EXAMINING INSTRUCTOR AND INSTRUCTIONAL EFFECTS ON STUDENTS' STATISTICS ATTITUDES

by

Chao Xu, MS

DISSERTATION

Presented to the Faculty of

The University of Houston-Clear Lake

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF EDUCATION

in Curriculum and Instruction

THE UNIVERSITY OF HOUSTON-CLEAR LAKE

DECEMBER, 2019

EXAMINING INSTRUCTOR AND INSTRUCTIONAL EFFECTS ON STUDENTS' STATISTICS ATTITUDES

by

Chao Xu

APPROVED BY

____
Suzanne Brown, PhD, Chair

____
Michelle Peters, EdD, Committee Member

____
Amy Orange, PhD, Committee Member

____
Debby Shulsky, EdD, Committee Member

RECEIVED/APPROVED BY THE COLLEGE OF EDUCATION:

____
Lillian Benavente-McEnery, EdD, Interim Associate Dean

____
Joan Y. Pedro, PhD, Interim Dean

Dedication

I dedicate this dissertation to my father Guogen Xu (许国根), my mother Yuping Zhu (朱玉萍), my wife Xiang Deng (邓湘), and my son Yifan Xu (许屹凡). Their unyielding love has inspired me to complete this research.

Acknowledgements

A thank you must first go to Dr. Suzanne Brown, the chair of my dissertation committee. Dr. Brown provided her expertise as well as support all along the way so I continued to believe in myself that I could accomplish this journey. Not only did Dr. Brown help me spear ahead my research efforts by keeping me on the right track, more importantly, communications with her made me feel like very much a part of this program and a part of this department. A big thank you goes to Dr. Michelle Peters, whose expertise in hierarchical linear model literally led me in to the fantastic world of educational research. I am grateful for her methodological support every time I was in need. I must also thank Dr. Amy Orange and Dr. Debby Shulsky for their expertise and assistance throughout this process. In all seriousness, I never knew I was ever capable of interviewing people until I took Dr. Orange's class. Also, I must thank Dr. Brenda Weiser, the EDCI program director, for walking me through the whole process.

A thank you must go to Dr. Candace Schau. Two years ago, I emailed her with a request for some data to begin with. To my amazement, Candace immediately sent me a huge set of data she and her colleagues had been collecting for years. This went to show what a spirit of open science truly is. For a moment, I felt I was the luckiest person in the world. I am deeply indebted to Candace for the great mentorship she was and still is providing. And I will always cherish the friendship we have developed through our collaborations. Most importantly, I wish her and Tom good health.

Last but certainly not the least, there is family to thank. A thank you must first go to my wife for her continual support and particularly for her not giving up on me when I used to be such a huge disappointment. She is a loving yet tough woman. I admire her greatly. In addition, I must admit that her PhD literally oriented me on the enhancement of my motivation to get my own doctorate. I also want to thank my son for getting me through the tough times by keeping me up and running. My life eventually began to have a purpose the time he was born. To my parents, words are just not enough to express my thankfulness. I love them evermore with all my heart!

ABSTRACT

EXAMINING INSTRUCTOR AND INSTRUCTIONAL EFFECTS ON STUDENTS' STATISTICS ATTITUDES

Chao Xu
University of Houston-Clear Lake, 2019

Dissertation Chair: Suzanne Brown, PhD

A long-standing hypothesis in statistics education has posited that instructors could have a large impact on students' attitudes toward statistics. This hypothesis remained untested over the years. Moreover, if instructors do have a large impact on students' statistics attitudes, then there is a need to provide explanations as to what dimensions of teaching competencies account for this impact. Drawing on a rich data set collected from 1,924 students clustered within 23 instructors across 11 post-secondary institutions in the United States, the hypothesis concerning instructor effects on students' statistics attitudes was tested in this study. Multilevel covariate adjustment models were employed to quantify the size of instructor effects. The analysis suggested that instructors varied considerably in their abilities to improve students' statistics attitudes. Furthermore, instructors' differential contributions to students' attitudes were found to be positively

associated with instructional practices most proximal to tasks involving data collection and analysis in proper contexts as well as with instructors' attitudes toward teaching statistics classes. Lastly, results showed that instructors who improved students' statistics attitudes were also effective at improving their expected course grades, a measure that strongly predicts student ratings of teaching. The teaching effectiveness measures explored in this study may be used to orient instructors on the development of new pedagogic skills centered on students' statistics attitudes. Altogether, these findings necessitate the need of future studies to identify and validate additional instructional dimensions that hold promise for improving students' statistics attitudes and also herald an exciting opportunity to expose students to a full range of instructional skills in the modern statistics classrooms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:
INTRODUCTION

Statistics education is becoming an essential component of higher education (Cobb, 2015). In the United States (U.S.), the number of students enrolled in college introductory statistics courses is rapidly growing and will continue to grow because, in the age of information, conceptual understanding of statistics is crucial in everyday life and data-analytical skills are becoming indispensable in many professions. In reality, most college students take only one introductory statistics course and, as a result, have very limited access to formal statistics instruction. Clearly, effective teaching practices are needed to support students' development in areas beyond their core academic skills over such a short span of time. These areas include students' attitudes toward statistics, which have recently emerged as an important course outcome (Ramirez, Schau, & Emmioğlu, 2012). Yet while researchers have hypothesized that instructors could have a large impact on students' statistics attitudes (Schau, 2003), there existed very little empirical evidence regarding whether and, to what extent, instructors actually contribute to this learning outcome. This study was intended to fill this gap.

