and situational. When success is internalized, the bias is personal, and when failure is externalized, the bias is situational (Jhangiani & Tarry, 2011). With SETs, a situational (external) attributional bias has been proposed to occur with students who believe they have performed poorly in a course (Jhangiani & Tarry, 2011). When attributing failures, students might be attempting to make sense of situations, shield or boost their self-esteem, and/or assign responsibility (Greenberg, Pyszczynski, & Solomon, 1982; Jhangiani & Tarry, 2011; Tetlock & Levi, 1982; Zuckerman, 1979). Tetlock and Levi (1982) wrote that the self-esteem hypothesis "assumes that people need to protect, confirm, and perhaps even enhance their feelings of personal worth and effectiveness" (p. 75). To attribute one's failure to lack of ability or aptitude elicits feelings of guilt, regret, and shame (Weiner, 2010).

Protecting self-esteem is especially important for those who expect success more than failure (Tetlock & Levi, 1982; Weary, 1979; Weiner, 2010). High success expectancy is one reason success is ascribed internally to a greater extent than is failure (Zuckerman, 1979). Several pieces of data support the idea that veterinary students expect success. First, admission to veterinary medical school is incredibly competitive. Of the 8,152 applicants nationwide to the class of 2022, only 3,456 (42.3%) were admitted, and 20% of admitted applicants were applying for a second or third application cycle (Association of American Veterinary Medical Colleges, 2020b). Second, the cumulative GPA of students admitted in 2019 ranged between 3.2–3.8 on a

4.0 scale (Association of American Veterinary Medical Colleges, 2020a). Finally, veterinary education in the United States is expensive, which likely increases pressure for students to perform well. In 2019, U.S. veterinary graduates reported approximately $175,000 in mean debt (Association of American Veterinary Medical Colleges, 2020b). Based on the evidence cited above, the success-driven characteristics of veterinary students suggest that this population would be likely to display attributional bias.

LaBelle and Martin (2014) looked at attribution theory in a different and perhaps more enlightening way. They incorporated instructional dissent into attribution theory using Goodboy's (2011) Instructional Dissent Scale, which divides dissent into three types: expressive, vengeful, and rhetorical (LaBelle & Martin, 2014). With expressive dissent, a student seeks to better their emotional state by expressing their emotions (LaBelle & Martin, 2014). Vengeful dissent involves enacting revenge on an instructor, and rhetorical dissent is an effort to convince an instructor to right a wrong (LaBelle & Martin, 2014). Expressive dissent occurs mostly when students discuss instructors or the course with peers, while rhetorical dissent occurs when students discuss their concerns directly with instructors (Goodboy, 2011). Goodboy (2011) found that instances of perceived classroom injustice triggered student dissent. If applied to SETs, most student comments are likely in the rhetorical dissent category because students' instructors (rather than students' peers) will read the comments. Vengeful dissent might also occur if students know that instructors' superiors read student comments.

Few studies have been reported on attributional bias as it directly relates to student performance. In a study on exam results of male, undergraduate students, Greenberg and colleagues (1982) found that students who scored highly on a test attributed the outcome to their

ability and effort, but the low-scoring students ascribed the outcome to bad luck. In Arnold's (2009) study, students who performed poorly in a course rated instructors' statistically lower than did students who performed well, but only after they had taken their final exam, supporting the theory of a self-serving bias.

In the current study, attributional bias was considered during data analysis and interpretation of results to provide insight into SET item scores and substance of student comments before and after final exams.

Significance of the Study

Students are more uniquely situated to consistently evaluate teaching than faculty peers or administrators. When students evaluate teaching, they are participating in longitudinal evaluation and have a more complete perspective than peers and administrators, whose perspective is often cross-sectional. The content knowledge of peers and administrators hinders their ability to fully appreciate the perspective of a novice (Hativa, 2013; Theall & Franklin, 2001). Data from retrospective studies indicate that students are as effective at evaluating teaching as recent graduates (Hativa, 2013; Theall & Franklin, 2001). However, low response rates threaten the reliability of SETs by introducing nonresponse bias (Clayson, 2009; Hativa, 2013). It is important to maintain adequate response rates to be able to trust results that are used to improve faculty teaching practices and inform promotion and tenure decisions. This study sought to extend the understanding of how timing, and thus knowledge of grade, affect SETs completed by veterinary students. Such knowledge could help guide veterinary college administrators, faculty, and assessment staff on best practices for timing of release and closure of evaluations to achieve consistent and trustworthy response rates.

Specifically, if SET response rates increased after final exams but no significant differences were seen in rating scores or comments, colleges should consider extending the deadline for students to submit SETs to help reduce nonresponse bias. However, if response rates did not significantly improve and/or rating scores or comments differed significantly between the before and after finals groups, colleges should consider other solutions to low response rates. If attributional bias seemed to be occurring, colleges could examine ways to better prepare students to provide constructive feedback, as well as recognize instructional dissent (LaBelle & Martin, 2014). Finally, this research could help guide practices at other health professional schools.

Terminology

The following definitions of terms and concepts are used within the study:

Attributional bias, also called self-serving bias, is a phenomenon in which people attribute their own successes to themselves and attribute their failures to someone else (Arnold, 2009; Korn et al., 2016; Robinson, 2017).

Didactic teaching is instructor centered. An instructor delivers content to students with the goal of knowledge transfer. In the context of this research, didactic teaching refers to pre-clinical teaching, which is delivery of foundational knowledge that students will apply later in their clinical education.

Formative feedback, when used in the context of SETs, is information received from students that allows instructors to identify strong and weak points in their teaching methods and identify strategies to improve their teaching (Algozzine et al., 2004; Schiekirka et al., 2012). Such feedback can occur at any time during instruction.

Nonresponse bias is a phenomenon that occurs when those who do not respond to surveys or questionnaires would have provided systematically different responses from those who did respond (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019).

Student evaluation of teaching (SET) is a form of accountability and assessment used to measure the quality of instruction (Beran et al., 2012; Clayson, 2009). In the context of this research, SETs are student evaluations of instructors' teaching.

Summative feedback, when used in the context of SETs, is information shared with an instructor's superior to allow evaluation of an instructor's teaching effectiveness (Algozzine et al., 2004; Schiekirka et al., 2012). This type of feedback is typically collected at the conclusion of a given instructional encounter or course.

A **teaching portfolio** is a collection of documents compiled by faculty as evidence to holistically demonstrate and explain their teaching experience and practices. Such evidence might include SET results, syllabi, student interviews, videos of class sessions, teaching scholarship and awards, and student achievement of learning outcomes (Algozzine et al., 2004; American Sociological Association, 2019; Berk, 2005; Seldin, 1999a; Zubizarreta, 1999).

Team teaching is the practice of more than one instructor teaching one course. This practice is common across veterinary and human medical schools because of the need for content experts within a given course (McNulty et al., 2010).

Summary

This chapter introduced the research problem, the conceptual framework for the research, the purpose of the study, and definitions of terms. It is important to assure to higher education stakeholders that colleges and universities are fulfilling their duty to educate students (Chapman

& Joines, 2017; Estelami, 2015; Lombardi, 2013; McCarthy, 2012). Although findings have been conflicting, SETs have generally been found to be valid and reliable measures of teaching effectiveness, for both summative and formative purposes (Adams, 2012; Addison & Stowell, 2012; Hativa, 2013; McCarthy, 2012). Nonetheless, nonresponse bias remains a concern, and permitting veterinary students to complete SETs after final exams could help increase response rates (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019). Of concern is that several researchers have shown an association between student grade and SET results (Clayson, 2009). This study seeks to determine how administering SETs to veterinary medical students after final exams (versus before final exams) affects results, with attributional bias theory as the conceptual framework. Little research exists on SETs in veterinary medical education literature, and no studies have been identified with regard to how timing and actual knowledge of grade affect SET results in veterinary medical schools.

Chapter 2 presents a review of historic literature and current research on SETs in higher education and health professional education. Chapter 3 describes the methodology, research design, and study procedures. Chapter 4 details results from data analysis and provides a summary of the results, and chapter 5 interprets and discusses the results and describes future opportunities for research.

CHAPTER TWO
LITERATURE REVIEW

This chapter begins with a historical overview of SETs in higher education, from their origin to their current use as summative and formative assessments of teaching quality (Addison & Stowell, 2012). Following a brief discussion of faculty perceptions of SETs, the chapter provides a summary of SET response rates and how they affect SET reliability and validity. Then, the most commonly researched biases are reviewed, including bias related to student knowledge of grades. Finally, the chapter ends with a description of how SETs are used in veterinary medical schools and how this study contributes to the literature.

A Brief History of Student Evaluations of Teaching

Student evaluations of teaching in higher education courses are thought to have originated in the 1920s with the work of Max Freyd at the University of Pennsylvania (Addison & Stowell, 2012). During the same decade, Hermann Henry Remmers and colleagues developed the first SET instrument (Addison & Stowell, 2012; Algozzine et al., 2004; Wachtel, 1998). Remmers was an educational psychologist at Purdue University, and thus the resulting SET instrument was called the Purdue Rating Scale for Instructors (Addison & Stowell, 2012; Algozzine et al., 2004; Gage, 1969; Smalzried & Remmers, 1943; Wachtel, 1998). The Purdue scale was meant to be used formatively by instructors to improve their teaching by enabling them to recognize their "specific limitations and assets" (Smalzried & Remmers, 1943, p. 363). Recognizing that teaching is a multidimensional, complex construct, Remmers and his colleagues closely examined personal qualities that the literature (at that time) linked to teaching success (Algozzine et al., 2004; Marsh & Roche, 1997; McCarthy, 2012; Smalzried & Remmers,

1943). Later, Smalzried and Remmers (1943) conducted a factor analysis on a version of the Purdue Rating Scale that contained 10 items that they called “traits” (Addison & Stowell, 2012). This early analysis included such traits as fairness in grading, personal appearance, and sympathetic attitude toward students (Smalzried & Remmers, 1943). Smalzried and Remmers (1943) found that their 10 items loaded onto two factors that they named the empathy trait and the professional maturity trait.

The work of Remmers and his colleagues dominated the SET literature and influenced research until the 1950s, when Wilbert J. McKeachie at the University of Michigan began to expand earlier work and shift the focus to examine summative use of SETs (Addison & Stowell, 2012). Importantly, McKeachie recommended that administrators consider the data from SETs in context with other forms of information about teaching, such as self-assessment and peer review, to, in essence, triangulate the data and provide administrators with a more comprehensive basis for evaluation of faculty (Addison & Stowell, 2012; McKeachie, 1969; Nasser & Fresko, 2002; Wilson & Ryan, 2012). He and others have insisted that the way in which SET data are used is the best approach to ensuring reliability and validity of results (Cashin, 1999; DeZure, 1999; Hammer, Peer, & Babad, 2018; Hativa, 2013; Ismail, Buskist, & Groccia, 2012; Keeley, 2012; Marsh & Roche, 1997; McCarthy, 2012; Schafer, Hammer, & Berntsen, 2012; Seldin, 1999b; Wilson & Ryan, 2012). Others have suggested adding teaching portfolios to evaluate teaching more holistically alongside SETs; these portfolios could include evidence such as student interviews, videos of class sessions, teaching scholarship and awards, and student achievement of learning outcomes (Algozzine et al., 2004; American Sociological Association, 2019; Berk, 2005; Seldin, 1999a; Zubizarreta, 1999).

By the 1970s, SETs were starting to become commonplace in higher education, and thus began what has been called the golden age of SET research (Algozzine et al., 2004; Royal, 2017). Herbert Marsh and colleagues had begun to contribute to SET research with the Student Evaluation of Educational Quality (SEEQ) instrument (Addison & Stowell, 2012). The SEEQ contained nine factors that Marsh developed based on reviews of other instruments, interviews with students and instructors, and psychometric analyses (Marsh & Roche, 1997). The nine factors, or constructs, of Marsh's instrument included (1) learning/value, (2) instructor enthusiasm, (3) organization/clarity, (4) group interaction, (5) individual rapport, (6) breadth of coverage, (7) examinations/grading, (8) assignments/readings, and (9) workload/difficulty (Algozzine et al., 2004; Beran et al., 2012; Keeley, 2012; Marsh & Roche, 1997). Marsh and others found that the SEEQ instrument was reliable, valid, and unaffected by potential biases such as class size and grading leniency (Addison & Stowell, 2012; Hativa, 2013; Keeley, 2012; Marsh & Roche, 1997).

From 1973 to 1993, the percentage of institutions employing SETs increased from 29% to 86% (Hoefer et al., 2012). Since the 1980s, SET research has been primarily concentrated on clarifying and strengthening previous findings, and several review articles and meta-analyses have been published (Algozzine et al., 2004; Benton & Cashin, 2010; Beran et al., 2012; Uttl, White, & Gonzalez, 2017). These articles debunked persistent misconceptions, confirmed correlations between SET scores and student learning, and included evidence-based recommendations on SET planning and implementation (Algozzine et al., 2004; Benton & Cashin, 2010; Beran et al., 2012; Uttl et al, 2017). Research into potential biases (related to both student and teacher) influencing SET data has also been a popular pursuit, as have studies about

statistical reliability and validity, particularly regarding student completion of SETs online, which has become the norm within the last 15 years (Algozzine et al., 2004; Hativa, 2013).

Initially, SETs were completed in person, on paper, at the end of a class session near the end of an academic term, and results were compiled manually (Algozzine et al., 2004). In the 1970s, schools began to collect SET data from students using a #2 pencil and paper optical mark recognition sheets (commonly known as bubble sheets) (Bogardus Cortez, 2016). These sheets were then processed using a scoring machine, such as Scantron, which compiled mean scores more quickly than manual entry (Bogardus Cortez, 2016). Although SET item scores could be compiled more easily using scanned mark sheets, students still hand-wrote comments on sheets of paper separate from the mark sheets, and these hand-written or transcribed comments were shared with instructors (Howell, 2002). The move to online SETs in the 2000s gave students more flexibility regarding response time and location of completion (Adams, 2012). Online SETs are also much more cost effective and efficient than paper SETs and allow faster data compilation and distribution of results (Adams, 2012). Students are also reported to provide more comments in online SETs because they view typing as easier than writing, and their instructor will not be able to recognize their handwriting (Adams, 2012). Nonetheless, with online SETs came a new controversy: lower response rates, and faculty and administrator fear of non-response bias and decreased reliability and validity (Adams, 2012; Hativa, 2013; Howell, 2002; Reisenwitz, 2016; Wilson & Ryan, 2012).

Faculty Attitudes toward SETs

A quick search in *The Chronicle of Higher Education* will illustrate the general higher education faculty attitude toward SETs. In the last 3 years, *The Chronicle* published 26 articles

related to SETs, and most of them portray SETs negatively. In a 2018 commentary piece, Nancy Bunge called them unreliable, contaminating, and troubling. Bunge (2018) pointed to perceived biases related to faculty gender and age, generous grading due to lowered standards, and students who put complete responsibility of their learning onto the instructor. Bunge's commentary seems to be representative of the attitudes of many faculty. Mary Lou Higgerson (1999) suggested that faculty fear of how the results will be used was a source of cynicism. Some researchers have pointed to the famous "Dr. Fox" study in 1973 as a source of faculty distrust in SETs (as cited in Hammer et al., 2018; Marsh & Roche, 1997). The Dr. Fox study was conducted at a teacher training conference for mental health educators, such as psychiatrists and social workers, and in a graduate course in philosophy (Naftulin, Ware, & Donnelly, 1973). In it, an actor posing as an expert delivered an engaging lecture that lacked actual content; resulting evaluations indicated Dr. Fox had stimulated participants' thinking and used enough examples to clarify his material (Hammer et al., 2018; Naftulin et al., 1973). Indeed, in 1975, an associate professor at Michigan State University's College of Veterinary Medicine wrote the following in a peer-reviewed article:

> The increasingly accepted concept of the student as consumer with a right to evaluate what he gets for his tuition dollar has eroded the previous immunity of the veterinary teacher to any but self-judgment of his competence. The presumption is that the veterinary student, as learner, is able to evaluate certain things about instructor performance. Those who must make decisions about retention, advancement, and tenure are too eager to accept this position, perhaps because it relieves them of the onerous chore of taking direct responsibility for evaluation themselves. (Getty, p. 25)

Empirical studies about faculty attitudes toward SETs are scarce, but in a 2019 study, many faculty participants perceived SETs as a platform for students to vent frustrations, be vindictive, and express anger, at the detriment of faculty promotion and tenure (Eckhaus & Davidovitch, 2019; Nasser & Fresko, 2002). In a 2018 study, 73% of the faculty members surveyed agreed with the myth (authors' description) that "students use evaluations in a vindictive manner," and 56.4% felt that SETs "mostly measure the lecturer's charisma" (Hammer et al., p. 524). At the same time, 62.7% felt that the feedback they receive from SETs helps them to "identify weak aspects of" their teaching (Hammer et al., 2018, p. 525). Interestingly, instructors who most believed the myths were the same instructors whose SET values were lowest (Hammer et al., 2018). In contrast, in an earlier survey about faculty perceptions of SETs, El Hassan (2009) found that the faculty participants in that study viewed SETs as useful for formative purposes. Still, approximately 30% of faculty believed student ratings are meaningless, and nearly half felt that students are not knowledgeable enough to assess instructional quality (El Hassan (2009).

The antagonistic attitudes of some instructors tend to also be transferred to the classroom, where students hear that their feedback does not matter (Seldin, 1999b). In one study, senior faculty and males held more negative attitudes than junior faculty and females (Hammer et al., 2018). Faculty participants in a 2002 study generally believed a link exists between the quality of teaching and SETs, but they also believed that more demanding instructors receive lower ratings and that students give higher ratings to nicer instructors (Nasser & Fresko). These same instructors felt that SET data were useful, particularly qualitative data (Nasser & Fresko, 2002). Faculty members with high evaluation scores tended to convey stronger confidence in the

validity of SETs (Nasser & Fresko, 2002). Additionally, in a 1993 study of veterinary faculty attitudes toward clinical SETs, the top two methods that faculty felt were meaningful in evaluating clinical teaching were peer evaluation and SETs (Fossum, Ruoff, Rushton, & Paprock, 1993). Conflicting faculty attitudes might explain why many of the SET studies in the last 20 years have been focused on dispelling myths associated with faculty viewpoints of SETs (Hativa, 2013).

Certainly, a concern with the use of SETs is that results can be damaging to instructor morale and self-image (Hammer et al., 2018; Nasser & Fresko, 2002; Warman, 2015). Instructors have reported that quantitative and qualitative SET data reduce their motivation, are harmful to promotion, and cause emotions of surprise, anger, sadness, and fear (Carmack & LeFebvre, 2019; Eckhaus & Davidovitch, 2019). Although most faculty who took part in a 2019 study agreed that poor evaluations and negative comments were rare, they noted that they tend to focus on and remember the negative more so than the positive (Carmack & LeFebvre). To help mitigate these reactions, several authors have recommended that faculty use professional teaching consultants, trusted colleagues, and/or mentors to help them make meaning of SET results and exercise reflective practice (Beran et al., 2012; Carmack & LeFebvre, 2019; Ismail et al., 2012; Marsh & Roche, 1997). Such reflective practice could help faculty to look at negative comments as teachable moments toward improvement in the classroom (Carmack & LeFebvre, 2019; Nasser & Fresko, 2002).

SET Reliability and Validity: Response Rate

For SETs, reliability most often refers to interrater agreement–within a specific class, if different students are apt to give similar scores for a specific SET item (Benton & Cashin, 2010).

Indeed, interrater reliability values increase as the number of responses increases (Benton & Cashin, 2010; Hativa, 2013). Therefore, the more students who participate, the better the reliability (Adams, 2012; Addison & Stowell, 2012; Benton & Cashin, 2010; Hativa, 2013). Quantitative validity asks if an instrument measures what it is supposed to measure (Benton & Cashin, 2010). An SET instrument is supposed to measure teaching effectiveness. To use SETs as valid evidence of teaching effectiveness, sufficient statistical power is needed to draw inferences from the data (Creswell, 2014; Royal, 2017). Statistical power refers to the probability of rejecting the null hypothesis, and 80% power and a 95% confidence level ($\alpha = .05$) are generally preferred (Royal, 2017; Tabachnick & Fidell, 2013). One way to increase reliability and validity for a survey instrument is to increase response rate (Benton & Cashin, 2010; Hativa, 2013; Malone et al., 2018; Richardson, 2005).

Most online surveys have a response rate between 20–47% (mean 33%), whereas face-to-face, paper survey responses have been reported to range from 33–75% (Nulty, 2008). Several researchers have speculated that SET response rates should be above 50% to be acceptable (Nulty, 2008; Richardson, 2005). However, a formula by Dillman (as cited in Nulty, 2008) that assumes random sampling indicates that in a class of 70 students, a 28% response rate is required for a liberal, 80% confidence level, and a rate of 91% is required for a stringent, 95% confidence level (Adams, 2012; Chapman & Joines, 2017). Nulty (2008) argues that Dillman's approach is not adequate for SETs since the sampling method is not random. For paper SETs, the sampling method would be convenience–limited to students who attended class on the day of SET completion (Nulty, 2008). For online SETs, the sampling method is self-selection.

Royal (2017) pointed to a new way of calculating a statistically sound response rate. He employed a margin of error (MOE) calculation method that uses not only sample size, level of confidence, and standard deviation, but also population size (Cavanaugh & Foster, 2010; Royal, 2017). The MOE is measured from 0 to 1.0, and a lower MOE value indicates higher reliability (Benton & Cashin, 2010). Because most national surveys have a 3% error range, Royal (2017) reasoned that an SET with a 5-point Likert-type scale would need an MOE value <0.12 to correspond to an error range of ± 3% (Royal, 2015). As the standard deviation of the sample increases, a higher response rate is needed (Bacon et al., 2016; Royal, 2017).

Usually, but not always, student response to an SET is voluntary (Adams & Umbach, 2010). In the last 15 years, online completion of SETs has become the norm over the prior practice of using paper SETs administered in the classroom (Algozzine et al., 2004; Berk, 2005; Hativa, 2013). A reduction in response rate was seen across higher education when SETs moved from paper to online (Adams, 2012; Chapman & Joines, 2017; McAlpin et al., 2014; McCarthy, 2012; Porter & Whitcomb, 2005; Reisenwitz, 2016; Wilson & Ryan, 2012; Young et al., 2019). The reason for the reduction in response rate when SETs moved online is largely unknown, but the decrease is speculated to be at least partially due to no longer having a captive audience, potential technical difficulties, students' belief that their feedback is unimportant, and the time cost of responding outside of class (Adams, 2012; Benton & Cashin, 2010; Estelami, 2015; Young et al., 2019).

Types of Bias

When compared to other measures of instructional effectiveness, SETs have generally been found to be valid and reliable, and minimally affected by common biases when used for

summative and formative purposes (Adams, 2012; Addison & Stowell, 2012; Hativa, 2013; McCarthy, 2012). They have also been found to be a function of the instructor and not the course itself (Addison & Stowell, 2012). Results from previous studies have debunked the idea that grading leniency positively affects SET data (Algozzine et al., 2004; Marsh & Roche, 1997; Nasser & Fresko, 2002). Several researchers have shown that SETs are more than measures of popularity, and that students are qualified judges of teaching (Nasser & Fresko, 2002). In a study of veterinary students in 1979, researchers found no correlation between SETs completed about *typical* instructors prior to a course and those completed about *specific* instructors at the end of the course (Engel & Fuqua). On the first day of the course, researchers asked students to evaluate typical instructors in similar courses, and on the last day, students evaluated the three specific instructors in that course (Engel & Fuqua, 1979). What has been found to positively affect (bias) student ratings includes prior student interest in the subject being taught and, perhaps surprisingly, a more difficult workload (Adams, 2012; Marsh & Roche, 1997). When examining student characteristics in relation to SET scores, Marsh and Roche (1997) found that the more students were interested in the subject prior to a course, the higher the SET scores. Additionally, SET scores were higher among students who perceived the course to have a high workload or level of difficulty (Marsh & Roche, 1997).

Several other potential sources of bias have been explored, and among these is bias related to faculty gender (Basow & Martin, 2012; Christopher & Leskosky, 1997). Researchers exploring gender bias in a veterinary student population looked at retrospective data from 1981 to 1991 and found that women faculty were more highly rated than men for every evaluation parameter (Christopher & Leskosky, 1997). All these ratings were statistically different except

those related to course handouts and emphasis of practical application (Christopher & Leskosky, 1997). Others have found that female instructors are sometimes rated lower by male students and higher by female students, and lower ratings have been seen with young, female instructors (Basow & Martin, 2012).

Nonresponse bias is a type of statistical error proposed to occur when those who do not complete the SETs could have provided systematically different responses from those who did complete the SETs (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019). Non-response for SETs threatens the reliability and validity of results (Benton & Cashin, 2010; Hativa, 2013; Malone et al., 2018; Richardson, 2005). Researchers have found that female students, students with higher GPAs, and those with more campus engagement are more likely to respond to SETs (Bacon et al., 2016; Porter & Whitcomb, 2005; Reisenwitz, 2016; Young et al., 2019). Estelami (2015) used data from late SET completers as representative of non-respondent data; in that study, scores given by late responders lowered the SET item that measures overall student satisfaction by 0.2. Studies such as those described above suggest that nonresponse bias should be a legitimate concern for interpreting SET data.

Barr, Spitmüller, and Stuebing (2008) categorized *survey* non-respondents as either active or passive. Active non-respondents make a conscious decision not to complete a survey, whereas passive non-respondents encounter extenuating circumstances, like overload of work and lack of time, that prevent them from completing a survey (Barr et al., 2008). Passive non-respondents are the most common of the two types (Barr et al., 2008). Still, active non-respondents tend to lack an understanding of how survey results will be used and to believe that survey results will not be acted upon (Young et al., 2019). Interestingly, in a study with 462 undergraduate student

participants, 29% of the sample never responded to a single survey on salient topics like drug and alcohol use and dining services (Porter & Whitcomb, 2005).

In survey research for the general population, women, more educated individuals, and those with high community involvement (or organizational citizenship) respond more to surveys than men, less educated people, and those with low community involvement (Barr et al., 2008; Chapman & Joines, 2017; Porter & Whitcomb, 2005). Similarly, several researchers have found that students with the following characteristics are more likely to respond to SETs: being female, having a higher GPA, and being more engaged on campus (Bacon et al., 2016; Porter & Whitcomb, 2005; Reisenwitz, 2016; Young et al., 2019). One way to identify potential nonresponse bias is to examine if respondent demographics are different from those who did not respond (Reisenwitz, 2016).

Student bias related to knowledge of expected or actual grade has also been explored (Aleamoni, 1999; Algozzine et al., 2004; Cho & Cho, 2017; Clayson, 2009). In a 1999 review, the author cited 24 studies that reported no significant relationship between grades and evaluations and 37 studies that reported significant positive correlations (Aleamoni, 1999). In that review, Aleamoni (1999) noted that across studies, a median positive correlation of .14 was found between expected/actual grades and SET scores (Algozzine et al., 2004). Marsh and Roche (1997) found the correlation to be between .20 and .30, and others saw correlations from .43 to .47 (Hativa, 2013). Although these correlations were statistically significant, a correlation (r) value between .30–.50 is generally interpreted in behavioral science research to be a low-to-medium correlation (Hativa, 2013; Hinkle et al., 2003). Therefore, for most of these studies, the correlation is small. Cho and Cho (2017) found that when students had more accurate knowledge

about their grades, bias increased. Students with higher GPAs also tend to give higher SET ratings (Estelami, 2015). In Reisenwitz's study (2016), the effect size, or practical significance, of GPA on response was in the moderate range (.45–.47). Other researchers have shown that students who expect better grades provide more favorable evaluations irrespective of instructor leniency (Algozzine et al., 2004; Clayson, 2009).

Compared to the research on correlation of *expected* grade with SET scores, few researchers have pursued the relationship between *actual* student grades and SET scores (Cho & Cho, 2017; Hativa, 2013). This focus is possibly due to the prevailing practice of SET completion before final exams (Hativa, 2013). Furthermore, researchers rarely know actual grades because most SETs are completed anonymously (Estelami, 2015). In Arnold's (2009) study, completion of SETs after final examinations was shown to affect SET scores in a group of undergraduate students. Significantly lower scores were seen after finals with items that focused on course relevance and clear objectives and materials (Arnold, 2009). Further, students who failed the course also rated instructor competence significantly lower after final exams than before (Arnold, 2009). Arnold (2009) focused on within-class variation in SET scores to better control for possible leniency bias between instructors. The instructors in that study also used multiple choice exams, thus allowing for more objective grading (Arnold, 2009).

In contrast, Canaday and colleagues (1978) conducted a study in a medical school using paper SETs and found that timing did not affect SET scores. They compared SETs immediately before and after final exams, as well as after students received their grades (Canaday et al., 1978). Researchers at a different medical school found that scores from SETs completed before final exams were not statistically different from those completed after final exams, but that,

overall, students gave fewer and less extensive open-ended comments after final exams versus before (McNulty et al., 2010). Similarly, in a study in a business school (including graduate and law students) in the United States, researchers found a significantly positive correlation between overall SET response and students' average final grade (Hoefer et al., 2012). Some researchers have looked at whether student learning is correlated with SET results (Clayson, 2009). In a 2009 meta-analysis of 17 studies from 1953 to 2007, Clayson found that the median correlation between learning and SET scores was .41 and ranged from -.75 to .91. These 17 studies were based on students' actual test results rather than on student perceptions of their own learning or expected grades (Clayson, 2009).

Use of SETs in Colleges of Veterinary Medicine

Within the last 15 years, health professional schools, including veterinary medical colleges, have also begun to examine the effectiveness of SETs as formative and summative assessments (Beran et al., 2012; McNulty et al., 2010; Royal, 2017; Schiekirka et al., 2012). Instruments used in undergraduate programs and traditional graduate programs are not thought to be suitable for graduate research programs, and, by extension, professional programs (Richardson, 2005). In the United States, there are 30 accredited veterinary schools (American Veterinary Medical Association, 2020b). Colleges of veterinary medicine use SETs for evaluating both didactic teaching in a classroom setting and clinical teaching in a hospital setting (Fossum et al., 1993). In health professional programs, teaching effectiveness, measured in part by SETs, affects student learning, the transfer of learning to the workplace, and, in veterinary medicine, eventually to the profession itself and society–clients, animal health and welfare, and food safety and security (Beran et al., 2012; Mosier, Cashin, & Elmore, 1993). Still, the

veterinary medical education literature on SETs is sparse (Beran et al., 2012; Royal, 2017). The *Journal of Veterinary Medical Education* has existed only since 1974. The literature that does exist is mostly focused on reliability and validity in a review format; empirical studies are uncommon (Beran et al., 2012; Hofmeister, 2021; Royal, 2017).

In a 1981 article out of Louisiana State University's School of Veterinary Medicine, researchers looked at criteria for rating teaching effectiveness and found that as students progressed in the program, they became less concerned with exams and more concerned with faculty attitude toward students (Cawunder & Tasker). The foci of a 2004 study in a French veterinary school were factors students and peer teachers felt should be evaluated in pre-clinical teaching (Toma, Begon, Fontain, & Rosenberg, 2004). Similarly, Ruoff, Fossum, Rushton, and Paprock (1992) developed a 3-factor questionnaire for evaluation of clinical teaching. In a 2021 qualitative study, non-participant student observers evaluated pre-clinical courses by developing a summary and critical analysis of their experiences after attending at least one lecture or laboratory by a different instructor each day for 2 weeks (Hofmeister). Hofmeister's (2021) recommendation was to use such case studies to provide additional insight into evaluation of teaching. A few veterinary medical education-related SET articles were published in the 1990s, but those authors focused on clinical teaching evaluation and SET practices (Fossum et al., 1993; Hubbell & Hudson, 1995; Mosier et al., 1993).

Veterinary education researchers tend to rely on studies conducted at other health professional schools, such as medical, dental, and nursing schools (Beran et al., 2012; Hofmeister, 2021; Ruoff et al., 1992). For example, at a medical school in the Netherlands, Stalmeijer, Dolmans, Wolfhagen, Muijtjens, and Scherpbier (2010) developed the Maastricht

Clinical Teaching Questionnaire (MCTQ) that is geared specifically toward clinical teaching. The MCTQ was created with apprenticeship learning in mind and is based on five factors: modeling, coaching, articulation, a safe learning environment, and exploration (Stalmeijer et al., 2010). Medical school studies on clinical education can be especially valuable due to the parallels between human medicine and veterinary medicine training programs (Ruoff et al., 1992). However, academic discipline is shown to be a significant variable in SET data analysis; therefore, whether clinical SET studies conducted at medical schools will translate well to pre-clinical veterinary medicine is unknown (Cashin, 1999; Clayson, 2009).

In a 1993 study, Mosier and colleagues looked at the use of teaching evaluation systems in 27 different veterinary schools in the United States and Canada. At that time, teaching evaluations were mostly used summatively rather than formatively (Mosier et al., 1993). At the site for the current study, SETs were first documented in 1998 (Bailey, 2016). However, SETs were first mentioned in the *Journal of Veterinary Medical Education* in 1975 as "one of those ideas whose time is come" (Getty, p. 25). Given that the Louisiana State veterinary school was reportedly using SETs in 1980, and the Alfort veterinary school in France began SETs in 1986, it is safe to assume that most veterinary schools implemented SETs sometime in the 1980s or 1990s (Cawunder & Tasker, 1981; Toma et al., 2004). Today, SETs are a requirement for accreditation through the American Veterinary Medical Association's (AVMA) Council on Education (2020a).

Some veterinary schools separate their SET processes from those of their university, including the site of the current study and The Ohio State University College of Veterinary Medicine (n.d.). Reasons for using separate SET systems are due to differences in veterinary

professional curricula in comparison to other graduate programs. Not only do veterinary schools use SETs for clinical teaching, but university timetable software often limits the number of instructors that may be evaluated in a semester. This is problematic for veterinary schools because veterinary courses are often team taught. Team teaching is common in veterinary schools across the country because it allows students to experience different perspectives and specialized content knowledge (Boston University Center for Teaching & Learning, n.d.).

Due to the team-teaching nature of veterinary courses, veterinary students have more instructors to evaluate than typical undergraduate and graduate students. At the site of the current study, each semester, pre-clinical veterinary students receive an average of 34 different SETs to complete: one for each course, one for each instructor who provides at least four lectures and/or labs in a given course, and one for the semester as a whole to identify fit within the curriculum. Malone and colleagues (2018) reported requesting over 50 SETs from each student, each semester at the University of Minnesota. In 2020, administrators at the Virginia-Maryland Regional and University of Georgia colleges of veterinary medicine reported 25–80% completion rates and described wide variability between class year and courses (S. Clouser, personal communication, May 18, 2020; S. Ryan, personal communication, May 18, 2020). In 2018, Kenneth Royal estimated in a personal communication (May 18, 2020) that most veterinary schools get a 20–25% response rate. Royal (2017) created lookup tables for common veterinary class sizes to assist faculty and administrators in understanding the response rate needed for valid data. For a class size of 80, at least a 30% response rate is needed for a margin of error of 0.117 (Royal, 2017). Interestingly, administrators at North Carolina State University's College of Veterinary Medicine (NCSU-CVM) saw their response rate increase when they

moved their SETs online (M. W. Hedgpeth, personal communication, May 18, 2020). To increase the response rate at NCSU-CVM, Royal developed a model that includes a designated time for students to complete online evaluations and alternated evaluation of instructors so that not every instructor is evaluated every year (personal communication, May 18, 2020).

Clinical veterinary students often receive SETs for both clinical instructors and clinical rotations. Clinical instructors might include not only the attending clinician but also interns, residents, and veterinary technicians (nurses). In contrast to pre-clinical student response rates, students in their fourth, clinical year at the study site complete SETs at a rate of about 98%. The increase in completion rate between pre-clinical and clinical SETs could be because, at the end of each 2-week rotation, the college withholds written summative feedback and preliminary grade reports until students complete their SETs. Students may choose to not complete SETs if they wish to wait until the university releases grades at the end of the semester, which could be months after completion of a given rotation. Ultimately, this means that if a student never completes an SET during their clinical training, they might never receive written feedback to help them improve. Many other veterinary schools use the same evaluation software as the study site (one45); therefore, it is likely that other schools use this same procedure. Although some researchers question the ethics of mandatory SETs, when schools have required students to complete SETs as an incentive to obtain their course grades, response rates have increased exponentially (Reisenwitz, 2016; Royal, 2017; Wilson & Ryan, 2012). Some medical schools withhold grades, progression, or transcripts if students do not complete SETs (Estelami, 2015; Royal, 2017).

The American Veterinary Medical Association's (2020a) Council on Education requires that all U.S. veterinary schools maintain a committee that reviews the curriculum, including "sufficient qualitative and quantitative information" to ensure "instructional quality and effectiveness" (p. 27). At the study site, complete evaluation results for courses and rotations are also reviewed by department heads and the academic dean. This process is likely similar at veterinary schools across the country (Fossum et al., 1993). Outside the United States, the University of Bristol (United Kingdom) shares SET summaries with its Faculty Quality Enhancement Team (Warman, 2015).

Contribution to the Literature

Few studies exist on how timing of SET administration after final exams affects SET results (Arnold, 2009; Canaday et al., 1978; Cho & Cho, 2017; Estelami, 2015; Hoefer et al., 2012; McNulty et al., 2010). Estelami (2015) pointed to this avenue of study as an interesting opportunity for additional research. In addition, the veterinary medical education literature on SETs is sparse, and, to-date, no studies have been identified regarding attributional bias nor the effect of timing and knowledge of expected or actual grade on SETs in veterinary schools. This study could help fill those gaps in the literature, as well as extend the understanding of how timing and knowledge of grade affect self-serving/attributional bias theory in a different population of students. Most previous studies regarding timing of SETs and student knowledge of grade were conducted with undergraduate and non-veterinary graduate students.

The veterinary student experience is quite different from that of undergraduate and non-veterinary graduate students, even with respect to class size. In the United States, veterinary class sizes range from 66 to 220 students per cohort, and core courses are taken as a cohort rather than

broken into sections (Association of American Veterinary Medical Colleges, 2020b). Although a few researchers have studied medical student SETs, veterinary student demographics differ from medical students (Canaday et al., 1978; McNulty et al., 2010). Females constitute 52% of medical school matriculants, whereas approximately 82% of veterinary students are female (Association of American Medical Colleges, 2020; Association of American Veterinary Medical Colleges, 2020b). Due to these differences in student population, it is unknown if allowing veterinary students to complete SETs after final exams will affect completion rate or if knowledge of grade and timing will result in bias in responses to SET items. That veterinary students receive instruction in communication and professionalism adds an extra dimension to how these students might differ from undergraduate and non-veterinary graduate students. In particular, Canaday et al.'s medical school study was published in 1978, before both online SETs and before professionalism was routinely taught as a non-technical competency in professional schools (Medical Schools Objectives Project Writing Group, 1999).

Degree level and academic discipline continue to be important variables in SET results across studies (Clayson, 2009; Hativa, 2013). Graduate students tend to rate instructors with higher scores than do undergraduate students (Hativa, 2013). One speculated reason for this higher rating is because graduate students are more motivated and invested in learning than undergraduates (Hativa, 2013). Another explanation is that traditional graduate courses are often smaller and more interactive than undergraduate courses (Hativa, 2013). However, the latter explanation would not apply to veterinary students in large classes in which instructors rely heavily on lecture (Hativa, 2013). Hativa (2013) also showed that instructors in science, technology, engineering, and math disciplines received lower SET scores than did instructors of

the “soft” disciplines of arts and humanities. In Clayson’s (2009) meta-analysis, the association between learning and SET score was strongest in education and liberal arts courses, but no association was seen in business courses. All sources included in that meta-analysis used exam scores as measures of student learning, and the more objective the exam, the smaller the association between learning and SET scores (Clayson, 2009).

Results from this study could also challenge or clarify support for attributional bias theory. Researchers in that field posit that those who expect success are more likely to exhibit attributional bias (Tetlock & Levi, 1982; Weary, 1979; Weiner, 2010). As shown, veterinary students have likely experienced academic success to be admitted to veterinary school due to the stringent admission requirements and competition, and therefore, they could be expected to display attributional bias.

Summary

Chapter 1 introduced the research problem, the conceptual framework for the research, and the purpose of the study. This chapter presented a review of historic literature and current research on SETs in higher education and health professional education.

Student evaluations of teaching in higher education have existed for nearly 100 years (Addison & Stowell, 2012). Originally, they were intended as formative assessments to be used by instructors to improve their teaching (Smalzried & Remmers, 1943). In the 1950s, the focus began to shift toward summative use of SETs (Addison & Stowell, 2012). By the 1970s, Marsh had begun to develop an instrument with nine constructs, and SETs were becoming commonplace in higher education (Algozzine et al., 2004; Marsh & Roche, 1997; Royal, 2017). Eighty-six percent of higher education institutions were reportedly using SETs by 1993 (Hoefer

et al., 2012). Online SETs, which became the norm in the 2000s, lowered cost and increased efficiency of the process (Adams, 2012).

Most SETs have been found to be valid and reliable instruments (Adams, 2012; Addison & Stowell, 2012; Hativa, 2013; McCarthy, 2012). However, non-response bias and student expectation of grade are legitimate concerns that could affect this reliability and validity (Aleamoni, 1999; Algozzine et al., 2004; Basow & Martin, 2012; Benton & Cashin, 2010; Cho & Cho, 2017; Christopher & Leskosky, 1997; Clayson, 2009; Hativa, 2013; Malone et al., 2018; Richardson, 2005). The few researchers who have examined the relationship between *actual* student grades and SET scores have found conflicting results (Arnold, 2009; Canaday et al., 1978; Cho & Cho, 2017; Hativa, 2013).