Research Problem

Statistics is gaining prominence at all three levels of education systems in the U.S. as well as in many other countries around the world (Cobb, 2015; De Veaux et al., 2017; Franklin, 2013; Usiskin & Hall, 2015). There are at least two reasons for this occurrence. On one hand, statistics as a discipline has evolved over the years. A few decades ago, many of the statistics problems found in everyday life could only be accessed by experts who were proficient in theory of statistics and probability. Today, the same set of problems can be handled easily by novices with an armamentarium of computational tools that are available to them. On the other hand, opportunities are abundant and everywhere in industry, academia and commerce for new graduates who combine at least some statistical skills with the domain knowledge gained from studying for their degrees. Almost every organization is now looking to bring in new hires with such skill set to handle the ever-increasing mass of data.

Historically, statistics was viewed as a branch of mathematical sciences primarily because this discipline takes root in probability (Cobb & Moore, 1997). Students resorted to probability distribution as a formal way to test hypothesis and make statistical inferences. Computation involved in solving statistics problems was procedural and typically laborious yet needed to be carried out with paper-and-pencil algebraic skills. This convention conveniently landed the study of statistics into school mathematics curricula (Usiskin & Hall, 2015). There was nothing particularly inventive about teaching these two disciplines together. Statistics educators also seemed quite content with this placement because mathematics is always the second most important school subjects, right behind the language art. It is tested everywhere and hence taught everywhere.

Students are living through an age of technological upheaval that has transformed the way in which knowledge is transferred in classrooms. Not surprisingly, the paradigm of teaching statistics has also shifted drastically toward a new one centered around modern data science concepts (De Veaux et al., 2017). Many of those major changes in statistics curricula stemmed from a widespread use of high-performance computers and a growing awareness for uncoupling statistics from mathematics (Cobb & Moore, 1997). The Common Core State Standards Initiative is the latest in a trail of nationwide efforts at revising teaching guidelines and learning objectives in this direction (Franklin, 2013).

The separation of statistics from school mathematics curricula was prompted by reconsideration of the relationship between the two disciplines. Students must have gained at least some mathematical understanding prior to studying statistics. However, statistics, just like economics, physics and engineering, has its own subject matter. Following this line of thought, statistics may then be taught as an independent school subject, not within mathematics nor as applied mathematics (Usiskin & Hall, 2015). Rather, mathematics should be thought of as a means to an end in the process of learning statistics rather than as an end in itself. Moreover, statistics works admirably with data. Due to the pivotal role that data contexts play, learning statistics demands a nonmathematical thought process. As Cobb and Moore (1997) contended, statistical thinking inevitably relates to the admonition that dwells on "data are not just numbers, they are numbers with a context" (p. 801).

In their landmark essay, expository of many transition points in statistics education at the turn of the century, Wild and Pfannkuch (1999) analyzed an array of nonmathematical thought processes that statisticians used in statistical inquiry. The inquiry process lies at the heart of statistical practice of professionals and has yet to be encultured into the ways in which students think and reason about statistics (Cobb, 2015). Wild and Pfannkuch further synthesized the types of nonmathematical thought processes into a four-dimensional "statistical thinking" framework – investigative cycle, interrogative cycle, general types of thinking, and dispositions.

Much of what has been done in Wild and Pfannkuch (1999) conforms to generic approaches to problem analysis and solving. For instance, the "statistical thinking" framework bears a striking resemblance to the "Learning by Design" cycles that have been intensely popularized in K-12 science and engineering classrooms (Brophy, Klein, Portsmore, & Rogers, 2008). The frequent crosstalk between different learning areas opens up the possibility that statistics can be taught in nonmathematical and even qualitative fashion (Ograjenšek & Gal, 2016), thus strongly suggesting the plausibility of teaching statistics across school curriculum (Usiskin & Hall, 2015).

The precise delineation and demarcation of the term “statistical thinking” laid the groundwork for the field of statistics education research by providing a fresh perspective from which to devise student learning outcomes (Cobb, 2015). Statistics education researchers, as those in any other branch of education research, view improving instruction and eventually student learning as one of the ultimate goals of teaching. Ideally, the ways in which statistics are taught are meant to spark a pursuit of knowledge in the students to such a degree that they continue to learn and use statistics outside the classroom and even after they leave introductory statistics courses. Accordingly, research in this domain is anticipated prompting strong education reform trends, whose purpose is to engender statistical literacy among all students, via evidence-based practice.