Despite widespread use of SETs, many faculty members perceive SETs as a way for students to vent frustrations and believe that students give the highest scores to the most popular instructors (Eckhaus & Davidovitch, 2019; Hammer et al., 2018; Nasser & Fresko, 2002). Nonetheless, 62.7% of instructors in a 2018 study felt that receiving SET feedback helps them to recognize areas for improvement in their teaching (Hammer et al., 2018, p. 525).

The next chapter, Chapter 3, describes the methodology, research design, and procedures for this study. Chapter 4 details the results from data analysis and provides a summary of the results, and Chapter 5 interprets and discusses the results and describes future opportunities for research.

CHAPTER THREE
METHODOLOGY

Student evaluations of teaching have generally been found to be valid and reliable measures of teaching effectiveness (Adams, 2012; Addison & Stowell, 2012; Hativa, 2013; McCarthy, 2012). Yet nonresponse bias persists as a concern; extending the SET completion window to after final exams was postulated to help increase veterinary student response rates (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019). Conversely, several researchers have shown an association between student grade and SET results (Clayson, 2009). The goal of this study was to determine how administering SETs to veterinary medical students after final exams (versus before final exams) affects results, with attributional bias theory as the conceptual framework.

Research Questions

With attributional bias informing and influencing this research, the following research questions guided this study:

RQ1. (*Quantitative*) How does veterinary student completion of SETs after final exams (versus before final exams) affect numerical scores on SET items?

RQ2. (*Quantitative*) How does administering SETs to veterinary students after final exams (versus before) affect completion rate and volume of comments?

RQ3. (*Qualitative*) How does veterinary student completion of SETs after final exams (versus before) affect the substance of comments?

RQ4. (*Mixed*) How does veterinary student completion of SETs after final exams (versus before) affect the intensity of comment themes and magnitude of sentiment?

Mixed Methods Approach

Plano Clark and Badiee (2015) indicated that one's research questions should dictate research design. Thus, the research design that best answered the research questions in this study was a convergent, parallel mixed methods approach, which incorporates quantitative and qualitative data collection simultaneously and then converges both data types during analysis (Creswell, 2014). The research design may be diagrammed as follows (Creswell, 2014; Morse, 2015):

QUAN + QUAL

The capitalization of QUAN (quantitative) and QUAL (qualitative) indicates that the research was driven equally by each of these methods (Morse, 2015). The + symbol represents the simultaneous collection of both types of data (Creswell, 2014; Morse, 2015). The data types met at two points, known as the analytic and the results points of interface (Morse, 2015). These two points are described more in the section on Data Analysis.

Creswell (2014) noted that a mixed methods approach provides more thorough insight into a research problem than either approach by itself. Quantitative research examines relationships between variables, and qualitative research explores meaning (Creswell, 2014; Plano Clark & Badiee, 2015). The use of mixed methods research in this study was informed by a pragmatic paradigm, a philosophical foundation that is concerned with solutions to problems (Biesta, 2015; Creswell, 2014; Plano Clark & Badiee, 2015). John Dewey, who is widely considered the father of the pragmatic worldview, held a transactional theory of knowledge based on the idea that "knowing is always about relationships between actions and consequences" (Biesta, 2015, p. 20, e-book).

In addition, mixed methods research provides an opportunity to triangulate sources of data toward convergence of findings (Creswell, 2014; Greene, Caracelli, & Graham, 1989). Integrating results from quantitative and qualitative methods allows for crosschecking the accuracy of multiple forms of data through statistical and textual analyses (Creswell, 2014; Greene et al., 1989). Results can be explained, confirmed, and enhanced to offset weaknesses and enhance the merits of both approaches (Plano Clark & Badiee, 2015). As Morse (2015) noted, mixed methods "allow the researcher to maintain the complexity of the phenomena within the research project" (p. 16, e-book).

Because of the bounded organization of this research, it might also be considered a case study (Creswell, Hanson, Plano Clark, & Morales, 2007). Creswell and colleagues (2007) considered the case study to be one of the most used qualitative designs and described its purpose as "developing an in-depth understanding about how different cases provide insight into an issue" (p. 239). Case studies typically contain multiple forms of data and are bounded by time and/or place (Creswell et al., 2007; Yin, 2013). Yin (2013) recognized that case studies frequently use mixed methods designs to triangulate data and bring evidence together. The study outlined here was bound by one location and a specific time period and used multiple forms of data, as described below.

This case study was also cross-sectional in that it occurred during a specific period of time rather than longitudinally (Creswell, et al., 2007; Levin, 2006). Cross-sectional studies are particularly useful to investigate associations between variables; however, they are also limited in that they contain data of only one moment in time (Levin, 2006). Nonetheless, cross-sectional studies do not undergo the participant dropout that occurs with longitudinal studies and work

particularly well when examining samples of students, who will eventually graduate and get lost to follow-up (Levin, 2006).

Prior Mixed Methods Use in SET Studies

McNulty and colleagues (2010) used a mixed methods design to study whether SETs were influenced by timing at a medical school. They examined SET Likert items and student comments to study the effect on student completion of SETs early in the semester versus later, with respect to knowledge of final grade (McNulty et al., 2010). They found that students who submitted SETs before the course ended provided more comments overall and more substantial comments, but no statistical differences were found in item scores (McNulty et al., 2010). Similarly, El Hassan (2009) explored student and faculty perspectives on SETs using surveys. The author statistically analyzed numerical survey item responses and conducted qualitative analysis using open-ended comments submitted with the surveys (El Hassan, 2009). Results from that study were that both students and faculty viewed SETs as useful, but students seemed to view them more positively than faculty (El Hassan, 2009). In a separate study by Onwuegbuzie et al. (2007), the researchers used students' open-ended comments to develop themes of student perceptions of the qualities of effective college instructors. They then quantified the themes and calculated a prevalence rate for each theme (Onwuegbuzie et al., 2007). Onwuegbuzie and colleagues (2007) found that the SET form used by their institution did not contain items that addressed some of the most prevalent themes that students identified as important characteristics of effective instructors.

Site and Sample

This study took place at a veterinary medical school at a large, public, Carnegie tier 1 research institution in the southeastern United States. The school contained a relatively small student body in comparison to other veterinary schools (Association of American Veterinary Medical Colleges, 2020b) and was fully accredited by the AVMA (2020b). For this study, data were collected during the 2016-2017 and 2017-2018 academic years and included fall and spring semesters.

The sampling method was purposive (Miles & Huberman, 1994). With purposive sampling, participants are selected who will enable generalization to the larger population of interest (Sharma, 2017). The population of interest in the current study was veterinary students. However, purposive sampling can introduce researcher bias to a study and possible non-representativeness of the sample compared to the population (Sharma, 2017). To help make the study more trustworthy, the researcher engaged in transparency with her own position and possible biases, as well as compared sample demographics (from student admissions data) to population demographic data (from the American Association of Veterinary Medical Colleges).

Participants

All veterinary students in their first 3 years of study at the research site had an opportunity to be included. This included 78 students in the class of 2018, 85 students in the class of 2019, 86 students in the class of 2020, and 91 students in the class of 2021, for a total of 340 students. Data from the class of 2018 (during the 2016-2017 academic year) were used only to establish a baseline response rate. Fourth-year students were excluded because they were not

enrolled in didactic courses and were instead completing 2- or 4-week clinical rotations that do not lend themselves to the same type of SETs nor the same type of analysis.

During the fall 2017 and spring 2018 semesters, for each course, a formula in Microsoft Excel (shown below) was used to randomly assign 262 students to receive their SET forms before or after final exams (Figure 3.1). Random assignment lessens bias because each individual has an equal chance of being selected; in this way, it allows for representativeness of the population (Sharma, 2017).

=CHOOSE(RANDBETWEEN(1,2),"Before","After")

Students then had the choice of whether to complete SETs at all, regardless of which group they were in (self-selection). Self-selection has the potential to introduce nonresponse bias (Bacon et al., 2016). Nonresponse bias creates a conundrum because it is part of the problem that this study was attempting to help solve. Ultimately, 2,926 SET forms were completed by 171 participants.

Response rates to SETs at this site had continually decreased from 2012 to 2017, during which years students had received a similar number of evaluations each semester (Figure 3.2). The content of the SET forms also remained the same during that time. In 2016 and 2017, about 40% of students in their first year completed SETs, whereas about 30% of students in their second and third years completed them (Bailey & Mawby, 2018). The assessment committee at the site had considered and experimented with multiple ways to improve response rates. One of these experiments was with SET timing. Similar to students at a German medical school, veterinary students at the study site had repeatedly requested to be given time to complete SETs after final exams (Schiekirka et al., 2012). When students made this request, they indicated that

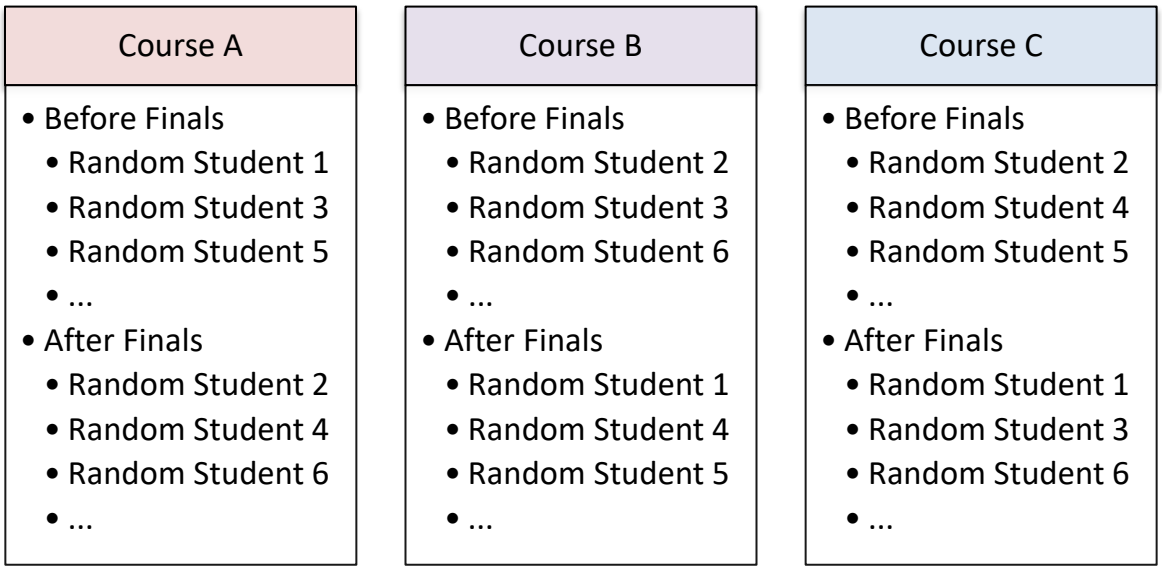

Figure 3.1. Representative random assignment of 262 veterinary students to receive their SET forms before or after final exams.

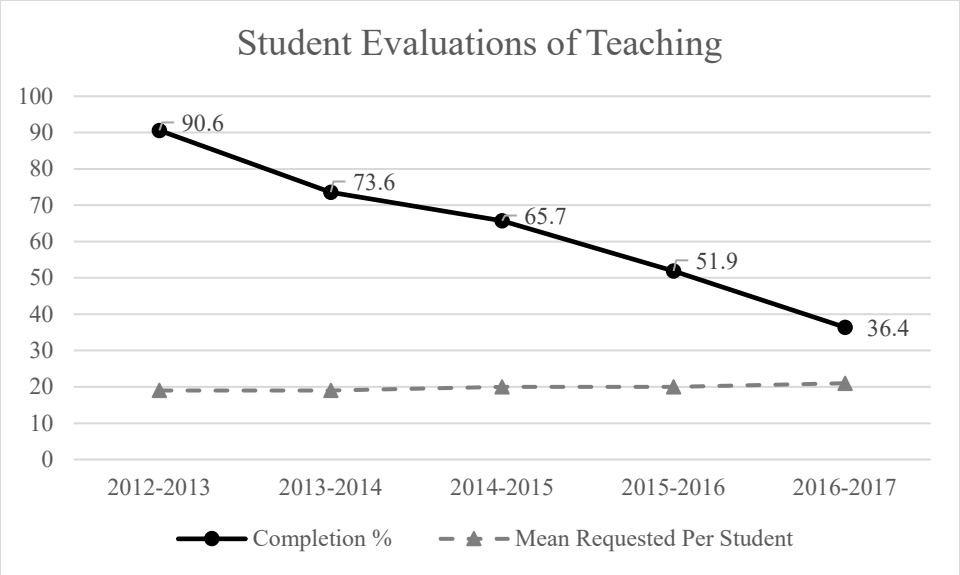

Figure 3.2. Summary of core course instructor evaluations completed and requested per student at the study site during academic years 2012-2013 to 2016-2017. Totals do not include course, general semester, or elective evaluations.

having time to complete SETs after finals would enable them to focus more on studying before finals and focus more on completing evaluations after finals. However, standard practice at most universities has been to close SETs before final exams (Arnold, 2009; Hativa, 2013; Young et al., 2019).

Instrument

The instrument used in this research is a form that has been used at the study site at least since spring semester 2011 and contains 10 items with a 5-point, Likert rating scale (Robinson & Leonard, 2019). As shown in Figure 3.3, the first nine items concern instructor facilitation of learning, objectives, availability, interaction, teaching aids, concept emphasis, evaluation, pace, and assignments. Rating scale choices are from strongly disagree (left placement) to strongly agree (right placement). The tenth item is about the instructor's overall teaching skills and contains rating scale choices from poor (left) to excellent (right). All items contain a "not applicable" option. The form also provides a comment box for students to give "relevant, constructive comments," and the qualitative data for this study resulted from student responses to this comment prompt.

Although this form has been used to provide feedback to instructors and in annual review, promotion, and tenure decisions for approximately 10 years, it has not undergone statistical analysis to determine its reliability and validity. However, the way in which SET data are used is an important aspect for ensuring a reliable and valid instrument, and many researchers have indicated that SETs, regardless of their origin, are valid and reliable (Hativa, 2013).

College of Veterinary Medicine	Evaluated By : evaluator's name Evaluating : person (role) or moment's name (if applicable) Dates : start date to end date

* indicates a mandatory response

Year 3 Instructor Evaluation Questions

	Not Applicable	1 Strongly Disagree	2	3	4	5 Strongly Agree
1. This instructor seems concerned with facilitating my learning.	○	○	○	○	○	○
2. Lectures/labs, etc. conducted by this instructor (regardless of course) were organized and had clear objectives.	○	○	○	○	○	○
3. This instructor was willing to discuss course material with me outside of class time.	○	○	○	○	○	○
4. This instructor encouraged interaction and answered student questions during class meetings.	○	○	○	○	○	○
5. Any teaching aids (handouts, visual materials, demonstrations, illustrations) used by this instructor were helpful and relevant to me in learning the subject matter.	○	○	○	○	○	○
6. Major concepts were emphasized during classroom presentations.	○	○	○	○	○	○
7. I was clearly informed by this instructor on how I would be evaluated (tested/graded).	○	○	○	○	○	○
8. The instructor covered material in this course at a pace that was reasonable for me.	○	○	○	○	○	○
9. Directions for course assignments by this instructor were clear and specific.	○	○	○	○	○	○

	Not Applicable	1 Poor	2	3	4	5 Excellent
10. My overall opinion of this individual's teaching skills is:	○	○	○	○	○	○

11. The grade I expect in this course is:

- ○ A
- ○ B+
- ○ B
- ○ C+
- ○ C
- ○ D
- ○ F
- ○ Satisfactory
- ○ No credit

Please provide any relevant, constructive comments regarding this instructor.

The following will be displayed on forms where feedback is enabled...
(for the evaluator to answer...)

*Did you have an opportunity to meet with this trainee to discuss their performance?

- ○ Yes
- ○ No

(for the evaluee to answer...)

*Did you have an opportunity to discuss your performance with your preceptor/supervisor?

- ○ Yes
- ○ No

Figure 3.3. Form veterinary students at the study site use to evaluate instructors in pre-clinical courses.

Data Collection

The data for this study already existed; therefore, this section will describe the prior collection procedures. Quantitative and qualitative data collection occurred simultaneously, as is expected with concurrent parallel designs (Creswell, 2014). The one45 software system, which is used by the site to administer SETs, was used to gather raw SET data.

The window of student SET completion was precisely 2 weeks for each group. The one45 software system was used to set up release and closing dates for online evaluations for each group in each course. One month before the first evaluations were released, the associate dean for academic affairs sent an e-mail to all students using pre-established college e-mail groups. This email told the students that based on their feedback, the academic affairs office was trying a new process for SETs that would cause different students to receive evaluations at different times from past years. The email further explained that the college must evaluate the reliability and biases associated with the proposed changes. The process was explained to avoid confusion among students who would not receive evaluations before final exams and might think they were mistakenly left off the distribution list. An explanation of why this new process was being implemented was also provided to let students know that their concerns and suggestions were being heard and addressed. None of the emails were returned due to erroneous addresses.

Using the same e-mail groups, the curriculum and assessment coordinator (the researcher) sent an e-mail reminder 1 week before the first evaluations were released and sent bulk e-mail reminders through the one45 system when 1 week and 1 day remained before the deadline. The same process was followed for the post-final exam SETs. Reminder flyers were placed in each classroom and in other shared spaces 1 day before the first SETs were released.

Raw quantitative data were exported to Microsoft Excel and loaded into SPSS version 26 for Windows statistical software. Similarly, students' open-ended comments were imported into NVivo 12 for Windows.

Data Analysis

The primary goal of this mixed methods data analysis was to identify patterns and differences (Plano Clark & Badiee, 2015). Data analysis for mixed methods studies include a quantitative and qualitative component (Creswell, 2014; Plano Clark & Badiee, 2015). For this study, data analysis might be best and most concisely explained in Figure 3.4. As described more completely below, the quantitative analyses involved SET item numerical scores, completion rate, number of students providing comments, and comment length. The qualitative analysis involved coding the comments for categories, themes, and sentiment. Thereafter, the codes were transformed into counts, allowing for additional quantitative analyses of theme intensity and sentiment. Within Figure 3.4, the location of the data transformation oval is where convergence of quantitative and qualitative findings occurred at two points of interface (Morse, 2015). Morse (2015) described the results point of interface as the transformed qualitative data being written into the quantitative results, which form the framework of the overall results. More specifically, transformation of qualitative data occurred by converting qualitative data into quantitative data via frequencies of occurrences of categories, themes, and sentiment. Codebooks were maintained to describe the quantitative and qualitative data in order to manage data, provide transparency, and make reproduction of the results possible (Saldaña, 2013; Tabachnick & Fidell, 2013). These codebooks contained variable names and numerical codes, statistical tests, and descriptions of emergent qualitative codes, categories, and themes.

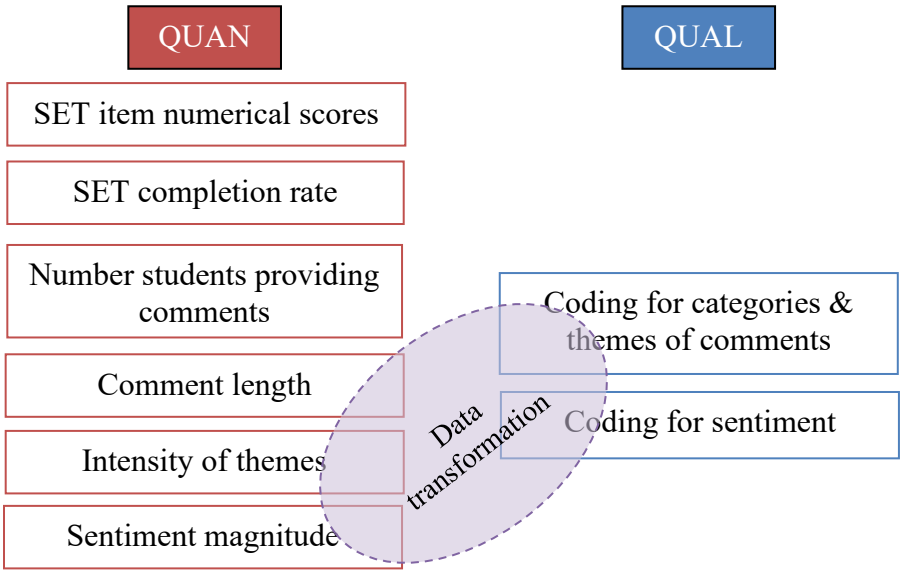

Figure 3.4. Research design showing data analysis.

Quantitative Data Analysis

Both descriptive and inferential statistics were used for the quantitative portions of the study. Descriptive statistics describe measures of central tendency, variance, and distribution, while inferential analysis enables sample-based findings to be generalized to a population (Onwuegbuzie & Combs, 2015). Table 3.1 contains a list of all quantitative variables and analysis types.

Data were first cleaned and examined to ensure all assumptions for each analysis had been met. For this process, the 12-step data cleaning method described by Morrow and Skolits (2017) was used. The dataset was examined for missing data, and a determination was made to perform casewise deletion for all analyses (Morrow & Skolits, 2017; Tabachnick & Fidell, 2013). Thereafter, the data were examined to determine if they met other statistical assumptions for planned statistical tests (Morrow & Skolits, 2017; Tabachnick & Fidell, 2013). When assumptions were not met, a determination was made of how much assumptions were violated and if the statistical tests remained robust despite violated assumptions or if non-parametric tests needed to be performed instead (Tabachnick & Fidell, 2013). The only non-parametric test used was the χ^2 because this test fit the data types and research questions being answered.

Because the data were non-independent but anonymous, prohibiting the pairing of datapoints before and after final exams, data were aggregated for some tests, where noted. Such aggregation transformed the raw data so that cases were no longer at the student level but were instead at the instructor level. Data that are truly independent have no duplication of participants between independent variable (IV) levels (van den Berg, 2020). Due to the anonymity of the data in this study, data from the same students and for the same instructors appeared in both groups of

Table 3.1
Quantitative Research Variables

Dependent variable	Independent variable(s)	Analysis method(s)
Item numerical scores*	Timing, student cohort, expected grade	Linear mixed models with an AR(1) covariance structure (timing, student cohort); Pearson *r* correlation (expected grade)
Completion rate	Timing, student cohort, semester, academic year	Simultaneous logistic regression (timing, semester, academic year); simultaneous multinomial logistic regression (student cohort)
Number students providing comments	Timing	χ^2
Comment length*	Timing, expected grade	Independent *t* test (timing); Pearson *r* correlation (expected grade)
Intensity of themes in comments	Timing	χ^2
Directional magnitude of comments	Timing	χ^2

Note. * Data aggregated at the instructor level.

the IV of timing, making the data non-independent. Additionally, the same instructors taught multiple courses, and aggregating the data this way helped control for instructor effect. Aggregated data is less noisy, meaning it is cleaner and enables more robust modeling (Garcia, Luengo, & Herrera, 2015; UCLA Statistical Consulting Group, n.d.). Hoefer et al. (2012) and McCullough and Radson (2011) used similar aggregation to perform SET item analysis. Independence of the data and data aggregation are addressed more fully later in this chapter.

Qualitative data from student comments were transformed into quantitative frequencies of comments provided, number of words in each comment, magnitude of praise/criticism, and intensity of discovered codes, categories, and themes (Plano Clark & Badiee, 2015).

SET Item Numerical Scores

For these analyses, the IVs were the timing of the SET (two levels–before and after finals) and student cohort year (three levels–years 1, 2, and 3) (Creswell, 2014). The dependent variables (DVs) were each of the 10 items within the SET instrument. Before running the analysis, data were aggregated in SPSS by student cohort, semester, year, instructor, course, and timing, creating new mean variables for each SET item. Therefore, the sample size for data analysis purposes was reduced from 2,926 to 238. After data aggregation for course A, there were four rows of data: one for each of two instructors both before and after final exams. For course B, there were eight rows of data: one for each of four instructors both before and after final exams.

Mixed effects models analyses were then used to determine if mean item scores were statistically different before versus after final exams. Mixed models analyses are useful with data that are non-independent and correlated (UCLA Statistical Consulting Group, n.d.). The specific

type of mixed models used in this study was linear mixed models using an autoregressive(1) (AR[1]) covariance structure (Kincaid, n.d.). A covariance structure refers to how mathematical matrices are constructed in mixed models. The AR(1) covariance structure relates to within-subject correlation and will hold constant the variation in the random effects in the models (Kincaid, n.d.). This structure allows for higher statistical power while enabling better control for Type I errors (JMP, 2020). Type I errors are those that occur when the null hypothesis is rejected although it is true (Lee & Lee, 2018). Fixed factors in the model (variables of interest) were cohort year, timing, and cohort year × timing interactions. Random effects were semester, year, instructor, and course, which allowed the test to control for the variance in those four variables and determine if statistical differences in SET item results are due to timing and not another random effect (Kincaid, n.d.). Specifically, including instructor and course as random factors permitted controlling for variation in instructor leniency and in course difficulty. Timing (before or after finals) was set as the repeated measure.

To make post-hoc comparisons for student cohort to determine if any cohorts differed statistically, least significant difference (LSD) pairwise multiple comparison tests were used; this test is similar to multiple *t* tests between all pairs of student cohorts (IBM Knowledge Center, n.d.). Calculation of an effect size, otherwise known as the practical value of the results, is not possible with mixed model analyses (Tabachnick & Fidell, 2013).

To examine SET items with respect to students' expected grade, data were aggregated by student cohort, semester, year, instructor, course, timing, and expected grade, creating new mean variables for the 10 SET items. Expected grade was included as a break variable, rather than as a mean variable, to maintain the seven different grade levels. Maintaining the seven different grade

levels rather than creating a mean variable enabled comparison between the seven grade types (A, B+, B, C+, C, D, and F). The sample size for data analysis purposes was reduced from 2,926 to 897. Because maintaining the seven grade levels created identical levels for the repeated variable of timing, mixed models analyses were not possible for grade as an IV. Therefore, Pearson correlations were used to determine possible relationships between students' expected grades, timing of delivery (IVs), and SET items (DVs).

SET Completion

Both descriptive and inferential statistics were examined for non-aggregated, SET completion data from the 2016-2017 and 2017-2018 academic years (Onwuegbuzie & Combs, 2015). This approach enabled examination of baseline completion during 2016-2017 and a comparison with 2017-2018 completion. Additionally, completion data for 2017-2018 before and after finals were examined. A simultaneous, binary logistic regression test was used to determine if timing, semester, and academic year (IVs) explained or predicted SET completion (DV). The 2016-2017 (control) and 2017-2018 (treatment) academic years were used to assess the probability of SET completion when students are given time after finals to complete them. Simultaneous, multinominal logistic regression was used to examine if student cohort (IV) predicted SET completion (DV).

Logistic regression is conceptually related to multiple regression; however, it uses the χ^2 test rather than the *F* test that is used in multiple regression (Keith, 2015). Results are produced as probabilities and odds (Huang & Moon, 2013). Specifically, logistic regression is commonly used when a DV is dichotomous, such as with SET completion–the SET was completed or not completed (Huang & Moon, 2013; Keith, 2015). In a simultaneous logistic regression, multiple

categorical or continuous IVs are incorporated into the model simultaneously rather than sequentially (Keith, 2015). A simultaneous approach was appropriate in this instance because there were no hypotheses about the order or importance of the IVs (Tabachnick & Fidell, 2013). Several post-hoc assumptions were also addressed; these included linearity, absence of multicollinearity, absence of outliers in the solution, and independence of errors (Tabachnick & Fidell, 2013). Outliers were indicated for the cohort multinomial regression and the outlying cases deleted, reducing the sample size from 10,018 to 9,923. The Durbin-Watson statistic indicated non-independence of errors for all IVs, which was expected, given the overlap of participants in both timing groups. This limitation is discussed fully later in this chapter.

The analyses provided the probability of students completing or not completing SETs, given the categories for timing, student cohort, semester, and academic year (Keith, 2015). This type of analysis has been used previously to analyze participation in surveys (Dillman et al., 2009; Haunberger, 2011; Porter & Whitcomb, 2005; Reisenwitz, 2016). More specifically, the null hypothesis for this logistic regression indicates that no relationship exists between SET completion and cohort year, semester, or timing.

Number of Students Providing Comments

The number of students (DV) providing comments on SETs before and after finals (IV) was compared using a nonparametric χ^2 test of independence with a 2×2 contingency table, with two rows containing timing (before and after finals) and two columns containing comment (included and not included). The χ^2 test determined if the two variables were statistically related based on their distribution (Sharpe, 2015). This statistical test indicates if students are more likely to provide comments either before or after finals. Of note is that this analysis was

conducted with non-aggregated data because aggregation prohibited the use of the χ^2 test by creating cells with expected frequencies less than 5 (McHugh, 2013).

Comment Length

The Student *t* test was used determine if the amount of information participants provided in comments, as determined by word counts (DV), differed statistically before and after final exams (IV). Before running the analysis, data were aggregated in SPSS by student cohort, semester, year, instructor, course, and timing, creating new mean variables. Therefore, the sample size for data analysis purposes was reduced from 13,560 to 238. This technique created means for word counts and grades for each row of data.

Because aggregated grade expectancy and word count data were continuous, a Pearson correlation test was conducted to examine possible correlations between the grade students expected (IV) and the number of words (DV) they provided in open-ended comments. When addressing the assumptions for the Pearson test, six outliers for word count were discovered and deleted from the dataset before moving forward with the analysis.

Sentiment Magnitude

After the data were coded for sentiment magnitude (criticism, praise, neutrality, and combination, as described below) and transformed to frequencies, a χ^2 test was used to determine if directional magnitude of criticism/praise (DV) differed statistically before versus after exams (IV).

Intensity of Themes

After the data were coded and themes developed, the intensity of the themes were examined in relation to the timing of data collection, as described below (Bergman, 2015;

Saldaña, 2013). Intensity refers to how often each theme appears in each dataset. Such an approach works well for mixed methods studies to supplement other findings toward richer answers to research questions (Saldaña, 2013). Therefore, after the data were coded for themes and transformed to frequencies, a χ^2 test was used to determine if frequency of concepts and themes (DV) differed statistically before or after final exams (IV).

Methodological Limitation: Anonymity and Duplication of Participants

Due to the anonymity of the data in this study and to the design of random assignment to groups, data from the same students and for the same instructors appeared in both groups of the independent variable of timing, making the data non-independent. Although it was possible to determine which students completed the SET forms, it was not possible to link their completion to their responses. This prohibited the use of statistical tests that use dependent data types. Overlap between before and after groups was calculated, and it was found that 24.1% of the 147 students in the before group were unique participants, whereas 13.5% of the 127 students in the after group were unique. In other words, 75.9% of students in the before group also appeared in the after group, and 86.5% of students in the after group also appeared in the before group. Data that are truly independent have no duplication of participants between IV levels (van den Berg, 2020).

The data used in this study violate the assumption of independence, thereby introducing bias to the results and a loss of robustness of statistical tests (Kenney & Judd, 1986). Nonindependence of data causes an inflation in the Type I error rate; it increases the probability of incorrectly rejecting a true null hypothesis (Kenney & Judd, 1986; Tabachnick & Fidell, 2013). To reduce the risk of committing a Type I error, the α level can be lowered to .01,

indicating the researcher is willing to accept a 1% chance of committing a Type I error (Tabachnick & Fidell, 2013). Therefore, in this study, α was lowered to .01. Because the main concern with Type I errors is seeing a statistical difference when one does not exist, nonindependence of data is not as worrisome with nonsignificant results. In other words, one can be more confident of the certainty of nonsignificant results.

Another way to reduce Type I error is to aggregate data. In this study, data were aggregated before analysis for comment length and item numerical scores so that they were no longer at the student level and were instead at the instructor level. Because the same instructors taught multiple courses, aggregating the data also helped control for instructor effect. Incorporating the grouping factor of timing into the mixed models analysis was also a solution to nonindependence due to groups (Kenney & Judd, 1986).

Qualitative Data Analysis

A constant, comparative method was employed to analyze textual comments provided by students; this analysis was achieved by using two simultaneous coding methods: eclectic and magnitude (Anfara, Brown, & Mangione, 2002; Bergman, 2015; Onwuegbuzie & Combs, 2015; Saldaña, 2013). When coding, researchers identify words or phrases that capture the essence of portions of text, whether that text be a word, phrase, or paragraph (Saldaña, 2013). The constant, comparative design permits the identification of patterns and categorization of findings (Anfara et al., 2002). In particular, eclectic coding, a type of open coding, is exploratory and descriptive and also allows for examination of magnitude (Saldaña, 2013). Because eclectic coding works in iterative cycles, after the first round of coding, categories were developed that encompassed

initial codes (Saldaña, 2013). After reflection, these categories were grouped and reduced into broader themes (Saldaña, 2013).

All manual coding, category, and theme development was completed with the assistance of NVivo 12 for Windows. Importantly, this use of a computer-assisted qualitative data analysis software does not automatically make the results more scientific; instead, NVivo helped to index and organize the data (Bergman, 2015). To help circumvent researcher bias, this coding was partially blinded (masked) by assigning a code to each instructor so that the instructors were not known during data analysis. However, given that the data did contain references to specific instructor names, completely blinded coding was not achievable. Additionally, timing of each comment was blinded by ensuring that timing was not visible during coding. Upon completion of qualitative data analysis, findings were compared to quantitative results to identify convergent and/or divergent results (Onwuegbuzie & Combs, 2015). In particular, of interest were how the data correlated in relation to the timing of data collection (Bergman, 2015; Onwuegbuzie & Combs, 2015).

In addition to theme development, magnitude coding was also used to tag each complete comment as praise, criticism, neutrality, or mixed (Saldaña, 2013). Magnitude coding enabled additional information about the data to indicate evaluative content and worked well to establish the "direction" student comments (Saldaña, 2013, p. 73). Matrices were then used to examine and display resulting data (Saldaña, 2013).

Although qualitative veterinary education research designs are in the minority, Schiekirka and colleagues (2012) used content analysis of focus group transcripts in a study of German medical student perceptions of SETs. They noted that their method was adjunctive to statistical

methods and could not generalize to larger populations of medical students (Schiekirka et al., 2012). Similarly, Maghães-Sant'Ana (2014) conducted a qualitative case study using syllabus content analysis and instructor interviews to examine ethics curricula in three European veterinary schools.

A concern with qualitative research is maintaining the rigor of the study (Anfara et al., 2002). Critics of coding argue that it lacks researcher objectivity (Saldaña, 2013). For qualitative design, the quantitative terms of objectivity, reliability, and internal and external validity are better represented with the terms confirmability, dependability, credibility, and transferability, respectively (Anfara et al., 2002). Strategies employed to maintain quality and rigor within this study included triangulation and purposive sampling, as well as researcher reflexivity and transparency in methods via an audit trail (Anfara et al., 2002).

Ethical Safeguards and Considerations

Creswell (2014) warned against "backyard research," in which the researcher has a close connection to the population and site being studied (p. 188). Although the data are easy to collect, it is also easy for a researcher to provide inaccurate information (Creswell, 2014). To address this shortcoming, Creswell (2014) points to the researcher as having a responsibility to ensure that multiple approaches for validation are used. Additionally, researchers must be particularly careful not to put participants at risk and to maintain the privacy of the data (Creswell, 2014). This study required approval from the University of Tennessee Institutional Review Board (IRB). The IRB process helps ensure that a study plans to protect human subjects and their data as much as possible. This study was approved by the IRB in exempt category 4 in

May 2018 but was amended with adjusted methods for expedited review and approved in September 2020 (UTK IRB-18-04470-XP).

Position Statement

A concern with qualitative coding is researcher subjectivity (Bergman, 2015). Ideally, two coders would compare their results to try to establish convergence, which increases the trustworthiness of the data (Bergman, 2015). Saldaña (2013) recommended discussing the coding process with colleagues and mentors and maintaining a reflective journal to lessen potential bias. Both these recommendations were carried out. Reflexivity helps the researcher and readers understand how researcher biases, background, and values have the potential to affect the interpretation of the results (Creswell, 2014). Because of the researcher's employment at the study site, she was an insider in the setting in which this research took place. Such positionality may be seen as both positive and negative. That the researcher has been employed at the site for 15 years confounds these positives and negatives. Her perspective brings with it inherent interest but also biases. She did not need to use gatekeepers to gain access to the research site or the participants. Although this facilitated ease in conducting this study, it could be perceived as taking advantage of participants. However, because these data are collected regularly apart from the purpose of this research, students were already being asked to provide these data.

Additionally, the researcher prepared the online SETs and disseminated the results. She often heard and read statements of disappointment from faculty when their SET response rates were low. Therefore, she wanted to see these rates increase and to understand how to better perform her job. Her own success was partially at stake, and this had the potential to affect the

way in which she approached the study, as well as how she interpreted the data. She remained cognizant of these biases during data analysis and strived to control her biases. In particular, she took care not to form any *a priori* suppositions with respect to student comments and to bracket preconceptions to diminish bias (Onwuegbuzie et al., 2007). Using triangulation of data and methods also helped offset biases and enhance validity via convergence of results (Greene et al., 1989).

Summary

Chapter 1 introduced the research problem, the conceptual framework for the research, and the purpose of the study. Chapter 2 presented a review of historic literature and current research on SETs in higher education and health professional education. This chapter described the methodology, research design, and study procedures.

The research design was a convergent, parallel mixed methods approach, that incorporated quantitative and qualitative data collection simultaneously and then converged data types during analysis (Creswell, 2014). Mixed methods research allows triangulation of data toward convergence of findings (Creswell, 2014; Greene et al., 1989). Because this study was bounded by time and research site, it might also be considered a case study (Creswell et al., 2007). The research design may be diagrammed as follows, showing quantitative and qualitative methods equally driving the study (Creswell, 2014; Morse, 2015):

QUAN + QUAL

Data were collected during the 2016-2017 and 2017-2018 academic years using an initially purposive sampling method (veterinary students) and then random assignment (before or after finals) and participant self-selection (SET completion) (Miles & Huberman, 1994).

Veterinary students in their first 3 years of study had the opportunity to be included. The instrument used was an SET form used by the study site through the one45 software system, which produced raw SET data.

Descriptive and inferential statistics were used for the quantitative portions of the study (Onwuegbuzie & Combs, 2015). Coding was conducted to identify themes within the qualitative data. Variables studied include item numerical scores, SET completion rate, number of students providing comments, comment length, directional magnitude of comments, comment themes, and intensity of comment themes.

The last two chapters (Chapters 4 and 5) detail results from data analysis; provide a summary, interpretation, and discussion of the results; and describe future opportunities for research.

CHAPTER FOUR
RESULTS

In pre-clinical veterinary medical courses, students have many more instructors to evaluate than typical undergraduate and graduate students due to the team-teaching nature of veterinary courses (Malone et al., 2018). A decrease in response rate has been shown to occur when 11 or more student evaluations of teaching (SETs) are administered (Adams & Umbach, 2010); therefore, the team-teaching nature of veterinary courses often results in lower response rates. In addition, the reliability of resulting SET data is affected by nonresponse (Benton & Cashin, 2010; Hativa, 2013; Malone et al., 2018; Richardson, 2005).

Prior research suggested that if students had an opportunity to complete SETs after final exams, it might help increase response rates and lower nonresponse bias (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019). However, it is still standard practice to close SET collection systems before final exams (Arnold, 2009; Hativa, 2013; Young et al., 2019). The timing of evaluations and student knowledge of final grade have been shown to influence SET scores in undergraduate and graduate student populations (Arnold, 2009; Cho & Cho, 2017; Hoefer et al., 2012). A possible rationale for this phenomenon is attributional bias, which asserts that people attribute their successes to themselves and their failures to someone else (Arnold, 2009; Korn et al., 2016; Robinson, 2017).

The purpose of this study was to determine how issuing SETs to veterinary students after final exams (versus before finals) would affect item scores, completion rate, and comments. Knowing how timing and student knowledge of grades affect SET results could help inform veterinary college personnel on best practices for timing of release and closure of evaluations to students to help reduce nonresponse bias. Additionally, if attributional bias seemed to be

occurring, college administrators could examine ways to better train students on how to give constructive feedback (LaBelle & Martin, 2014).

With attributional bias informing and influencing this research, the following research questions guided this mixed methods study:

RQ1. (*Quantitative*) How does veterinary student completion of SETs after final exams (versus before final exams) affect numerical scores on SET items?

RQ2. (*Quantitative*) How does administering SETs to veterinary students after final exams (versus before) affect completion rate and volume of comments?

RQ3. (*Qualitative*) How does veterinary student completion of SETs after final exams (versus before) affect the qualitative substance of comments?

RQ4. (*Mixed*) How does veterinary student completion of SETs after final exams (versus before) affect the intensity of comment themes and magnitude of sentiment?

Study Participants

The initial sampling method for this study was purposive to the population of interest–doctor of veterinary medicine students in U.S. veterinary schools (Miles & Huberman, 1994; Sharma, 2017). Then, for each course, 262 students were randomly assigned to receive their SET forms before or after final exams to better enable generalization of the results to the population of interest. Students then had the choice of whether to complete SETs at all, regardless of which group they were in (self-selection). Ultimately, 171 students completed SETs during the study period.

The mean age of the student population was 25.68 years (SD = 3.33), and 82.13% of participants were female, while 17.87% were male. In comparison, all other students (in cohort

years 1, 2, and 3) enrolled in U.S. colleges of veterinary medicine during the same period had a mean age of 24.08 (*SD* = 0.72); 81.1% self-identified as female, 18.9% as male, and < 1.0% as non-binary (K. Young, personal communication, January 19, 2021). Table 4.1 shows comparative age and gender data for participants and the national population.

Analysis by Research Question

*RQ1 (*Quantitative*). How Does Veterinary Student Completion of SETs After Final Exams (Versus Before) Affect Numerical Scores on SET Items?*

The effect of timing of completion of SETs on the numerical scores on SET items was examined statistically with linear mixed models using an AR(1) covariance structure (Kincaid, n.d.). Student-level data were aggregated to the instructor level prior to analysis using IBM SPSS version 26. Data met the assumptions required for the analysis, except for normality. Kurtosis values were above the acceptable level of |2.0|, ranging from 2.06 to 4.73. These values indicated non-normality of the data; however, when the sample size is over 100, such deviations from normality do not diminish the robustness of the analyses (Tabachnick & Fidell, 2013). Therefore, the analyses were continued.