Statistics education research in the early era fell short of this ideal, as empirical studies have focused predominantly on the development of instructional approaches and assessment methods that only benefited student learning of basic statistical knowledge (Zieffler et al., 2008). However, statistical literacy goes well beyond knowledge base and also includes a set of dispositions (i.e., attitudes, beliefs, and critical stance) and statistical reasoning (Gal, 2002; Tobías-Lara & Gómez-Blancarte, 2019). Garfield, Hogg, Schau, and Whittinghill (2002) contended that students’ statistics attitudes are at least as important as their achievement in the first statistics courses. The authors postulated positive statistics attitudes as being crucial in enhancing learners’ ability to solve statistics problems in real-life situations, as are positive academic attitudes that were frequently found to prompt students to use the knowledge gained in classes (Pianta & Hamre, 2009). Even when students forget the basic knowledge, good attitudes likely still remain and keep them re-learning and perhaps learning more.

In spite of the importance, students’ statistics attitudes are seldom assessed as a learning outcome as opposed to statistics achievement and statistical reasoning.

Instructors usually pay little, if any, attention to this dimension of student learning. Efforts to promote and improve students' statistics attitudes are hence lackluster. Recent studies have found that introductory statistics courses in the U.S. did not improve students' attitudes toward statistics, at least on average (Schau & Emmioğlu, 2012). In some cases, students' attitudes even changed in the negative direction although the decrease over time was not monotonic (Kerby & Wroughton, 2017). The reason for this decrease is still largely speculative since little is known about what factors actually contribute to improvements in students' statistics attitudes (Schau & Emmioğlu, 2012), though statistics instructor and instruction appeared to be one determinant (Schau, 2003).

Education economists regard teacher quality as the most important input to both the cognitive and non-cognitive returns to public education (e.g., Hanushek, 2003, 2011; Hanushek & Rivkin, 2010). Individual differences in teaching effectiveness are generally considered as one of several main determinants of student success in academic performance. For instance, Odden, Borman, and Fermanich (2004) found that the size of estimated teacher effects on students' academic achievement was considerably larger than the effect sizes produced by almost all other educational inputs such as school resources, teaching materials supplies, class size reduction and so forth. Similarly, Nye, Konstantopoulos, and Hedges (2004) also found that effective teachers might be a more important factor contributing to student outcomes than schools. Notably, both studies were less successful in identifying teacher and teaching attributes that could predict teacher effects that were large enough to be deemed important.

Unlike their academic achievement, students' academic attitudes are stable and rather resistant to change even in the presence of strong instructional interventions (Cohen, Garcia, & Goyer, 2017). The difficulty with estimating teacher effects on students' academic attitudes may be due to this recalcitrance. If the magnitudes of changes associated with teachers in scores on the outcome data are small, so is the variance in the measurements across teachers. Technically, it is commonly referred to as small between-teacher variation; if this occurs, there would not be enough teacher-specific variation to be productively modelled (Bauer, 2003).

Recent direct investigations of the teacher's contributions to students' non-cognitive skills shed light on the dimensions of teaching practices that hold promise for improving students' academic attitudes. Using exploratory factor analysis, several studies have identified that attitudes-improving classrooms required teachers to frequently interact with students, to precisely deliver the content and to provide both emotional and intellectual supports (Blazar & Kraft, 2017). These teaching competencies are mathematics-specific with strong connections to students' non-cognitive competencies even beyond their academic attitudes. Therefore, statistics educators who are interested in improving students' attitudes may draw on these findings.

Considerable effort has been devoted in recent years to understanding the interactions between statistics instruction and student learning of statistics. Such effort has resulted in a set of guidelines for assessment and instruction in statistics classrooms (GAISE College Report ASA Revision Committee, 2016). However, the literature of statistics education is still noted for its relative inattention to students' attitudes, a non-cognitive outcome that has begun to receive much attention in current research on K-12 mathematics education.

Findings from at least two earlier studies have hinted at the possibility that instructors could have a non-negligible impact on students' statistics attitudes (Schau, 2003; Waters, Martelli, Zakrajsek, & Popovich, 1988). However, it is difficult to determine to what degree the empirical evidence from these two studies supports this claim. Few, if any, works have followed up on the findings from these two studies. As a result, the hypothesis remained untested concerning whether and to what extent instructors can have an impact on students' statistics attitudes. This problem is intractable unless a large-scale data set becomes available for which there exist students linked with a large number of statistics instructors.

In conclusion, with such data set already being available (Schau & Emmioğlu, 2012), this hypothesis needs to be put to test. If instructor effects turn out to be large on students' statistics attitudes, it is then possible to model the effects productively as a function of instructor and instructional characteristics. This approach allows for identification of instructor-specific characteristics that are potentially associated with improvements in students' statistics attitudes. Thus, this line of research is of both theoretical and practical interest to education researchers and statistics instructors.