Results from the mixed models indicate that timing of SET completion had no significant main effect on any of the SET items (Table 4.2). However, there were significant ($p < .05$) main effects of cohort year on items 2, 3, 8, 9, and 10, with pairwise comparisons shown here:

2) Lectures/labs, etc. organized with clear objectives, $t(169.04) = 2.48$, $p = .014$

3) Instructor willingness to discuss course material outside of class time, $t(201.26) = 2.66$, $p = .008$

8) Instructor pace of covering the material, $t(168.55) = 2.90$, $p = .004$

Table 4.1

Comparative Demographic Data for Study Site and National Student Population

Cohort year	Site		National	
	Age (*M*)			
1	24.40		23.34	
2	25.81		24.12	
3	26.91		24.78	
Total	25.68		24.08	
	Gender (*N*)			
	Female	Male	Female	Male
1	80	11	2,711	647
	(87.91%)	(12.09%)	(80.73%)	(19.27%)
2	68	18	2,681	604
	(79.07%)	(20.93%)	(81.61%)	(18.39%)
3	69	16	2,580	612
	(81.18%)	(18.82%)	(80.83%)	(19.17%)
Total	217	45	7,972	1,863
	(82.82%)	(17.18%)	(81.06%)	(18.94%)

Table 4.2

Results of the Linear Mixed Model to Test the Main Effects of Student Cohort and Timing (Before/After Final Exams) on SET Items

Independent Variable	Predictor	*df*	*F*	*p*
1. This instructor seems concerned with facilitating my learning	Student cohort	2, 117	3.09	.05
	Timing	1, 114	.06	.81
	Timing × Student cohort	2, 114	1.90	.16
2. Lectures/labs, etc. conducted by this instructor were organized and had clear objectives	Student cohort	2, 119	3.53	.03*
	Timing	1, 114	.26	.61
	Timing × Student cohort	2, 114	.77	.47
3. Instructor was willing to discuss course material with me outside of class time	Student cohort	2, 120	4.60	.01*
	Timing	1, 116	.32	.57
	Timing × Student cohort	2, 117	.14	.87
4. Instructor encouraged interaction and answered student questions during class meetings	Student cohort	2, 114	2.14	.12
	Timing	1, 110	.04	.85
	Timing × Student cohort	2, 110	1.29	.28
5. Teaching aids used were helpful and relevant to me in learning the subject matter	Student cohort	2, 114	1.09	.34
	Timing	1, 109	.08	.78
	Timing × Student cohort	2, 110	.72	.49
6. Major concepts were emphasized during classroom presentations	Student cohort	2, 116	2.43	.09
	Timing	1, 111	.13	.72
	Timing × Student cohort	2. 112	.17	.85
7. I was clearly informed on how I would be evaluated (tested/graded)	Student cohort	2, 113	1.71	.19
	Timing	1, 110	.11	.74
	Timing × Student cohort	2, 110	.71	.49
8. Instructor covered material at a pace that was reasonable for me	Student cohort	2, 117	4.24	.02*
	Timing	1, 111	3.75	.06
	Timing × Student cohort	2, 112	.73	.48
9. Directions for course assignments by this instructor were clear and specific	Student cohort	2, 118	4.37	.02*
	Timing	1, 114	.07	.79
	Timing × Student cohort	2, 114	.64	.53
10. Overall opinion of individual's teaching skills	Student cohort	2, 117	3.35	.04*
	Timing	1, 112	.35	.56
	Timing × Student cohort	2, 113	.67	.51

Note. Form n = 2,926. Aggregated n = 238. Participant n = 171.

*$p < .05$

9) Clarity and specificity of directions for course assignments, $t(191.93) = 2.61$, $p = .010$

10) Overall opinion of instructor teaching skills, $t(172.49) = 1.97$, $p = .050$.

For these items, students in their third year (form $n = 885$; participant $n = 46$) rated instructors significantly higher than did students in their first year (form $n = 1218$; participant $n = 78$). These differences are shown in Table 4.3. No significant differences were found with second-year students.

To examine the relationship between students' expected grades, timing, and SET items, Pearson correlations were conducted. These results are shown in Table 4.4. Both before final exams and overall, there were significant, but small, positive correlations between grades and all 10 SET items. All but the following SET item had significant positive correlations after final exams: (3) Instructor was willing to discuss course material with me outside of class time. These results indicate that as students' expected grades increased, so did most SET item scores (Figure 4.1). However, the significant correlation r values ranged between .17 and .38. A general rule of thumb for interpreting r is that any correlation < .50 is considered a small correlation with little practical relevance (Hinkle et al., 2003).

RQ2* (Quantitative). *How Does Administering SETs to Veterinary Students After Final Exams (Versus Before) Affect Completion Rate and Volume of Comments?

To determine if issuing SETs to veterinary students after final exams affected completion rate and volume of comments, several statistical analyses were conducted. This section includes a summary of all findings related to SET completion rate, the number of students providing open-ended comments in SETs, and the length of those comments. Timing (before versus after finals) was an independent variable (IV) for all analyses. For completion rate analyses, student

Table 4.3

Statistically Significant Results of Linear Mixed Models to Test the Effects of Individual Student Cohort on SET Items

SET Item	*M* (*SE*)		
	1st year *n* = 78	2nd year *n* = 48	3rd year *n* = 46
2. Lectures/labs, etc. conducted by this instructor were organized and had clear objectives	4.39 (.07)*	4.53 (.05)	4.63 (.05)*
3. Instructor was willing to discuss course material with me outside of class time	4.62 (.17)**	4.73 (.17)	4.80 (.17)**
8. Instructor covered material at a pace that was reasonable for me	4.41 (.07)**	4.55 (.05)	4.65 (.05)**
9. Directions for course assignments by this instructor were clear and specific	4.52 (.18)*	4.63 (.18)	4.72 (.18)*
10. Overall opinion of individual's teaching skills	4.45 (.07)*	4.56 (.05)	4.66 (.05)*

Note. Asterisks indicate significant differences within rows. Rating scale choices were from (1) strongly disagree to (5) strongly agree for items 2, 3, 8, and 9. Scale choices were from (1) poor to (5) excellent for item 10. *SE* = standard error.

Form *n* = 2,926. Aggregated *n* = 238.

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

Table 4.4

Pearson Correlations between SET Items, Expected Grade, and Timing

SET Item	Timing		
	Before	After	Overall
1. Instructor seems concerned with facilitating learning	.33***	.17**	.26***
2. Lectures/labs, etc. conducted by this instructor were organized and had clear objectives	.32***	.20***	.26***
3. Instructor was willing to discuss course material with me outside of class time	.27***	.14	.21***
4. Instructor encouraged interaction and answered student questions during class	.31***	.24***	.28***
5. Teaching aids used by this instructor were helpful and relevant to me in learning the subject matter	.35***	.25***	.30***
6. Major concepts were emphasized during classroom presentations	.36***	.22***	.29***
7. I was clearly informed by instructor on how I would be evaluated	.33***	.30***	.32***
8. Instructor covered material at a pace that was reasonable for me	.35***	.34***	.35***
9. Directions for course assignments by this instructor were clear and specific	.38***	.25***	.32***
10. Overall opinion of individual's teaching skills	.33***	.24***	.29***

Note. Asterisks indicate significant, but small, correlations.
Form n = 2,926. Aggregated n = 897. Participant n = 171.
**$p < .01$. ***$p < .001$.

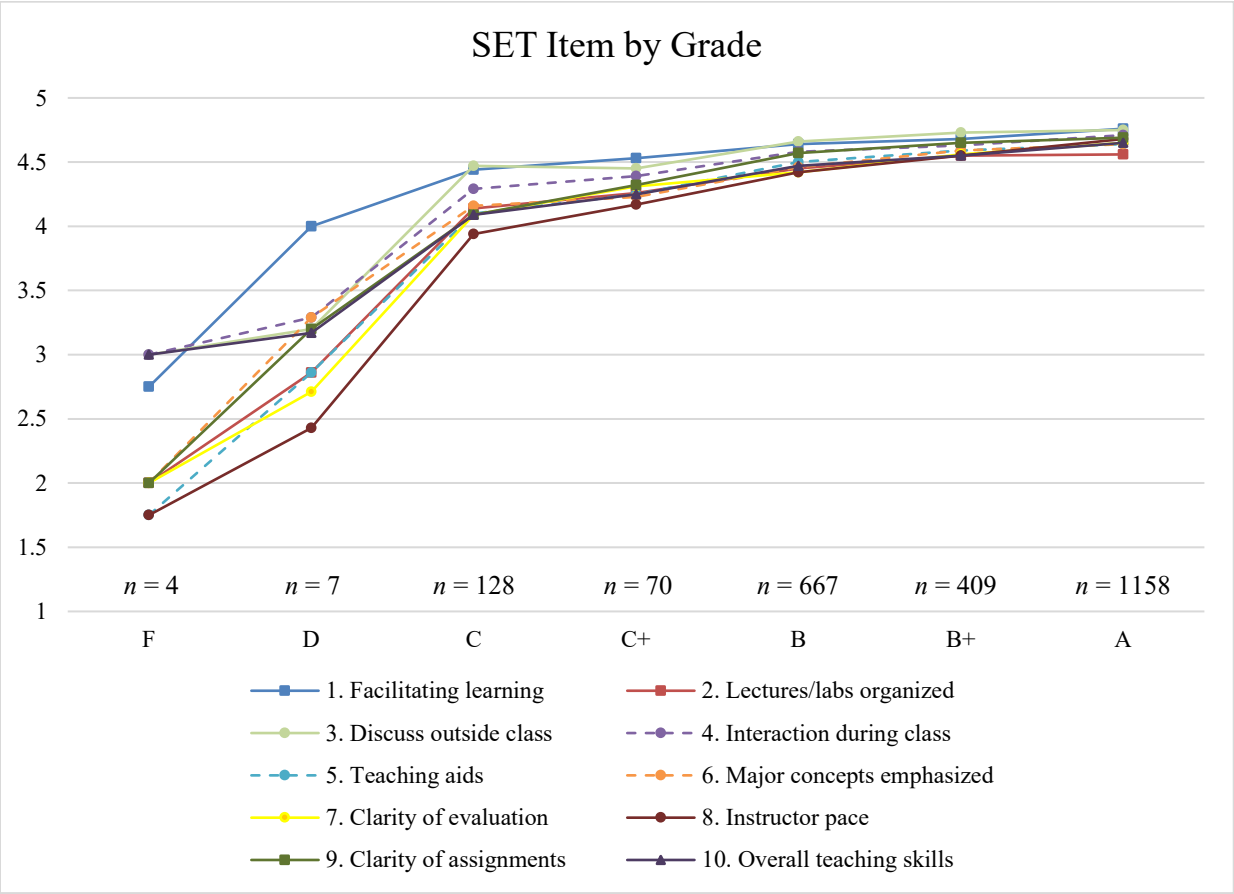

Figure 4.1. Mean SET item score (1–5) by student-reported, expected grade.

cohort, semester, and academic year were also examined, and for comment length, students' expected grade was considered as an additional IV.

Completion rate. The overall completion rate for SETs during the 2017-2018 academic year was 28.4% (compared to a completion rate of 32.60% during 2016-2017). Table 4.5 shows completion data grouped by student cohort and semester. To determine the probability of student completion of SETs, given the categories of timing and semester, a simultaneous, binary logistic regression was conducted. The model indicated that 92.8% of the not completed evaluations were correctly classified, and 31.8% of the completed evaluations were correctly classified. Overall correct classification was 75.5%. Model parameters are shown in Table 4.6.

Each predictor was significant on the outcome variable of SET completion, $\chi^2(6, n = 10{,}018) = 1205.93$, $p < .001$, Cox & Snell $R^2 = .11$. The Type I error rate was 7.2% (non-completers incorrectly classified as completers), and the Type II error rate was 31.8% (completers incorrectly classified as non-completers). The results indicate that students were less likely (*OR* = .57 [.51–.64]) to complete evaluations after rather than before final exams, and they were less likely (*OR* = .47 [.42–.53]) to complete SETs during spring semester than in fall semester, $p < .001$.

To determine the probability of student completion of SETs, given the category of student cohort, a simultaneous, multinomial logistic regression was conducted, $\chi^2(2, n = 9{,}923) = 982.69$, $p < .001$, Cox & Snell $R^2 = .09$. Student cohort was also a significant predictor of completion of SETs, $p < .001$. Students in their first year ($n = 78$) were much more likely (*OR* = 5.28 [4.69–5.93]) to complete SETs than students in their third year ($n = 46$; Table 4.6).

Table 4.5

Summary of Core Course Instructor Evaluations During Academic Year 2017/2018

		Faculty evaluations		
Student cohort	Semester	Requested	Completed (rate)	Mean requested per student
First year Class of 2021	Fall (n = 89)	1424	905 (63.6%)	16
	Spring (n = 88)	792	313 (39.5%)	9
Second year Class of 2020	Fall (n = 85)	1782	443 (24.9%)	21
	Spring (n = 83)	2156	380 (17.6%)	26
Third year Class of 2019	Fall (n = 81)	2184	439 (20.1%)	27
	Spring (n = 81)	1680	364 (21.7%)	21

Table 4.6

Logistic Regression Model Parameters for Semester- and Timing-Based Predictors by SET Completion

Predictor	SET completion			
	df	χ^2	*OR*	95% CI
Semester	1	160.01***	0.47	[.42, .53]
Timing	1	96.87***	0.57	[.42, .64]
Cohort	1	776.63***	5.28	[4.69, 5.93]

Note. n = 10,018. *OR* = odds ratio; CI = confidence interval.

***p < .001.

To assess the probability of SET completion when students are given time after finals to complete them (compared to the 2016-2017 academic year), an additional simultaneous, binary logistic regression was completed, $\chi^2(1, n = 20{,}314) = 152.30$, $p < .001$, Cox & Snell $R^2 = .007$.

Academic year was also a significant predictor of completion of SETs at the $p < .001$ level. Students during the 2017-2018 (treatment) academic year were less likely to complete SETs than students during the 2016-2017 (control) academic year, $\chi^2(1, n = 20{,}314) = 152.30$, $p < .001$, odds ratio = .69 (.65 to .73).

Number of students providing comments. Of the 2,926 SETs students submitted before and after final exams, 1,149 (39.3%) included comments. Table 4.7 shows these data by timing of SET delivery.

To determine if the number of students submitting comments was significantly different due to timing, a χ^2 test was run as a 2 × 2 contingency table. None of the cells had an expected frequency less than 5, which indicates the appropriateness of the χ^2 test (McHugh, 2013). Although a slightly higher percentage of students provided comments after final exams, there was no statistically significant relationship between timing of SET delivery and students providing comments, χ^2 (1, $n = 2{,}926$) = 1.39. This result indicates that students were just as likely to provide comments before final exams as they were afterward.

Comment length. The 1,149 SET forms that included comments contained 50.24 mean words per comment. After aggregation of data to the instructor level, an independent *t* test was used to compare means between the before and after final exams groups. Although students' comments were lengthier after final exams ($M = 52.13$, $SD = 50.68$) than before ($M = 48.45$, $SD = 38.93$), these differences were not statistically significant.

Table 4.7

Students Who Provided Comments in SETs Before and After Final Exams (Timing)

Timing	SETs completed	Contained comments	Words per comment (*M*, *SD*)
Before	1,542	590 (38.3%)	48.45 (38.93)
After	1,384	559 (40.4%)	52.13 (50.68)
Total	2,926	1,149 (39.3%)	50.24 (45.05)

A Pearson correlation test was then conducted on the aggregated data to examine possible correlations between students' expected grade and comment length. There was no statistically significant correlation (r = -.05).

*RQ3 (*Qualitative*). How Does Veterinary Student Completion of SETs After Final Exams (Versus Before) Affect the Substance of Comments?*

Students' open-ended comments were analyzed using eclectic coding, incorporating simultaneous, magnitude, emergent, and elaborative strategies. Three different categories of codes were developed: (1) provisional (*a priori*) codes–based on SET item prompts, (2) emergent codes–based on concepts that emerged organically in the data, and (2) sentiment codes. Sentiment was classified *a priori* as praise, criticism, combination, or neutrality. Other categories are described more fully in the following sections. Names of courses and instructors have been changed to preserve anonymity.

Thematic findings. Taking into consideration all 1,149 student comments, three themes emerged from the data, with attribution theory informing thematic analysis. The themes that emerged were *classroom climate*, *achievement striving and goal attainment*, and *operational deliverables*. Following the structure in Table 4.8 and Figure 4.2, the following three sections present these thematic findings by richness and interrelatedness of the categories within each theme. Each section also contains an analysis of the themes grouped by timing (before and after final exams).

Classroom climate.

The most prominent theme to emerge from the data, *classroom climate*, is best defined by two categories: instructor-focused concepts and student-focused concepts. Students described

Table 4.8

Summary of Inductively Developed Themes in SET Comments

Themes	Associated Concepts	Description
Classroom Climate	Instructor-focused Level of knowledge Facilitation of learning Promotion or detraction of understanding and retention Level of care and support Level of dedication and effort Level of enthusiasm for teaching Level of enthusiasm for topic Engagement as speaker Personality of instructor Style of instructor	Student expresses perceived level of instructor knowledge; how/if instructor seems concerned with facilitating student learning; how well instructor promotes understanding and retention; perceived level of instructor support, self-sacrificing devoting, and commitment; instructor enthusiasm for teaching and subject matter; level of engagement as a speaker; personality of instructor (unrelated to caring, kindness); instructors' way of using their abilities
	Student-focused Appreciation Level of enjoyment	Student expresses gratitude and taking (or not) pleasure from instructor teaching
Achievement Striving and Goal Attainment	Rigor Studying Low stakes exams and alternative assignments Grades Clarity of expectations Clarity of assessment Exam format	Students' perceived rigor of course content and assessments; student study methods and how instructor methods facilitate studying; exam or exam question format; course grades or points; instructor clarity of expectations and information about testing and grading
Operational Deliverables	Organization Teaching aid effectivity Pace Curriculum Quantity of material Clinical applicability of material Emphasis of concepts	Students' perception of instructor's organization of lectures and/or laboratories; helpfulness and relevance of teaching aids; reasonableness of instructional pace; course content, time devoted to content, and timing of delivery in relation to other content; amount of course material; applicability of material to clinical education and practice; if and how well major concepts were emphasized

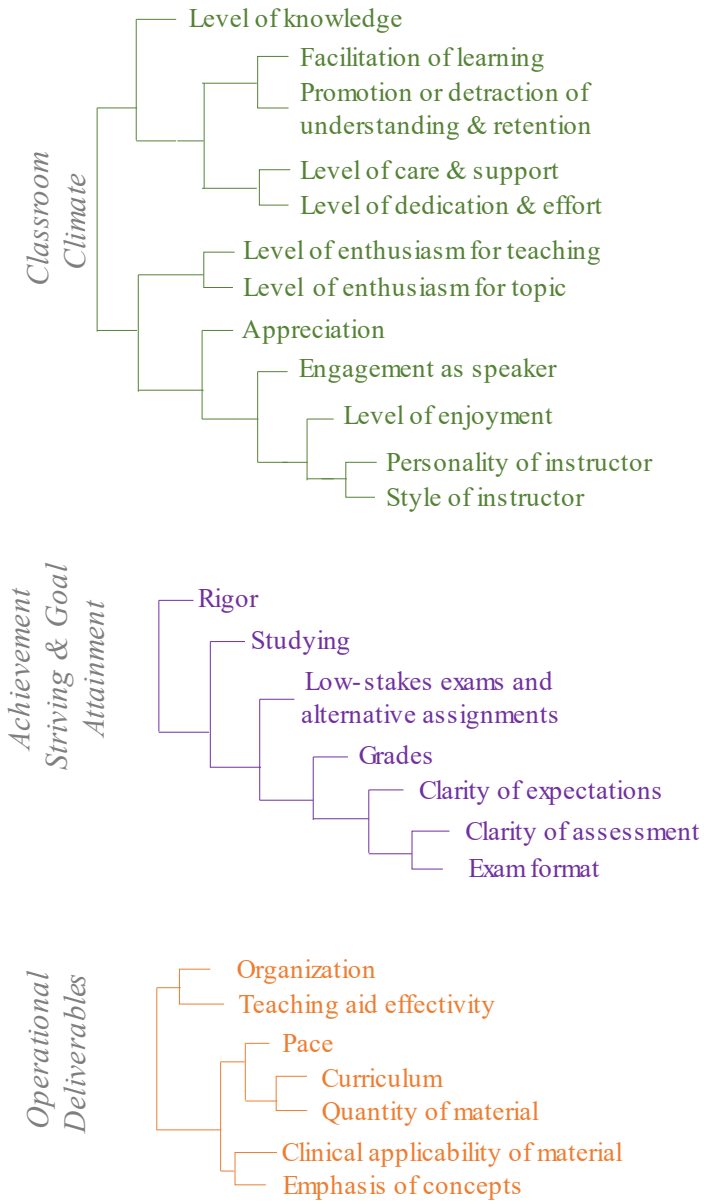

Figure 4.2. Themes and concepts clustered by word similarity (produced by NVivo 12).

their classroom climate based on their perceptions of the instructor's teaching, as well as their own feelings of gratitude and enjoyment.

Instructor-focused. Students discussed how instructor characteristics shaped how much and how well they learned. Not only did students describe methods instructors used to promote or detract from their understanding and retention of information, but they also expressed their perceptions of instructor personality, especially regarding caring and patience. Students reported being able to recognize when instructors were enthusiastic about their content or about teaching, in general. Instructor enthusiasm was often transferred to students. One student noted that the instructor's "love for learning makes getting excited about ophthalmology [pseudonym] (and cool gross lesions) easy!" Instructor engagement as a speaker motivated students to actively engage with the material, especially students who had little interest or previous instruction in the subject: "I was really nervous about bacteriology [pseudonym] before starting the semester but Dr. Shelton [pseudonym] made learning everything really interesting and fun!"

Although students did comment on their perception of instructor knowledge, they focused much more on how instructors seemed concerned with facilitating student learning and were supportive of and devoted to students and teaching. Regarding one teacher, a student wrote, "After doing poorly on the first exam, unlike many professors, he didn't say, 'Work harder,' or 'You didn't study.' INSTEAD, he said, 'We'll work together, and you can do this.'" In contrast, another student wrote about a different instructor: "Dr. Hamilton [pseudonym] was also very impatient and not encouraging at all during our . . . lab, which many people were nervous about. His presence only made us more anxious."

Student-focused. Student-focused concepts included students expressing appreciation and taking pleasure (or not) from an instructor's teaching. Students often thanked their instructors and acknowledged specific teaching methods that they appreciated. For example, when discussing pace, one student said, "Dr. Mayes [pseudonym] talks so fast!! She knows it and tries to slow down and I appreciate that." Although many students simply stated generalizations, such as, "I really enjoyed the lectures by Dr. Beason [pseudonym]," others expressed fondness for specific instructor attributes. Illustrating these aspects of emotion, students wrote, "Overall, I loved the focus Dr. Paul [pseudonym] had during her lectures," and "I also really liked submitting the . . . assignment to go with the exam instead of just having test questions."

Timing–before and after finals. Students' open-ended comments regarding classroom climate were similar regardless of whether students commented before or after final exams, with one exception. The most pronounced variation in thematic concepts existed with *personality of instructor*. Before final exams, student comments seemed to focus more on the instructor's personality than they did after final exams. At times, different students described the same instructor in vastly different terms. For example, "Dr. Hodges [pseudonym] can be very harsh and borderline rude at times" starkly contrasts with "Dr. Hodges is one of the most personable individuals I've ever met." Still, student comments about the instructor's personality were largely commendable rather than critical, regardless of the timing of SET delivery.

Achievement striving and goal attainment.

The theme of *achievement striving and goal attainment* emerged because it encompassed the interrelatedness of the data, but the naming of the theme was inspired by the foundations of

attributional theory. Weiner (1986) used Atkinson's theory of achievement striving to inform the development of his attributional theory of motivation and emotion. This connection is described in depth in Chapter 5.

Few students commented specifically about grades and point systems. Instead, they described how an instructor's methods affected their ability to study, perform well on exams, and identify concepts that are most important for a practicing veterinarian to know. Consequently, clarity of instructor expectations permeated this theme in the data. Students discussed how instructor-provided notes, presentation slides, and lectures facilitated or hindered their study efforts. One student noted, "I also felt like I knew what to focus on when studying because she emphasized major concepts." In contrast, another student commented that "Every concept taught by him was considered a major concept, making studying for tests extremely difficult." Students expected that the emphasis of topics in class would correspond to what was emphasized on exams, and they frequently noted when that did not occur: "I felt mislead [sic] for the exam. I focused far too much on initial presentation and narrowing down differential diagnoses, rather than the specific details of the . . . procedure itself (which comprised the majority of the final)." Nonetheless, students also praised instructors whose emphasis and exams correlated well: "Her test questions are very challenging but they are fair and go off the information taught."

Beyond striving for achievement on assessments, students seemed focused on attaining their goal of becoming a competent veterinarian. Many of their comments about the format of exams reflected student perceptions of knowledge that would be beneficial for a veterinary practitioner. Comments such as, "test questions often contain multiple right answers and focus on tiny details with limited clinical relevance" were differentiated from comments like, "I do like

that the exam questions are often clinically related and I think that makes it more interesting and easier to correlate to future use."

Timing–before and after finals. There was little distinction between student comments before versus after final exams regarding *achievement striving and goal attainment*. As with *classroom climate*, there were exceptions: comments related specifically to clarity of assessment and exam format seemed more pronounced after final exams. Even though students shifted the focus of their comments toward assessments after finals, these comments did not become more critical than they were before final exams, as detailed further in the section on comment sentiment.

Operational deliverables.

If *classroom climate* is related to the instructor and *achievement striving and goal attainment* are related to student outcomes, then *operational deliverables* might be thought of as the curriculum and the logistics that connect the instructor with student outcomes. Indeed, this theme emerged from students' communicating about course material and its organization and clinical applicability, as well as the pace at which instructors delivered this material. Summing up their thoughts, one student wrote, "The pace was occasionally too quick for me, but the amount of notes and their organization completely mitigates that." Remarking on clinical application, a student described how the instructor "offered a lot of textbook info, but also just very PRACTICAL information that I feel will be helpful in practice." Students described the efficacy of teaching aids, such as case discussions, presentation slides, videos, photographs, instructor-provided notes, laboratories, interactive classroom technology, games, review sessions, and storytelling. One student noted:

> I would love to see more visualization of the surgical techniques Dr. Fidell [pseudonym] discussed. A lab on a cadaver would be most beneficial. Learning surgery from a lecture is not the most efficient, definitely doesn't make me feel like I can go out and do any of the procedures described.

Students also were able to reflect on their courses at a macro level to discuss how well the course sessions fit in sequence with each other, as well as with other courses within the semester and within the entire curriculum. For example, one student recommended condensing a course to "avoid too much duplication with neurology," while another felt that the "material would have been better understood and retained long-term if this course was taught at a later date once the foundation of our veterinarian education had been established (maybe 2nd semester or 2nd year)." Students recognized when their instructors emphasized major concepts and wrote about when they perceived their instructors to be focused on minutia. One student felt that the instructor's "teaching style focuses on details that aren't clinically relevant, and he goes over things that only a pharmacologist [pseudonym] would care about or use. . .. many minute details he talks about that he emphasizes as much as big concepts."

Timing–before and after finals. Like the *classroom climate* and *achievement striving and goal attainment* themes, few distinctions emerged between student comments about *operational deliverables* before final exams versus afterward. The exceptions were that student observations about clinical applicability, teaching aid effectiveness, and instructor pace were more conspicuous before finals than they were after. Yet, like the two themes discussed above, student praise and criticism of these concepts were analogous before and after finals exams.

Thematic outliers.

Although the three themes described above were pervasive in the open-ended comments, students extended their perceptions about their instructors by writing about their admiration for them and making note of instructors' kindness and inspiring manner. Noted one student: "Dr. Gregg [pseudonym] made me want to be a pathologist [pseudonym]." Instructors' level of patience was also mentioned on both ends of the continuum, with one student writing that the instructor was sometimes "frustrated with my classmates when they were trying to understand grading and testing," while another student felt that a different instructor was "the epitome of patience as we ask questions again and again and often repeat our questions."

Students described feelings of regret, guilt, and sadness that an instructor was retiring and that students "gave little energy back" to the instructor in a class that was scheduled toward the end of the day. A concept atypical to the open-ended comments, student consumerism, was also mentioned after final exams but not before. A characteristic statement illustrating consumerism included, "This sort of learning experience is not what I am paying for."

Summary.

Overall, students provided insightful open-ended feedback about the essence of the learning environment and how their instructors affected the classroom climate. Additionally, students were able to illustrate how instructors facilitated or impeded achievement of student goals and provide their perceptions on the value of the curriculum and how it was packaged. Rarely did the substance of student comments depend on the timing of student completion of SETs (before or after finals), but some divergence did occur based on timing. Before final exams, student comments centered around the personality of their instructors, clinical

applicability of material, and teaching aid effectivity; after final exams, students focused on exam format and clarity of assessments. Nonetheless, these comments were similar in sentiment regardless of the timing of SET delivery–they were largely praise rather than criticism.

RQ4: (*Mixed*) How does veterinary student completion of SETs after final exams (versus before) affect the intensity of comment themes and magnitude of sentiment?

This section describes the quantitative findings of the qualitative data once they were transformed into frequencies of occurrence. Both intensity (frequency) of themes and magnitude of sentiment were analyzed.

Intensity of themes. Upon examining the probability of differences in themes between the before and after finals groups using a χ^2 test, only one significant result was obtained. Before final exams, for the theme of instructor-focused classroom climate, students were significantly more likely to comment on concepts like instructors' facilitation of learning; levels of enthusiasm, dedication, support, and knowledge; personality, and style than they were after final exams, χ^2 (9, n = 1,023) = 21.62, p = .010.

The following χ^2 comparisons were also made for the before and after groups based on findings in the qualitative data, but no significant results were found:

- Classroom climate (student-focused), χ^2 (1, n = 363) = .46, p = .496
    - Personality of instructor and level of engagement as a speaker, χ^2 (1, n = 273) = .96, p = .327
- Achievement striving and goal attainment: χ^2 (6, n = 720) = 7.12, p = .310
    - Exam format and clarity of assessment, χ^2 (1, n = 394) = .70, p = .403
- Operational deliverables, χ^2 (6, n = 1,073) = 10.95, p = .090

- Clinical applicability of material, teaching aid effectivity, and pace, χ^2 (2, n = 638) = 4.54, p = .103

Comment sentiment. As shown in Figure 4.3, overall, students provided more praise than criticism, both before and after final exams. Table 4.9 shows a word similarity index between data coded by sentiment and by category. Praise was more highly correlated with *classroom climate*, whereas criticism was more highly associated with *achievement striving and goal attainment* and *operational deliverables*. When the combination category was combined with praise and criticism, respectively, to create two new categories, student comments were 36.3% more positive than critical before final exams and 38.9% more positive than critical after finals.

The results of a χ^2 test to determine differences between actual and expected results indicated there was no statistical difference in sentiment before final exams versus after, χ^2 (1, n = 627) = 0.2, p = .903. These results supported the results of the qualitative analysis and indicated that student praise and criticism were statistically similar regardless of when students provided open-ended comments.

Before final exams, student praise occurred most frequently in comments in which students were expressing appreciation for the instructor and enjoyment of some aspect of the instructor's teaching. One example is a student acknowledging that they "thoroughly enjoyed listening to her and soaking up all the wisdom she provides." When students were critical, they were much less likely to express appreciation and enjoyment. After finals, students were slightly more critical of clarity of assessment and exam format, as evidenced here: "his exam questions . . . seemed to focus all on one concept, which was surprising to me because he taught a good deal of information that ranged from physiology concepts to multifactorial issues."

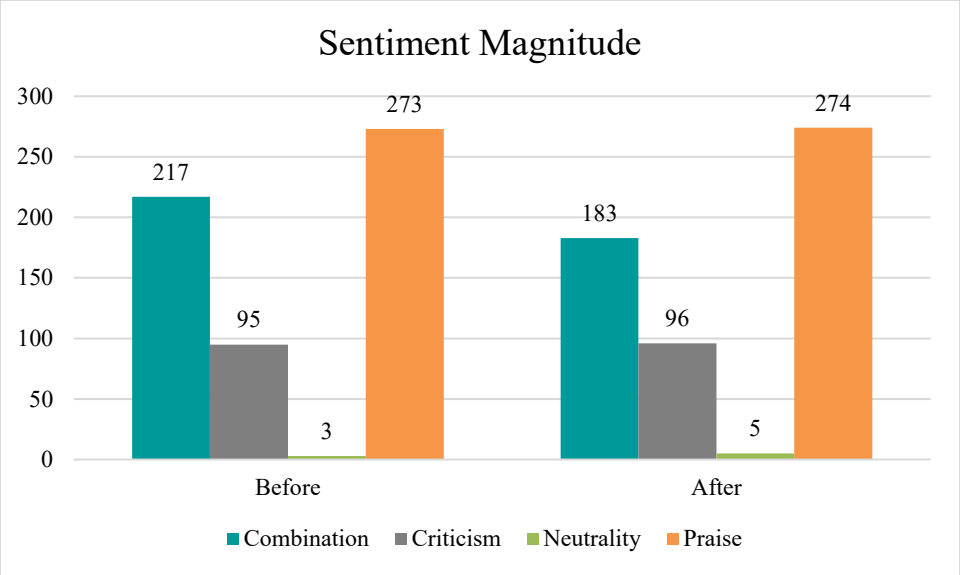

Figure 4.3. Intensity of sentiment magnitude in SET comments before and after final exams

Table 4.9
Correlation Matrix of Word Similarity for Sentiment and Concept Coding

Theme & Concepts	Sentiment (*r*)	
	Praise	Criticism
Classroom climate		
Promotion or detraction of understanding & retention	.78	-
Level of dedication and effort	.72	-
Personality of instructor	.78	-
Style of instructor	.84	-
Achievement striving and goal attainment		
Grades	.46	.72
Clarity of expectations	-	.82
Clarity of assessment	.51	.79
Exam format	.50	.76
Operational deliverables		
Teaching aid effectivity	-	.69
Pace	.52	.67
Quantity of material	-	.71

Note. Pearson correlation coefficients (*r*) produced by NVivo 12 for Windows. *p* values not produced because significance tests are inappropriate for use in qualitative data analysis (M. Shakeb, personal communication, April 16, 2021).

Summary

This chapter presented the quantitative results of SET data analysis and the qualitative inferences of students' open-ended comments. Some statistical differences were seen with student timing, cohorts, semester, and expected grade, but overall, both paths of analysis suggest little difference between SET responses after final exams compared to before final exams. The open-ended comments provided rich data, resulting in themes that paralleled with and supported quantitative results. The three themes that emerged were *classroom climate*, *achievement striving and goal attainment*, and *operational deliverables*. Chapter 5 will analyze the findings and results, outline the implications of the study, discuss study limitations, and suggest paths for future research.

CHAPTER FIVE
DISCUSSION AND RECOMMENDATIONS

Previous chapters established the problem and outlined study methods and findings. In this chapter, the findings and results are examined and connected to existing research and theory. Each section includes a discussion of the findings and a connection to the existing body of literature. As detailed in Chapter 4, generally, there were few differences in student evaluation of teaching (SET) scores or students' open-ended comments regardless of whether students completed the SETs before or after final exams. However, students were more likely to complete SETs before final exams, and making SETs available after final exams did not increase the completion rate when compared to the previous academic year (control). Differences were seen with expected grade, student cohort, and semester variables, and those differences are explored below.

Sections in this chapter are arranged by research method type and further subdivided by research question topic. Each subsection connects the findings to existing research and explores the relationship of the results with attribution theory and attributional bias, where applicable.

Quantitative and Mixed Method Findings

The findings discussed below are based on the following three research questions:

RQ1. (*Quantitative*) How does veterinary student completion of SETs after final exams (versus before final exams) affect numerical scores on SET items?

RQ2. (*Quantitative*) How does administering SETs to veterinary students after final exams (versus before) affect completion rate and volume of comments?

RQ4. (*Mixed*) How does veterinary student completion of SETs after final exams (versus before) affect the intensity of comment themes and magnitude of sentiment?

Changes in Numerical Scores of SET Items

Timing of SET availability (before or after final exams) had no effect on SET item scores. This finding supports two previous studies conducted in medical schools (Canaday et al., 1978; McNulty et al., 2010). Interestingly, the Canaday et al. study was published in 1978, before online SETs and before professionalism was routinely taught as a non-technical competency in professional schools (Medical Schools Objectives Project Writing Group, 1999). The effect of timing on SET scores seems to be unrelated to SET administration method and instruction in professionalism within the curriculum.

Notably, there were positive correlations between students' expected grades and SET items. As students' expected grades increased, so did most SET item scores (Figure 4.1). In a previous related study, Arnold (2009) found SET scores increased when grades increased but only *after* final exams. Distinguishing these results from Arnold's (2009) study is that SET item scores were lower both before and after final exams, indicating no support for a timing effect. These results are consistent with previous reports of significant positive correlations between expected grades and evaluations (Aleamoni, 1999; Algozzine et al., 2004; Clayson, 2009). Such results endorse the reciprocity hypothesis, in which students with higher grades are expected to provide more positive evaluations (Clayson, 2009). Nonetheless, like findings in previous studies, the correlation values in the current study ranged between r = .14–.38 (Aleamoni, 1999; Algozzine et al., 2004; Hativa, 2013). Correlations < .50 are considered small and have little practical relevance (Hinkle et al., 2003). In a meta-analysis that used exam scores as measures of

student learning, the more objective the exam, the smaller the association between learning and SET scores (Clayson, 2009). Most of the exams for participants in this study comprised multiple choice questions, which are arguably one of the most objectively scored types of exam questions (Downing, 2002). Exam objectivity could help account for the small correlations between grades and SET scores.

Cohort differences also were seen with several SET items. Students in their third year rated instructors significantly higher than students in their first year for the following items: 2) Lectures/labs, etc. organized with clear objectives, 3) Instructor willingness to discuss course material outside of class time, 8) Instructor pace of covering the material, 9) Clarity and specificity of directions for course assignments, and 10) Overall opinion of instructor teaching skills. These results suggest that third-year students might be more familiar than first-year students with how lectures, laboratories, and course assignments are structured in veterinary school. The additive experience, familiarity with instructors, and knowledge of third-year students might also help explain their higher ratings for instructor pace and comfortableness in seeking instructor assistance outside of class time. Third-year students' overall opinion of instructor teaching skills might be influenced by the variation in the third-year curriculum, which is focused more on clinical applicability of content covered in the 2 years prior.

Attributional bias. Taken together, the quantitative results described above suggest that attributional bias was likely not occurring due to timing of SET within this population of veterinary students but could be occurring due to students' expected grade. With attributional bias, people attribute their successes to themselves and their failures to someone else (Arnold, 2009; Korn et al., 2016; Robinson, 2017). With SETs, a situational attributional bias

(externalizing failure) has been proposed to occur with students who have performed poorly in a course (Jhangiani & Tarry, 2011). The success-driven characteristics of veterinary students suggest that they would be likely to display attributional bias, which is more common in highly successful people (Tetlock & Levi, 1982; Zenner et al., 2005).

Decrease in Completion Rate

Low response rates and nonresponse bias are a concern with SETs (Bacon et al., 2016; Reisenwitz, 2016; Royal, 2017; Young et al., 2019), and providing veterinary students the opportunity to complete SETs after final exams was explored as a way to increase response rates for pre-clinical courses. For several years, students at this study site had requested to be given time to complete SETs after final exams and had indicated they would be more likely to complete them at that time. Making SETs available after final exams did not increase response rates in the current study. In fact, response rates decreased. When compared to the 2016-2017 academic year (control), students in 2017-2018 (treatment year) were 69% less likely to complete SETs.

Completion of SETs was also higher during fall 2017 semester (versus spring 2018) and within the first-year cohort (versus third year). One reason for this outcome is likely because, for this research site, fewer instructors taught in the first semester than in later semesters. Therefore, first-year students were being asked to complete fewer SETs than in later semesters. These findings might also be considered in terms of the "honeymoon stage" (Killinger, Flanagan, Castine, & Howard, 2017, p. 6). During the first year of the 4-year program and during the first (fall) semester of a new academic year, student zeal is likely higher than in any subsequent semester (Killinger et al., 2017). With first-year students, the newness of veterinary school and

the emphasis placed on professional responsibility could contribute to a higher completion rate. Additionally, research into passive non-respondents (more common than active, intentional non-respondents) indicates that an overload of work and lack of time prevent people from completing surveys (Barr et al., 2008). Indeed, veterinary student anxiety and academic stress has been shown to increase from first to third semesters (Reisbig et al., 2012).

These results suggest that veterinary schools might need to seek other ways to increase response rates to SETs among their pre-clinical students to increase representation and accuracy of the results and decrease nonresponse bias; the more respondents, the higher the reliability of the results (Benton & Cashin, 2010; Hativa, 2013; Richardson, 2005). Some alternative methods to consider to increase response rates are discussed in the section on recommendations.

Equal Volume of Comments

In a previous study conducted at a medical school, students tended to give fewer and less substantial open-ended comments after final exams (McNulty et al., 2010). In contrast, in the current study, there were no statistical differences in the number of students providing comments nor in comment length between before and after finals groups, even though slightly more students provided open-ended comments after final exams. Additionally, no statistical association existed between students' expected grade and comment length.