Significance of the Study

This study represents a unique contribution to the field of statistics education research. The contributions are multifaceted. First, value-added methodologies are scarcely applied to analysis of student outcome data in higher education despite a mountain of empirical evidence related to how much student learning varies across teachers in K-12 education. With growing recognition that measuring student achievement is essential for improving the quality of higher education, teacher value-added measures have recently begun to receive attention in the tertiary settings (e.g., Carrell & West, 2010; Chingos, 2016; Liu, 2011; Galbraith, Merrill, & Kline, 2012; Shavelson et al., 2016; Steedle, 2012). Those measures can be used to inform departments about the ways to improve instructor quality and therefore increase enrollment. This work added to the previous studies by exploring practices, issues and challenges of measuring instructor effects on college students' academic attitudes.

Second, students' academic attitudes have emerged as one primary dimension of student learning, and were shown to serve as a strong predictor of student long-term outcomes such as educational attainment, adult earnings and health status (Chetty et al., 2011; Moffitt et al., 2011). More interestingly, Chetty and his colleagues found that the predictive power associated with students' academic attitudes on those long-term outcomes was substantially larger than that with their test scores. As a result, the findings of this study on students' statistics attitudes contributed to the rapidly growing interest in including measures of academic attitudes in assessment of student success as well as evaluation of teacher quality.

Third, the teacher effectiveness literature has focused predominantly on students' academic performance on high-stakes standardized tests (e.g., Hanushek & Rivkin, 2010; Rockoff, 2004). However, high-quality teachers are expected not only to raise students' test scores but also to improve their academic attitudes (Blazar & Kraft, 2017). The present study examined the relationships between students' statistics attitudes and certain dimensions of instructional characteristics. Thus, it added to decades worth of multidimensional teaching theory by expanding the focus to include instructors' abilities to improve students' statistics attitudes.

Fourth, it is noted that all studies regarding teacher effects on students' non-cognitive competencies were conducted in pre-kindergarten or K-12 settings (Blazar, 2018; Blazar & Kraft, 2017; Cheng & Zamarro, 2018; Chetty et al., 2011; Jennings & DiPrete, 2010; Kraft, 2019; Ladd & Sorensen, 2017; Ruzek, Domina, Conley, Duncan, & Karabenick, 2015). However, according to most theories of adult development, affective and behavioral competencies are also malleable at later ages (Cohen et al., 2017; Rogaten et al., 2019). Whether and to what extent those findings in early childhood settings can generalize to higher education remained unclear. This study extended the body of research conducted only in pre-kindergarten and K-12 settings by examining the impact of teachers on additional attitudes captured by college students.

Finally, little evidence existed on the types of instructional approaches that might contribute to development of students' statistics attitudes (Carlson & Winquist, 2011; Ramirez et al., 2012; Schau & Emmioğlu, 2012). This study filled this gap by first estimating instructor effects on students' statistics attitudes and then identifying certain dimensions of instructional practices that potentially held promise for improving students' statistics attitudes. The findings of this study may enable statistics educators to design and implement timely as well as targeted course interventions tailored for college students who take introductory statistics courses. This type of course interventions is considered particularly important to this population of students because they typically have very little exposure to formal statistics instruction (Schau, 2003).

Research Purpose and Questions

The purpose of this study was to examine instructor and instructional effects on students' statistics attitudes. A conceptual overview of the present study is diagrammed in Figure 1.1, which visualizes the following seven research questions that guide this study:

1. Does taking introductory statistics courses change students' statistics attitudes?
2. Is there an unexplained variation across statistics instructors regarding students' statistics attitudes?
3. Do instructors' attitudes toward teaching predict students' statistics attitudes?
4. Does instructors' emphasis on data contexts predict students' statistics attitudes?

5. Does instructors' emphasis on assessment predict students' statistics attitudes?
6. Does instructors' emphasis on conceptual understanding predict students' statistics attitudes?
7. Do instructors who improve students' statistics attitudes also improve expected course grades?