A potential reason for the difference between results in a veterinary school versus a medical school is the gender makeup of the two. Several researchers have found that female students are more likely to respond to SETs (Bacon et al., 2016; Porter & Whitcomb, 2005; Reisenwitz, 2016; Young et al., 2019). Females represented 49% of participants in the medical school study, whereas 82% of the students in the current study were female (McNulty et al.,

2010). If females are also more willing than males to extend their effort to provide qualitative feedback, this could account for the similarity in student comment volume between the before and after finals groups.

Qualitative Method Findings

The qualitative findings discussed below are based on the following research question:

RQ3. How does veterinary student completion of SETs after final exams (versus before final exams) affect the substance of comments?

Substance of Open-Ended Comments

Three themes emerged from students' open-ended comments on the SET form: *classroom climate*, *achievement striving and goal attainment*, and *operational deliverables*. Comments within the *classroom climate* theme were largely related to student perceptions of the instructor, whereas comments within the theme of *achievement striving and goal attainment* were related to student outcomes. *Operational deliverables* comments described the curriculum and logistics that connected the instructor with student outcomes.

Shaping students' comments on *classroom climate* was the adage that people don't care about how much you know until they know how much you care. The *classroom climate* theme encompassed concepts such as how well instructors promoted the understanding and retention of information and how students perceived instructor enthusiasm and engagement. The focus on instructor personality in this theme could be construed as support for faculty perceptions that SETs largely measure instructor charisma (Hammer et al., 2018). However, in the current study, student perceptions of instructor personality were generally favorable and referred mostly to instructors' level of caring and patience; rarely did a student write negative feedback about

instructor personality. A construct that permeated the classroom climate theme was student appreciation and enjoyment.

Within *achievement striving and goal attainment*, students communicated about how instructors' methods directed their studying, performance on exams, and ability to recognize the most important concepts for a practicing veterinarian to know. The theme of *achievement striving and goal attainment* emerged because it encompassed the interrelatedness of the data, but the naming of the theme was inspired by the foundations of attributional theory. Weiner (1986) used Atkinson's theory of achievement striving to inform the development of his attributional theory of motivation and emotion. Weiner (1986) identified achievement striving as one of the most dominant sources of motivation. Perceived causes of success include students' own ability but also instructor competence (Weiner, 1986).

Operational deliverables may be thought of as the curriculum and logistics connecting the instructor with student outcomes. This theme emerged from student comments about course material, its organization and clinical applicability, and instructors' pace of delivery. Students also examined their courses from a macro perspective to consider how well course sessions were sequenced and how well a given course fit with other courses within the semester and within the entire curriculum. Notable is that among SET items, pace of delivery and usefulness of teaching aids were scored lowest by students who expected the lowest grades in each course.

Attributional theory and bias. Considering the theme of *achievement striving and goal attainment* together with *classroom climate*, it is evident that students overwhelmingly expressed gratitude toward their instructors for assisting them in achieving their goal of becoming a veterinarian. Appreciation is a consequence of anticipating successful completion of a goal

(Weiner, 1986). The results of the current study therefore support attributional theory but not attributional *bias* in students' comments. Although students have been reported to search for causality of failure more than causality of success, causal attribution has been shown to occur for success, as well (Zuckerman, 1979). In particular, Weiner (1979) found that if success was perceived to be based on assistance from others, the resulting emotion was most often reported to be gratitude. Furthermore, females have been shown to ascribe achievement outcomes more to external factors than males (Zuckerman, 1979). The participants in this study were 82% female.

Summary

The three thematic concepts fluctuated little between before and after finals groups. Before final exams, students seemed to focus more on instructor personality than after final exams. Student praise was most frequent in instances in which students expressed appreciation for the instructor and enjoyment of the instructor's teaching. Comments related to clarity of assessment and exam format were more evident after final exams. Even though students shifted their comments toward exams after finals, critical comments did not become more pronounced than they were before final exams. Generally, students provided more praise than criticism, both before and after final exams. No statistical difference existed in sentiment magnitude before final exams versus after. Here again, there was no support for students exhibiting attributional bias based on completion of SETs after final exams.

These results also refute prior empirical studies about faculty attitudes toward SETs. Faculty have been reported to perceive SETs as a platform for students to vent frustrations, be vindictive, and express anger (Eckhaus & Davidovitch, 2019; Hammer et al., 2018; Nasser & Fresko, 2002). Additionally, faculty have reported feeling that SETs mostly assess an instructor's

charisma (Hammer et al., 2018). Although instructor personality was a major concept in student comments, these comments were overwhelmingly complementary of instructor enthusiasm, engagement, and devotion. These results support previous findings that SETs are more than measures of popularity, and that students are capable assessors of teaching (distinguished from teaching *effectiveness*), permitted that SET results are contextualized as student perceptions of their own learning (Kreitzer & Sweet-Cushman, 2021; Nasser & Fresko, 2002).

Practical Implications and Recommendations

Students are uniquely situated to consistently evaluate teaching because they participate in longitudinal evaluation, whereas faculty peers and administrator perspectives are often cross-sectional. To improve the reliability of results, colleges must determine ways to increase response rates and ensure that their teaching evaluation program has faculty, administrator, and student support. Low response rates threaten the reliability of SETs by introducing nonresponse bias (Clayson, 2009; Hativa, 2013). Additionally, interrater reliability values increase as the number of responses increases (Benton & Cashion, 2010; Hativa, 2013). Discussed below are evidence-based methods to increase response rates and recommendations for refining and improving programs to assess teaching effectiveness.

Increasing Response Rates

Online SETs allow for more flexibility for students regarding when SETs may be completed. Most often, SETs are administered and due before the final exam period begins (Arnold, 2009; Estelami, 2015; Hativa, 2013; Hoefer et al., 2012; Young et al., 2019). In this study, opening SETs after final exams did not increase response rate in the population of students under study, suggesting that veterinary schools should seek other methods.

Due to the team-teaching nature of veterinary courses, veterinary students have more instructors to evaluate than typical undergraduate and graduate students. At the site of this study, each semester, pre-clinical veterinary students receive an average of 34 different SETs to complete. Malone and colleagues (2018) reported requesting over 50 SETs from each student, each semester at the University of Minnesota. To increase the response rate at the North Carolina State University College of Veterinary Medicine, Kenneth Royal developed a model that included a designated time for students to complete online evaluations and alternated evaluation of instructors so that not every instructor is evaluated every year (personal communication, May 18, 2020). For alternating evaluations, instructor rank should be considered so that those with upcoming promotion and tenure evaluations will have enough SET data to help inform peer and administrator review. Random sampling of students is also a way to decrease the number of SETs being requested per student and possibly increase completion rate. As shown in the current study, response rate was highest in the first semester in the first year, when students had fewer instructors to evaluate. However, it is unknown if the number of evaluations or the ardor of new veterinary students is responsible for that high completion rate.

Another way to reduce the number of SETs students are requested to complete would be to consolidate forms. With an intentional process evaluation, college administrators could determine approaches to merge SET forms and still maintain the validity of the data obtained. As outlined by Royse, Thyer, and Padgett (2016), a formal process evaluation helps explain why desired outcomes (higher response rates) were not achieved. Part of a process evaluation is examining the forms used to help assure quality (Royse et al., 2016). Possible considerations are to merge course and instructor evaluation forms into a single form to be completed for each

course, merge all course evaluations into a single course evaluation form for each cohort, or to merge instructor evaluation forms into a single evaluation form to be completed for each course.

In several studies, the use of various types of incentives was examined to determine their effect on response rates to online surveys, including SETs. Personalized correspondence and salience of the survey topic were both shown to increase responses (Chapman & Joines, 2017). Others have noted that positive incentives, such as extra credit or making SET completion a course assignment, helped achieve higher response rates (Chapman & Joines, 2017). Negative incentives have also been explored. Some medical schools withhold grades, progression, or transcripts if students do not complete SETs, but the ethics of this technique could be questionable (Estelami, 2015; Royal, 2017). Other universities have withheld early access to grades unless SETs were completed (Chapman & Joines, 2017). The site for this study also withholds early access to grades for clinical students and consistently sees response rates above 80%. However, many university policies prohibit the use of incentives, so college administrators should consult existing policies at their own institutions (Chapman & Joines, 2017).

Chapman and Joines (2017) examined the practices of 205 instructors who achieved SET response rates of 70% or higher and found that higher SET response rates were associated more with classroom climate than with incentives. Previous researchers have speculated that students believe that their feedback is unimportant; therefore, the benefit of completing SETs is outweighed by the time cost of responding outside of class (Adams, 2012; Benton & Cashin, 2010; Chapman & Joines, 2017; Estelami, 2015; Young et al., 2019). Several researchers found that reminding students about the importance of SETs and stressing students' social responsibility have helped increase web-based response rates (Chapman & Joines, 2017). When

students feel that their instructors take student feedback seriously and that improvements are being made based on that feedback, they become more engaged and motivated as evaluators (Chapman & Joines, 2017). The top three strategies used by instructors in Chapman and Joines's (2017) study were talking about the importance of SETs in class, creating a classroom climate reflective of mutual respect, and telling students how SETs are used to modify a course. However, in a nursing school, an approach similar to that of Chapman and Joines (2017) did not affect response rates (Gordon, Stevenson, Brookhart, & Oermann, 2018). Gordon et al. (2018) found that grade incentives were predictors of increased response rates in a population of nursing students. Nonetheless, incentives such as extra credit, dropping a low assignment grade, and bringing snacks have the potential to bias data and could be considered ethically problematic (Chapman & Joines, 2017; Kreitzer & Sweet-Cushman, 2021).

An alternative to SETs is the use of non-participant student observers. In a 2021 qualitative study, students observed a pre-clinical veterinary course in which they were not enrolled and thereafter developed a summary and critical analysis of their experiences; these students attended at least one lecture or laboratory by a different instructor each day for 2 weeks (Hofmeister). Hofmeister's (2021) recommendation was to use such case studies to provide additional insight into evaluation of teaching. Given the heavy course load of veterinary students, this approach might not be feasible for long-term application.

Refining and Improving Programs to Assess Teaching Effectiveness

Faculty and administrator support and mutual understanding are required for a successful program to assess teaching effectiveness, especially since the oppositional attitudes of some instructors tend to also be transferred to the classroom, where students hear that their feedback

does not matter (Seldin, 1999b). Users of SET data should be educated on the value of such information and on the reliability of student feedback (Adams, 2012; Addison & Stowell, 2012; American Sociological Association, 2019; Hativa, 2013; McCarthy, 2012). Education on the proper use of SETs in combination with other measures of teaching quality is also important (Addison & Stowell, 2012; McKeachie, 1969). Faculty and administrators value evidence and knowledge; developing a way forward requires showing that programmatic decisions about SETs have been based on trustworthy evidence and input from stakeholders. Carol Pearson (2012) stated that for change to take place, leaders need to not only educate, but also to inspire and motivate. To help ensure instructor acceptance, leaders of SET programs should emphasize open communication, mutual goals, and common interests for the greater good (Bolman & Gallos, 2011).

However, faculty and administrator concerns must be heard and addressed before attempting to educate them about the value of SETs. Mapping a plan that is best for the organization and all its stakeholders will require listening and careful consideration of the interests of the organization and the interests of individuals (Bazerman & Tenbrunsel, 2011). Elisabeth Kubler-Ross (as cited in Hodgson & Ilkiw, 2017) described the first stage of any type of change as initiating shock and denial and noted that during this first stage, communication is crucial to try to gain eventual acceptance. Bolman and Gallos (2011) suggested being diligent in collecting feedback and beginning any feedback session with a question like, “What do you hope we can accomplish, and how should we proceed to make sure that happens?” (p. 38). Engagement of the faculty enables them to take ownership of the process (Higgerson, 1999).

Additionally, in medical schools, faculty governance and dedicated leadership have been found to aid change (Hodgson & Ilkiw, 2017).

The timing of the introduction of a teaching assessment program will also affect its acceptance by instructors (Bolman & Gallos, 2011). If the faculty is not ready to make a change, or if timing is not thoroughly considered, it might be difficult to gain the support of the faculty (Gini & Green, 2013). Discussion of changes to a teaching effectiveness program just before the end of a semester or annual performance reviews might seem too delayed because the changes could not occur in enough time to implement them before their immediate use. Improper timing might also result in quick, intuitive, off-the-cuff decisions rather than slower and more thoughtful decisions based on logic and cost/benefit analysis (Bazerman & Tenbrunsel, 2011, p. 35). Timing also might be influenced by upcoming accreditation reviews and site visits. The American Veterinary Medical Association's Council on Education (2020a), the U.S. accrediting body for veterinary schools, requires SETs for schools to maintain accreditation, and site visits occur every 7 years. Ultimately, the organizational climate must also be scanned to determine if constituents are ready for a discussion of an evaluation of teaching program (Higgerson, 1999). This might be achieved with a survey.

Student input also should be considered when revising a program to evaluate teaching. Shapiro and Stefkovich (2005) described learners as being at the heart of the educational process, a process that should include their feedback on organizational policy and procedures (American Association of University Professors, 1990). Aside from gathering student perspectives, we must also educate students on their role in the evaluation of teaching (Marincovich, 1999). Students must understand how to provide constructive, useful feedback; this might require student training

sessions or other educational strategies (Marincovich). Marincovich suggested that when instructors model constructive feedback, students tend to take on that same attitude. In addition, learners should understand that their participation in the SET process is a classic "social dilemma" (Bazerman & Tenbrunsel, 2011, p. 64). Completing SETs is much like a clinical trial, in which the decision to participate requires people to contribute now to help others in the future rather than immediately help themselves (Bazerman & Tenbrunsel, 2011). In this way, a successful program will require learners to understand that they have a professional responsibility to participate in the process. Learners should also be involved in deciding how SET data will be used, which will help them become vested in the process and better understand the importance of SETs (Cashin, 1999). At many universities, aggregated SET item scores are made available to students (Beran et al., 2012).

Methodological Limitations

This section discusses the limitations in the methods and what was done to ensure the accuracy of the results and increase their generalizability to other populations.

SET Form

Although the SET form used to collect the data has been used to provide feedback to instructors and in annual review, promotion, and tenure decisions for approximately 10 years, it has not undergone statistical analysis to determine its reliability and validity. However, Huck (2008) explained that reliability may be established by assessing a survey's validity, and a frequently used, non-statistical type of validity is content validity. To establish content validity, the creators of the SET form conducted a thorough literature review and faculty focus groups (Beran et al., 2012; Fink, 2017; N. Howell, personal communication, February 6, 2021). Having

such expert review helps ensure consistency between the conditions under which the survey is created and the conditions under which it is administered (Huck, 2008).

Many researchers have indicated that SETs, regardless of their origin, are reliable measures of instruction (Beran et al., 2012). Most validation studies have shown evidence of content validity (Beran et al., 2012). Additionally, much empirical research has shown that factors widely believed to affect SETs have little to no influence on student ratings. The consistency of scores across students, time, and context indicates a reliable instrument, and SET scores completed at the end of a course and again several years later have shown high reliability (Beran et al., 2012).

Use of Ordinal Data in Inferential Statistics

Some controversy exists with using ordinal data (scale items), rather than interval data, in inferential statistics. Some statisticians maintain that it is inappropriate to use ordinal data in statistical tests that use measures of central tendency like means (McCullough & Radson, 2011; Wu & Leung, 2017). Kuzon, Urbancheck, and McCabe (as cited in Wu & Leung, 2017) went as far as to call it "one of the seven deadly sins of statistical analysis" (p. 528). Their reasoning is that interval data have equal magnitudes between points; however, the points between Likert scale ratings do not necessarily have equal magnitudes (Lærd Statistics, 2018; McCullough & Radson, 2011). For example, the difference between 35 and 36 degrees on a thermometer (interval) is the same as the difference between 38 and 39 degrees, but the difference between 1 and 2 on a Likert scale (ordinal) is likely not the same as the difference between 4 and 5 on that same scale. Other researchers and statisticians have argued that meaningful results can be acquired with Likert data, especially as the number of points increases (Wu & Leung, 2017). In

practice, SET data are frequently displayed and analyzed with means, so the decision was made to go forward with the analysis and take care to avoid declaring that a student who marks 4 out of 5 is twice as satisfied as a student who marks 2 out of 5 on an SET (Estelami, 2015; McCullough & Radson, 2011; McNulty et al., 2010).

Generalizability of Results

A limitation of the study is that it was restricted to one veterinary school with three cohorts with approximately 85 students in each cohort. It is possible that the results will not generalize to other veterinary schools within the United States. The demographics of the sample have been examined to determine if significant differences exist between study site students and those throughout the United States during the same time period. No significant differences were found. With similar demographics, the results of this study are more generalizable to the larger veterinary student population. Still, generalizability of the results should be approached with caution. Increasing the number of participatory colleges would help increase the external validity of the study.

Hawthorne Effect

Also of genuine concern is the Hawthorne effect, in which the awareness of being studied might influence behavior (McCambridge, Witton, & Elbourne, 2014). Because the associate dean for academic affairs at the study site sent an e-mail explaining the process to students, they were aware of why evaluations were being conducted in this manner. If this cohesive body of students wished to sway the response rates, they could easily have done so. To alleviate the Hawthorne effect, this process should likely be carried out over the course of at least 3 years to

encompass an entire student cohort's progression through the pre-clinical curriculum. However, the short time available for dissertation work precludes such an extended study.

Future Research

Future research should attempt to identify whether non-respondents are making a conscious decision not to complete SETs (active non-respondents), or if they are prevented from completing the SETs due to extenuating circumstances, like overload of work and lack of time (passive non-respondents) (Barr et al., 2008). Such knowledge could help identify the best ways to increase response rates. Passive non-respondents might need their SET burden reduced, whereas active non-respondents might need a better understanding of how survey results are used and acted upon (Barr et al., 2008; Young et al., 2019).

Modifying the current case study to include more veterinary medical schools over a longer period would help improve the generalizability of the results. Additionally, setting up the SET randomization so that the participants in each timing group were truly independent would enable more parsimonious and robust statistical analyses. However, that approach would mean that some students would receive all SETs before final exams, and some students would receive all SETs after final exams. Alternatively, participant responses could be matched, but that would eliminate the anonymity of SETs. Finally, further research should investigate the outcome of making SETs available to students before finals and leaving the completion window open until after finals.

Summary and Conclusions

The purpose of this study was to determine how issuing SETs to veterinary students after final exams would affect item scores, completion rate, and comments, in relation to attributional

theory and bias. Knowing how timing and student knowledge of grades affect SET results can help inform veterinary college personnel on best practices for timing of release and closure of evaluations to help reduce nonresponse bias. In this study, timing of completion had no statistical effect on SET item score, number of students providing comments, nor comment length. An analysis of the intensity of themes before finals showed that students were more likely to comment on concepts like instructors' facilitation of learning; levels of enthusiasm, dedication, support, and knowledge; personality, and style than they were after final exams. Otherwise, students' open-ended qualitative comments were similar in theme and sentiment before and after final exams. However, when students completed SETs after final exams, response rates *decreased*.

Non-timing variables that were found to affect SET item scores included student cohort and expected grade, while cohort and semester affected completion rate. Attributional bias did not seem to occur due to timing of SET delivery; however, there was evidence to support the occurrence of attributional bias due to students' expected grade since students' expected grades were positively correlated with SET item scores.

Support for attributional *theory* was seen in the qualitative data. Considering the theme of *achievement striving and goal attainment* together with *classroom climate*, it is evident that students expressed gratitude toward their instructors for assisting them in achieving their goal of becoming a veterinarian. The results of the current study therefore support attributional theory but not attributional bias in students' comments. Causal attribution has been shown to occur for success, particularly if success is perceived to be based on assistance from others; the resulting emotion was most often reported to be gratitude (Weiner, 1979; Zuckerman, 1979).

The veterinary medical education literature on SETs is sparse. No prior studies were identified regarding timing and knowledge of expected or actual grade on SETs in veterinary schools, nor attributional bias in SETs. Most previous studies regarding timing of SETs and student knowledge of grade were conducted with undergraduate and non-veterinary graduate students. This study helps fill those gaps in the literature, as well as extends the understanding of how timing and knowledge of grade affect attributional bias theory in a different population of students.

Because SETs are used by administrators in making decisions about faculty promotion, tenure, and salary, it is important to maintain an acceptable response rate, collect SET responses at the most ideal time, and ensure that results are used alongside other documentation of teaching effectiveness (Algozzine et al., 2004; Berk, 2005; Seldin, 1999a; Zubizarreta, 1999). The results of this study suggest that veterinary college administrators should examine methods (other than SET completion after final exams) to raise response rates and to better train students on how to give constructive feedback (LaBelle & Martin, 2014). This research could help guide practices at other health professional schools like those training medical, dentistry, pharmacy, and nursing students. In health professional programs, teaching effectiveness, measured in part by SETs, affects student learning, the transfer of learning to the workplace, and eventually to the profession itself (Beran et al., 2012; Mosier et al., 1993).

REFERENCES

Adams, C. (2012). On-line measures of student evaluation of instruction. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 50–59). doi:10.1093/bja/aes563

Adams, M. J. D., & Umbach, P. D. (2010). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, *53*, 576–591.