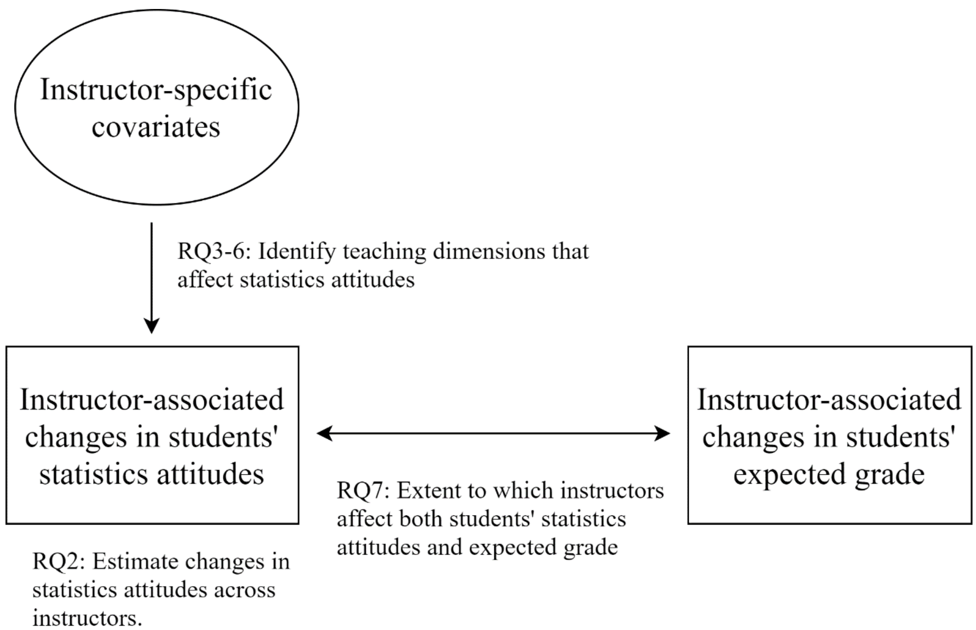

Figure 1.1 Conceptual Overview of the Present Study

Definitions of Key Terms

College Introductory Statistics Courses: A growing number of college students are enrolled in introductory statistics courses either with no mathematics prerequisite or with a college algebra-only prerequisite. These courses are usually mandatory for students with a variety of majors such as biology, chemistry, environmental sciences, engineering, psychology, medicine and so forth. Students come to the course with various quantitative backgrounds and different levels of mathematics preparation. This type of courses is typically taught in mathematics and statistics departments and are thus also called service courses (Schau & Emmioğlu, 2012).

Students' Statistics Attitudes: This psychological trait broadly refers to a set of dispositions, emotions and beliefs toward learning and using statistics, both in and outside the statistics classrooms (Schau, 2003).

Affect: Students' positive or negative feelings toward statistics (Schau, 2003).

Cognitive Competence: Students' attitudes about intellectual knowledge and skills when applied to statistics (Schau, 2003).

Value: Students' attitudes about the usefulness, relevance, and worth of statistics in their personal and professional life (Schau, 2003).

Difficulty: Students' perceptions about how difficult statistics is as a school subject (Schau, 2003).

Interest: Students' level of individual interest in statistics (Schau, 2003).

Effort: Amount of effort the student expends to learn statistics (Schau, 2003).

Teacher value-added: Value-added refers to a teacher's contribution to the outcome scores of their students during a particular period of time (e.g., a school year). Importantly, the contribution needs to be comparable to the same performance measures of other teachers (Amrein-Beardsley & Holloway, 2019).

Value-added model: Value-added model typically assumes a two-level structure in which students are clustered into individual teachers to form the first level while teachers constitute the higher level. At its simplest, student outcome of interest at the end of a period is regressed on the same outcome at the beginning of the period with teachers being modelled as random effects assumed to arise from the same probability distribution. Further correction is made to include student characteristics and other factors that are beyond teacher's control (Kraft, 2019).

Instructor effects: Instructor effects refer to the magnitudes of impact that instructors have on student learning. The effects are typically operationalized either as the percentage of variation that teachers contribute to the total variation in learning outcome measures (Nye et al., 2004) or as the standard deviation of the instructor-level variance component (Kraft, 2019).

Instructor effectiveness: Instructor effectiveness refers to a distribution of instructor effects on the student outcome of interest. This distribution enables the researchers to compare average performances of a teacher's students to predictions of these students' performances had they been assigned to an average teacher. It also enables the schools to identify teachers who are effective at improving the student outcome of interest. As a result, educational policymakers see measures of teacher effectiveness as a part of outcome-based accountability or an improved approach to objectively evaluating teacher quality (Blazar & Kraft, 2017).

Instructional effects: Instructional effects refer to the underlying relationships between certain dimensions of instructional practice and instructor-associated changes in student outcomes of interest. The relationships are only suggestively rather than conclusively causal despite the use of causal word "effects" (Blazar & Kraft, 2017).

Conclusion

This chapter identified the need for this study and provided an overview of the research problem, significance of the problem, research purpose and questions, and definitions of key terms pertinent to this study. The next chapter will present a review of the current literature with the focus on the following three topics – the relationships between students' statistics attitudes and their characteristics, students' statistics attitudes as a learning outcome, and teacher and teaching effects on students' non-cognitive competencies in general.

CHAPTER II:

LITERATURE REVIEW

Statistics is considered as a service to the welfare of a modern society in the age of information (Cobb, 2015). In this spirit, researchers from different disciplines collaborate actively on various topics in a collective effort to improve the teaching and learning of statistics. One major challenge facing the statistics education community is to define course outcomes that are useful to predicting students' willingness to learn and use statistics. Clearly identified course outcomes in this field are statistics achievement, reasoning and attitudes. As with any other learning outcomes, students' statistics attitudes are impacted by many educational variables. Some commonly researched variables include (a) student characteristics such as gender, age and prior experiences with mathematics; (b) other course outcomes; and (c) teacher and teaching attributes (Ramirez et al., 2012; Schau, 2003). The first two sections of this chapter reviewed the relationships among students' statistics attitudes and various educational inputs and outputs. The third section focused on instructor and instructional effects on students' academic attitudes and behaviors in general since this line of research is noticeably lacking in the field of statistics education research.

Statistics Attitudes and Student Characteristics

Student Demographics

To date, evidence is mixed on the impact of student demographics on statistics attitudes. Coetzee and van der Merwe (2010) found that student age appeared to influence students' attitudes toward statistics. Students' attitudes were measured by the *Survey of Attitudes Toward Statistics* (SATS) (Schau, 2003). The SATS-28 was administered to a sample of 235 students enrolled in an introductory psychology statistics course at a large university in South Africa. Of note were the constituents of the sample surveyed. A large percentage of students in this class were not typical university freshman students. The average age of the participants was 31 years. The researchers examined mean differences in subscale scores. The results suggested that most of the students perceived statistics to be very difficult yet worthy of investing reasonable amount of efforts due to its utilitarian value. The average *Difficulty* score was conversely correlated to age, indicating that younger students perceived statistics to be more difficult. However, Cashin and Elmore (2005) reported an opposite finding with a sample of 342 students whose average age was slightly over 29 years.

Due to the small sample sizes and idiosyncratic locations (i.e., a large South African university versus a medium-sized university in the Midwest) in either study, it is unclear which finding is more generalizable to other student populations, particularly for the population in the U.S. In fact, age may influence students' statistics attitudes in a considerably more complicated manner than it appeared in either study because there are too many factors that potentially confound age. For instance, Onwuegbuzie (2004) reported that graduate students tended to have higher statistics anxiety/more negative attitudes toward statistics than undergraduate students. This anxiety was highly associated with students' academic procrastination due to their fear of failure.

Evidence from studies on the impact of gender on students' statistics attitudes remains inconclusive, albeit with a very interesting pattern. For all studies completed outside of the U.S., statistically significant gender differences were found on some of the six attitudinal components (e.g., Coetzee & van der Merwe, 2010). Namely, male students reported significantly higher average scores on *Affect*, *Difficulty*, and *Cognitive Competence* while female students were more willing to invest more effort to study statistics. On the other hand, earlier studies conducted in the U.S. concluded that male and female students did not differ in attitudinal scores, at least on average (Carnell, 2008; Cashin & Elmore, 2005).

In a recent study, van Es and Weaver (2018) re-examined the relationship between gender and students' statistics attitudes using a total of 611 students enrolled in introductory statistics courses offered by Cornell University's Dyson School of Applied Economics and Management. The percentages of male and female students in this sample were approximately balanced (51% and 49%). The results from multivariate analysis of variance (MANOVA) corroborated the findings of the studies from outside of the U.S.; that is, female students reported significantly lower average scores than males in three attitude subscales – *Affect*, *Cognitive Competence*, and *Difficulty*.

van Es and Weaver (2018) did not find significant differences in attitudinal scores between racial/ethnic groups. However, Latino and African American students reported the largest difference between final and expected grades while Asian students' final grades generally met the self-expectation. The mismatch between students' statistics attitudes and expected grades seemed to somewhat contradict theories of students' motivation. The misalignment between theories and empirical evidence calls for additional studies.

Prior Achievement in Mathematics

Cashin and Elmore (2005) examined whether students' prior experiences with mathematics predicted students' statistics attitudes. The sample consisted of 342 students enrolled in two elementary statistics courses at a Midwestern university of medium size. Students were from two different classes; that is, 116 students were enrolled in an upper-level undergraduate course and 226 were in a graduate course. Survey data were collected by the administration of both pre- and post-versions of the SATS-28 (Schau, Stevens, Dauphinee, & Vecchio, 1995). Previous achievement in mathematics was measured using the number of previous mathematics courses that the students had taken prior to their participation in the research project. Data from two samples were then pooled together for the subsequent multiple regression analyses. Scores on all four subscales – *Difficulty*, *Effort*, *Cognitive Competence* and *Affect* – of the SATS-28 obtained at the pretest sessions were found more or less positively correlated with previous experiences with mathematics. In addition, the researchers also suggested a combination of *Difficulty*, *Cognitive Competence* and *Affect* subscales. This finding, in line with what was suggested in Vanhoof, Kuppens, Sotos, Verschaffel, and Onghena (2011), implied that there might exist additional constructs in one or more of the following scenarios: an overall attitude construct (Gignac, 2008; Rodriguez, Reise, & Haviland, 2016), multilevel constructs of students' attitudes (Geldhof, Preacher, & Zyphur, 2014; Stapleton, Yang, & Hancock, 2016), and undesired constructs due to method effects (Maul, 2013; Spector, Rosen, Richardson, Williams, & Johnson, 2019). Addition construct-centric approaches are needed to tease apart the measure of students' statistics attitudes by the SATS-36 (Xu & Schau, 2019).