Addison, W. E., & Stowell, J. R. (2012). Conducting research on student evaluations of teaching. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 1–12). doi:10.1093/bja/aes563

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153–166.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., . . . Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, *52*, 134–141. doi:10.3200/CTCH.52.4.132-141

American Association of University Professors. (1990). *Statement on government of colleges and universities*. Retrieved from https://www.aaup.org/report/statement-government-colleges-and-universities

American Sociological Association. (2019, September). *Statement on student evaluations of teaching.* Retrieved from https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_feb132020.pdf

American Veterinary Medical Association. (2020a). *Accreditation policies and procedures of the AVMA Council on Education.* Retrieved from https://www.avma.org/sites/default/files/2020-10/COE-pp-September-2020.pdf

American Veterinary Medical Association. (2020b). *Accredited veterinary colleges.* Retrieved from https://www.avma.org/education/accredited-veterinary-colleges

Anfara, V. A., & Mertz, N. T. (2015a). Closing the loop. In V. A. Anfara & N. T. Mertz (Eds.), *Theoretical frameworks in qualitative research* (2nd ed., pp. 227–236). Thousand Oaks, CA: SAGE Publications.

Anfara, V. A., & Mertz, N. T. (2015b). Setting the stage. In V. A. Anfara & N. T. Mertz (Eds.), *Theoretical frameworks in qualitative research* (2nd ed., pp. 1–20). Thousand Oaks, CA: SAGE Publications.

Anfara, V. A., Brown, K. M., & Mangione, T. L. (2002, October). Qualitative analysis on stage: Making the research process more public. *Educational Researcher*, 28–38.

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research*, *48*, 215–224.

Association of American Medical Colleges. (2020). *Applicants, first-time applicants, acceptees, and matriculants to U.S. medical schools by sex, 2010-2011 through 2019-2020*. Retrieved from https://www.aamc.org/system/files/2019-10/2019_FACTS_Table_A-7.2.pdf

Association of American Veterinary Medical Colleges. (2020a). *Admitted student statistics*. Retrieved from https://www.aavmc.org/becoming-a-veterinarian/what-to-know-before-you-apply/admitted-student-statistics/

Association of American Veterinary Medical Colleges. (2020b). *Annual data report 2019–2020*. Retrieved from https://www.aavmc.org/assets/site_18/files/data/2019%20aavmc%20annual%20data%20report%20(id%20100175).pdf

Bacon, D. R., Johnson, C. R., & Stewart, K. A. (2016). Nonresponse bias in student evaluations of teaching. *Marketing Education Review*, *26*, 93–104.

Bailey, M. (2016). *UTCVM Assessment history*. Retrieved from https://vetmed.tennessee.edu/academics/Documents/Assessment%20History%202016.pdf

Bailey, M., & Mawby, D. (2018, October). *Timing of student evaluations of teaching in a veterinary medical education setting*. Paper presented at the Assessment Institute Conference, Indianapolis, IN.

Barr, C. D., Spitmüller, C., & Stuebing, K. K. (2008). Too stressed out to participate? Examining the relation between stressors and survey response behavior. *Journal of Occupational Health Psychology*, *13*, 232–243.

Basow, S. A., & Martin, J. (2012). Bias in student evaluations. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40–49). doi:10.1093/bja/aes563

Bazerman, M. H. & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it*. Princeton, NJ: Princeton University Press.

Benton, S. L., & Cashin, W. E. (2010). Student ratings of teaching: A summary of research and literature. *IDEA Paper #50.*

Beran, T. N., Donnon, T., & Hecker, K. (2012). A review of student evaluation of teaching: Applications to veterinary medical education. *Journal of Veterinary Medical Education*, *39*, 71–78. doi:10.3138/jvme.0311.037R

Bergman, M. M. (2015). Hermeneutic content analysis: Textual and audiovisual analyses within a mixed methods framework. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of mixed methods in social & behavioral research* (pp. 379–396). Thousand Oaks, CA: SAGE Publications.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*, 48–62.

Biesta, G. (2015). Pragmatism and the philosophical foundations of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of mixed methods in social & behavioral research* (pp. 95–118). Thousand Oaks, CA: SAGE Publications.

Bogardus Cortez, M. (2016, December). What is a Scantron: How Scantron technology paved the way for testing tools. *EdTech: Focus on Higher Education*. Retrieved from https://edtechmagazine.com/higher/article/2016/12/scantron-technology-creates-more-efficient-testing

Bolman, L. G., & Gallos, J. V. (2011). *Reframing academic leadership*. San Francisco, CA: Jossey-Bass.

Boston University Center for Teaching & Learning. (n.d.). *Team-teaching: Teaching guide*. Retrieved from http://www.bu.edu/ctl/guides/team-teaching-teaching-guide/

Bunge, N. (2018, November 27). Students evaluating teachers doesn't just hurt teachers. It hurts students. *The Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/article/Students-Evaluating-Teachers/245169

Canaday, S. D., Mendelson, M. A., & Hardin, J. M. (1978). The effect of timing on the validity of student ratings. *Journal of Medical Education*, *53*, 958–964.

Carmack, H. J., & LeFebvre, L. E. (2019). "Walking on eggshells": Traversing the emotional and meaning making processes surrounding hurtful course evaluations. *Communication Education*, *68*, 350–370. doi:10.1080/03634523.2019.1608366

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 25–44). Boston, MA: Anchor Publishing.

Cavanaugh, J. E., & Foster, E. C. (2010). Margin of error. In N. J. Salkind (Ed.), *Encyclopedia of research design*. Retrieved from https://dx-doi-org.proxy.lib.utk.edu/10.4135/9781412961288.n229

Cawunder, P., & Tasker, J. B. (1981). Students ratings of teacher effectiveness: What are the criteria? *Journal of Veterinary Medical Education*, *8*, 21–22.

Chapman, D. D., & Joines, J. A. (2017). Strategies for increasing response rates for online end-of-course evaluations. *International Journal of Teaching and Learning in Higher Education*, *29*, 47–60.

Cho, D., & Cho, J. (2017). Does more accurate knowledge of course grade impact teaching evaluation? *Education Finance and Policy*, *12*, 224–240.

Christopher, M. M., & Leskosky, L. (1997). Effect of faculty gender on veterinary student teaching evaluations. *Journal of Veterinary Medical Education*, *24*, 36–42.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, 16–30.

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*, 52–65.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: SAGE Publications.

Creswell, J. W., Hanson, W. E., Plano Clark, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The Counseling Psychologist*, *35*, 236–264.

DeZure, D. (1999). Evaluating teaching through peer classroom observation. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 70–96). Boston, MA: Anchor Publishing.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*, 1–18.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*, 235–241.

Eckhaus, E., & Davidovitch, N. (2019). How do academic faculty members perceive the effect of teaching surveys completed by students on appointment and promotion processes at academic institutions? A case study. *International Journal of Higher Education*, *8*, 171–180.

El Hassan, K. (2009). Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research & Development*, *28*, 319–333.

Engel, H. N., & Fuqua, R. (1979). Are students biased in their judgement of faculty? *Journal of Veterinary Medical Education*, *6*, 91–96.

Estelami, H. (2015). The effects of survey timing on student evaluation of teaching measures obtained using online surveys. *Journal of Marketing Education*, *37*, 54–64.

Fink, A. (2017). *How to conduct surveys* (6th ed.). Thousand Oaks, CA: Sage Publications.

Fossum, T. W., Ruoff, W. W., Rushton, W. T., & Paprock, K. E. (1993). Patterns of and criteria for evaluating clinical teaching performance: Perceptions of a national sample of teaching veterinary clinicians. *Journal of Veterinary Medical Education*, *20*, 24–27.

Gage, N. L. (1969). A tribute to Hermann Henry Remmers, 1892–1969. *Educational Psychologist*, *6*, 37.

Garcia, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Switzerland: Springer International Publishing.

Getty, S. M. (1975). Student evaluation of teaching: An appraisal and review of recent studies. *Journal of Veterinary Medical Education*, *2*, 25–28.

Gini, A., & Green, R. M. (2013). *10 virtues of outstanding leaders: Leadership and character* (1st ed.). Somerset, NJ: John Wiley & Sons.

Goodboy, A. K. (2011). The development and validation of the instructional dissent scale. *Communication Education*, *60*, 422–440.

Gordon, H., Stevenson, E., Brookhart, A., & Oermann, M. H. (2018). Grade incentive to boost course evaluation response rates. *International Journal of Nursing Education Scholarship*, 20180031.

Greenberg, J., Pyszczynski, T., & Solomon, S. (1982). The self-serving attributional bias: Beyond self-presentation. *Journal of Experimental Social Psychology*, *18*, 56–67.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*, 225–274.

Hammer, R., Peer, E., & Babad, E. (2018). Faculty attitudes about student evaluations and their relations to self-image as teacher. *Social Psychology of Education*, *21*, 517–537.

Hativa, N. (2013). *Student ratings of instruction: Recognizing effective teaching*. Charleston, SC: Oron Publications.

Haunberger, S. (2011). To participate or not to participate: Decision processes related to survey non-response. *Bulletin de Méthodologie Sociologique*, *109*, 39–55.

Higgerson, M. L. (1999). Building a climate conducive to effective teaching evaluation. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 194–212). Boston, MA: Anchor Publishing.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Hodgson, J. L., & Ilkiw, J. E. (2017). Curricular design, review, and reform. In J. L. Hodgson, & J. M. Pelzer (Eds.), *Veterinary medical education: A practice guide* (pp. 3–23). Hoboken, NJ: Wiley Blackwell.

Hoefer, P., Yurkiewicz, J., & Byrne, J. C. (2012). The association between students' evaluation of teaching and grades. *Journal of Innovative Education, 10*, 447–459.

Hofmeister, E. H. (2021). Nonparticipant student observation of faculty classroom teaching. *Journal of Veterinary Medical Education*, *48*, 48–53.

Howell, N. (2002, October 24). Outcomes assessment: College of Veterinary Medicine. *UT Academic Program Review*. Knoxville, TN: University of Tennessee.

Huang, F. L., & Moon, T. R. (2013). What are the odds of that? A primer on understanding logistic regression. *Gifted Child Quarterly*, *57*, 197–204.

Hubbell, J. A. E., & Hudson, W. A. (1995). Development and validation of criteria to evaluate clinical teaching in veterinary medicine. *Journal of Veterinary Medical Education*, *22*, 52–55.

Huck, S. W. (2008). *Reading statistics and research* (5th ed.). Boston, MA: Allyn and Bacon.

IBM Knowledge Center. (n.d.). GLM post hoc comparisons. Retrieved from https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/common/idh_glmu_post.html

Ismail, E. A., Buskist, W., & Groccia, J. E. (2012). In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 79–91). doi:10.1093/bja/aes563

Jhangiani, R., & Tarry, H. (2011). *Principles of social psychology* (1st ed.), Victoria, B.C.: BCcampus. Retrieved from https://opentextbc.ca/socialpsychology/

JMP. (2020). Repeated covariance structures. Retrieved from https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/repeated-covariance-structures.shtml#ww1273585

Keeley, J. W. (2012). Choosing an instrument for student evaluation of instructors. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 13–21). doi:10.1093/bja/aes563

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). New York, NY: Routledge.

Kenney, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*, 422–431.

Kincaid, C. (n.d.). Paper 198-30: Guidelines for selecting the covariance structure in mixed model analysis. *SAS Support*. Retrieved from https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/198-30.pdf

Killinger, S. L., Flanagan, S., Castine, E., & Howard, K. A. S. (2017). Stress and depression among veterinary medical students. *Journal of Veterinary Medical Education*, *44*, 3–8.

Korn, C. W., Rosenblau, G., Rodriguez Buritica, J. M., & Heekeren, H. R. (2016). Performance feedback processing is positively biased as predicted by attribution theory. *PLoS One*, *11*(2), e0148581.

Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluation student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics.* Advance online publication. doi:10.1007/s10805-021-0999-2

LaBelle, S., & Martin, M. M. (2014). Attribution theory in the college classroom: Examining the relationship of student attributions and instructional dissent. *Communication Research Reports*, *31*, 110–116.

Lærd Statistics. (2018). *Types of variable*. Retrieved from https://statistics.laerd.com/statistical-guides/types-of-variable.php

Lee, S., & Lee D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, 71, 353–360.

Leveille, D. E. (2006). *Accountability in higher education: A public agenda for trust and cultural change*. Berkeley, CA: Center for Studies in Higher Education.

Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-Based Dentistry*, *7*, 24–25.

Lombardi, J. V. (2013). *How universities work*. Baltimore, MD: Johns Hopkins University Press.

Maghães-Sant'Ana, M. (2014). Ethics teaching in European veterinary schools: A qualitative case study. *Veterinary Record*, *175*, 592. doi:10.1136/vr.102553

Malone, E. D., Root Kustritz, M. V., & Molgaard, L. K. (2018). Improving response rates for course and instructor evaluations using a global approach. *Education in the Health Professions*, *1*, 7–10.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 45–69). Bolton, MA: Anker Publishing Company.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187–1197.

McAlpin, V., Algozzine, M., Norris, L., Hartshorne, R., Lambert, R., & Algozzine, B. (2014). A comparison of web-based and paper-based course evaluations. *Journal of Academic Administration in Higher Education*, *10*, 49–58.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*, 267–277.

McCarthy, M. A. (2012). Using student feedback as *one* measure of faculty teaching effectiveness. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 30–39). doi:10.1093/bja/aes563

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation & Research in Education*, *24*, 183–202.

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, *23*, 143–149.

McKeachie, W. J. (1969). Student ratings of faculty. *American Association of University Professors Bulletin*, *55*, 439–444.

McNulty, J. A., Gruener, G., Chandrasekhar, A., Espiritu, B., Hoyt, A., & Ensminger, D. (2010). Are online student evaluations of faculty influenced by the timing of evaluations? *Advances in Physiology Education*, *34*, 213–216.

Medical Schools Objectives Project Writing Group. (1999). Learning objectives for medical student education–guidelines for medical schools: Report 1 of the medical schools objectives project. *Academic Medicine, 74*, 13–18.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Morrow, J. A., & Skolits, G. (2017, November). *Twelve steps of quantitative data cleaning: Strategies for dealing with dirty evaluation data*. Session presented at the meeting of the American Evaluation Association, Washington, DC.

Morse, J. (2015). Procedures and practice of mixed method design: Maintaining control, rigor, and complexity. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of mixed methods in social & behavioral research* (pp. 339–352). Thousand Oaks, CA: SAGE Publications.

Mosier, D. A., Cashin, W. E., & Elmore, R. G. (1993). Administrative evaluation of teaching in veterinary medical colleges. *Journal of Veterinary Medical Education*, *20*, 40–44.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, *48*, 630–635.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*, 187–198.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*, 301–314.

Onwuegbuzie, A. J., & Combs, J. P. (2015). Emergent data analysis techniques in mixed methods research: A synthesis. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of*

mixed methods in social & behavioral research (pp. 397–430). Thousand Oaks, CA: SAGE Publications.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*, 113–160.

Pearson, C. S. (2012). *The transforming leader: New approaches to leadership for the twenty-first century*. Oakland, CA: Berrett-Koehler Publishers.

Plano Clark, V. L., & Badiee, M. (2015). Research questions in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of mixed methods in social & behavioral research* (pp. 275–304). Thousand Oaks, CA: SAGE Publications.

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement, and personality. *Research in Higher Education*, 46, 127–152.

Reisbig, A. M. J., Danielson, J. A., Wu, T.-F., Hafen, Jr., M., Krienert, A., Girard, D., & Garlock, J. (2012). A study of depression and anxiety, general health, and academic performance in three cohorts of veterinary medical students across the first three semesters of veterinary school. *Journal of Veterinary Medical Education*, *39*, 341–358.

Reisenwitz, T. H. (2016). Student evaluation of teaching: An investigation of nonresponse bias in an online content. *Journal of Marketing Education*, *38*, 7–17.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*, *30*, 387–415.

Robinson, J. (2017). Exploring attribution theory and bias. *Communication Teacher*, *31*, 209–2013.

Robinson, S. B., & Leonard, K. F. (2019). *Designing quality survey questions*. Thousand Oaks, CA: SAGE Publications.

Royal, K. (2015). *Margin of error calculator*. North Carolina State University College of Veterinary Medicine. Retrieved from https://legacy.cvm.ncsu.edu/c/l/online_calculator/workspace/index.html

Royal, K. (2017). A guide for making valid interpretations of student evaluation of teaching (SET) results. *Journal of Veterinary Medicine Education, 44*, 316–322.

Royse, D., Thyer, B. A., & Padgett, D. K. (2016). *Program evaluation: An introduction to an evidence-based approach* (6th ed.). Boston, MA: Cengage Learning.

Ruoff, W. W., Fossum, T. W., Rushton, W. T., & Paprock, K. E. (1992). Developing criteria for evaluation of teaching performance in the veterinary teaching hospital. *Journal of Veterinary Medical Education*, *19*, 102–107.

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Schafer, P., Hammer, E. Y., & Berntsen, J. (2012). Using course portfolios to assess and improve teaching. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 71–78). doi:10.1093/bja/aes563

Schiekirka, S., Reinhardt, D., Heim, S., Fabry, G., Pukrop, T., Anders, S., & Raupach, T. (2012). Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. *BMC Medical Education*, *12*, 45.

Seldin, P. (1999a) Current practices–good and bad–nationally. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 1–24). Boston, MA: Anchor Publishing.

Seldin, P. (1999b) Self-evaluation: What works? What doesn't? In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 97–115). Boston, MA: Anchor Publishing.

Shapiro, J. P., & Stefkovich, J. A. (2005). *Ethical leadership and decision making in education* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Sharma, G. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research*, *3*, 749–752.

Sharpe, D. (2015). Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation*, *20*, 8.

Smalzried, N. T., & Remmers, H. H. (1943). A factor analysis of the Purdue Rating Scale for Instructors. *The Journal of Educational Psychology*, *34*, 363–367.

Stalmeijer, R. E., Dolmans, D. H. J. M., Wolfhagen, I. H. A. P., Muijtjens, A. M. M., & Scherpbier, A. J. J. A. (2010). The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. *Academic Medicine*, *85*, 1732–1738.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Upper Saddle River, NJ: Pearson.

Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology*, *18*, 68–88.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, *109*, 45–56.

The Ohio State University College of Veterinary Medicine. (n.d.). *Faculty student evaluation of instruction survey course request*. Retrieved from https://vet.osu.edu/faculty-student-evaluation-instruction-survey-course-request

Toma, B., Begon, D., Fontain, J.-J. Rosenberg, D. (2004). Informative evaluation of the teaching skills of the faculty at Alfort Veterinary School. *Journal of Veterinary Medical Education*, *31*, 45–54.

UCLA Statistical Consulting Group. (n.d.). Introduction to linear mixed models. Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

Uttl, B., White, C. A., Wong Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42.

van den Berg, R. G. (2020). *SPSS independent samples t-test tutorial*. SPSS Tutorials. Retrieved from https://www.spss-tutorials.com/spss-independent-samples-t-test/comment-page-10/

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*, 191–211.

Warman, S. M. (2015). Challenges and issues in the evaluation of teaching quality: How does it affect teachers' professional practice? A UK perspective. *Journal of Veterinary Medical Education*, *42*, 245–251.

Weary, G. (1979). Self-serving attributional biases: Perceptual or response distortions? *Journal of Personality and Social Psychology*, *37*, 1418–1420.

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, *71*, 3–25.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York, NY: Springer-Verlag.

Weiner, B. (2010). Attribution theory. *International Encyclopedia of Education*, *6*, 558–563.

Wilson, J. H., & Ryan, R. G. (2012). In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 22–29). doi:10.1093/bja/aes563

Wu, H., & Leung, S.-O. (2017). Can Likert scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, *43*, 527–532.

Yin, R. K. (2013). How to do better case studies: (With illustrations from 20 exemplary case studies). In L. Bickman, & D. J. Rog (Eds.), *The SAGE handbook of applied social research methods* (pp. 254–282). Thousand Oaks, CA: SAGE Publications.

Young, K., Joines, J., Standish, T., & Gallagher, V. (2019). Student evaluations of teaching: The impact of faculty procedures on response rates. *Assessment & Evaluation in Higher Education*, *44*, 37–49.

Zenner, D., Burns, G. A., Ruby, K. L., DeBowes, R. M., & Stoll, S. K. (2005). Veterinary students as elite performers: Preliminary insights. *Journal of Veterinary Medical Education*, *32*, 242–248.

Zubizarreta, J. (1999) Evaluating teaching through portfolios. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practice guide to improved faculty performance and promotion/tenure decisions* (pp. 162–182). Boston, MA: Anchor Publishing.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*, 245–287.

VITA

Misty R. Bailey is from Washburn, TN, and holds B.A. and M.A. degrees in English from the University of Tennessee, Knoxville. She has experience as an instructor of English composition, literature, and English as a second language (China). Her higher education experience includes 10 years of practice as a science editor in veterinary and biomedical sciences, and 5 years of experience in veterinary medical education assessment. Bailey has published several articles in the field of veterinary medical education and is a peer reviewer for the *Journal of Veterinary Medical Education*. She also has had several veterinary medical education grants funded as either a principal investigator or co-investigator. Bailey is a member of the Phi Zeta Veterinary Honor Society.