Dempster and McCorry (2009) examined the nature of the relationships between previous academic experiences and attitudes toward statistics among undergraduate psychology students. Statistics courses are typically an integral part of the college psychology curriculum. However, psychology students often think statistics courses are dreadful and therefore many of them hold very negative attitudes toward statistics. The researchers administered SATS-28 to the students at the beginning of the first year and again at the end of the second year, a point where the students finished all the required statistics courses. Eighty-two students completed tests at both time points. Of these participants, 78% were female. This percentage accurately reflected the composition of male and female students whose major was psychology in many countries. In addition, previous experiences with mathematics were also assessed by using the additional items appended to the SATS-28.

Multilevel modeling approach was employed to investigate the nature of the relationship between students' statistics attitudes and prior achievement in mathematics. Initial findings suggested that *Cognitive Competence* at the first session served as a factor that mediated the associations between self-reported experiences with mathematics and *Cognitive Competence* at the second session. The mediating effect due to *Cognitive Componence* at the first session was further tested using regression models and a specialized *t* test (i.e., Sobel test). Having identified *Cognitive Competence* as a mediating factor, statistics education researchers might be able to design course interventions that target a known attitude component.

Academic attitudes, or non-cognitive competencies in general, are latent psychological traits. Therefore, these types of psychological traits are measured with considerable uncertainty. Due to the reliability issue, latent variable modeling techniques are commonly seen in research involving students' academic attitudes and behaviors. Chiesi and Primi (2010) used structural equation modeling (SEM), a common latent variable modeling technique (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005; Muthén, 2002; Rosseel, 2012), to test the hypothesis that course achievement could be a function of both mathematics background and students' statistics attitudes. The sample consisted of 487 undergraduate psychology students enrolled in one introductory statistics course at the University of Florence in Italy.

The results suggested that previous experiences with mathematics had an indirect impact on pre-course attitudes via the amount of mathematics knowledge students had mastered prior to the enrollment. Particularly, the extent to which *Affect* and *Difficulty* increased depended on previous experiences with mathematics. In addition, *Cognitive*

Competence was found independent of previous experiences with mathematics. This last finding leads back to the one previously reported in Dempster and McCorry (2009). Therefore, it is interesting to explore how *Cognitive Competence* mediates the associations between post-course mathematics knowledge and other components of students' statistics attitudes.

Many theoretical frameworks have attempted to explain the relationship between past achievement in quantitative disciplines and students' statistics attitudes. Hood, Creed, and Neumann (2012) conducted a study to *a posteriori* align the six components of students' statistics attitudes with Eccles' expectancy value theory (EVT) of achievement motivation (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). A sample of 149 undergraduate psychology students participated at a large urban university in Queensland, Australia. Multiple measurements were conducted. Students' grades in the previous research methods and statistics course were used to measure the past achievement in mathematics and statistics; students' statistics attitudes were measured using the SATS-28; students' expectancies of success were measured using a three-item questionnaire. Path analysis was employed to examine the relationships among the foregoing constructs.

As hypothesized, a large percentage of the variance (approximately 40%) in academic achievement was accounted for by the three constructs. Past performance made the largest direct contribution (around 22%). Besides the direct effect, past performance also made indirect contribution to achievement through the pathway that consisted of expectancies and two statistics attitude components – *Effort* and *Cognitive Competence*. However, the effect size was small with only less than 1% of the variance in the achievement outcome being indirectly accounted for by past performance via students' statistics attitudes. Of note, how much effort students were willing to invest directly accounted for about 8% of the variance in achievement. The findings suggested that effort was also an indispensable component of students' statistics attitudes. Thus, this study succeeded in aligning the SATS with EVT, albeit *a posteriori*.

Paul and Cunnington (2017) extended the previous study by employing path analysis and sophisticated statistical approaches such as Bayesian modeling, an approach that has been heavily popularized across many disciplines (e.g., Deng, Liu, Shao, Rayner, & Yang, 2008). The results suggested that students' past academic performance affected their statistics achievement via *Cognitive Competence*. The small sample size (n = 107) in this study may cast much doubt on both the generalizability and replicability of the findings obtained through the use of Bayesian learning network, as this statistical approach is based on sample covariance structure and thus known to be sensitive to sample size. However, the quantitative results were further triangulated with analysis of the qualitative data from focus group discussions. This mixed methods approach attested to the general validity of this study to a certain degree.

Lastly, there is a paucity of cross-country comparison studies on the relationships between student characteristics and attitudes toward statistics. Wang, Palocsay, Shi, and White (2018) developed a Chinese pre-version of the SATS-36. Students' attitudes were measured and compared only at the beginning of the courses in order to eliminate instructor and course effects. Drawing on a final sample of 292 U.S. and 277 Chinese respondents, the researchers used t-test to compare attitudes toward statistics between students from U.S. and Chinese enrolled in introductory business statistics classes. Results showed that U.S. students were willing to expend more effort in statistics classes than their Chinese counterparts. Chinese students showed positive feelings about statistics whereas U.S. students had rather neutral feelings toward statistics. In addition, Chinese students also had greater appreciation of the utilitarian value of statistics than

U.S. students. Students from both countries perceived statistics to be very challenging yet U.S. students viewed statistics to be easy relative to Chinese students. Lastly, students from both countries showed no differences in their confidences in their ability to learn statistics nor in levels of their interest in statistics.

Altogether, these findings clearly suggested that student characteristics need to be contextualized when it comes to research on students' statistics attitudes, particularly for those characteristics perceived to be relevant for students' prior experiences in mathematics and other quantitative school subjects. For example, probability and statistics have already become an integral part of the school mathematics curriculum in the U.S. whereas topics related to these two disciplines are rarely seen in Chinese school mathematics curricula. Therefore, students from two countries might enter college business statistics classes with different levels of familiarity with statistics. This could cause the differences in the perceptions of statistics (e.g., *Difficulty*) between U.S. and Chinese students. Moreover, achievement-attitude anomaly assumed to arise from reference bias could cast a shadow over the general validity of this cross-country comparison study where students' statistics attitudes were of primary interest (Duckworth & Yeager, 2015; West et al., 2016).

Statistics Attitudes as a Course Outcome

Tempelaar, Schim van der Loeff, and Gijselaers (2007) conducted a study to analyze the relationships among students' attitudes toward statistics, prior reasoning abilities and course performance. Undergraduate students (n = 1,618, 64% male and 36% female) enrolled in introductory statistics courses at the Maastricht University in the Netherlands were administered both the SATS-36 and the Statistical Reasoning Assessment (SRA). The SATS-36 measured students' statistics attitudes while SRA measured both statistical and probabilistic reasoning about chance events defined by seven latent components. In addition, students' course performance was assessed using grades from both quizzes and exams.

A full SEM was utilized to examine the relationships among the three constructs. The results suggested that students' prior statistical reasoning ability was weakly correlated with their course performance whereas students' attitudes toward statistics strongly predicted their course performance for the entire sample of introductory statistics students. The findings confirmed multiple predictions made from the EVT, and thereby demonstrated the strong predictive validity of the SATS-36. Emmioğlu and Capa-Aydin (2012) came to the same conclusion via a meta-analytic study. More interestingly, not only did students' statistics attitudes predict short-term academic achievement, this construct was also found to predict long-term results on statistics examination (Vanhoof et al., 2006).

Given that students' statistics attitudes are at least as important as their academic performance, Schau and Emmioğlu (2012) asked whether introductory statistics courses in the U.S. improved students' attitudes. Using over 2,000 undergraduate students enrolled in service courses from multiple post-secondary institutions across the U.S., the researchers measured attitudes toward statistics at both the beginning and the end of the courses. Average scores of pre- and post-attitudes scores were obtained, and gain scores were then calculated by simply taking the difference. The results suggested that scores for *Effort* and *Interest* decreased by approximately one-half point on average at the end of each semester. Moreover, scores for the rest of attitude subscales remained largely unchanged. The findings indicated that introductory statistics courses in the U.S. fell short of improving students' statistics attitudes. In particular, at the end of introductory statistics courses, students generally became less willing to expend effort to learn more statistics and showed diminished interest in using statistics in everyday life.

An interesting question raised by the previous study is whether students' statistics attitudes remain largely the same throughout the semester. The answer to this question is of great value to statistics instructors because they may capture the period when students' statistics attitudes are on the rise and then maintain the tendency for continued improvements in attitudes. Kerby and Wroughton (2017) administered the SATS-36 at three different time points – the beginning, the middle and the end of an introductory statistics course. The results of the repeated measures analysis showed that five attitude components – *Difficulty, Effort*, *Cognitive Competence*, *Interest*, and *Value* – did change over time on average. This significant change was most likely due to the additional mid-semester measurement. The first half of the semester seemed to be key to improvements in students' statistics attitudes because students' attitudes appeared to be more sensitive to instructional activities over this period than over the second half of the semester. Additionally, Kerby and Wroughton (2017) also analyzed the changes in attitudes at the individual level. The results indicated that a large fraction of students changed their attitudes toward statistics during the second half of the semester, mostly in the negative direction. This finding revealed how attitudes changed in a complicated manner over time and provided great insights into the adequate timing for which teachers can intervene student learning of statistics.

Not only do students outcomes change over time in terms of differences in sample means, changes can also take place in the sample covariance structure of the outcomes. The latter is typically overlooked yet can have significant implications for enrichment of statistics curriculum materials as well as for current research on students' non-cognitive competencies. Xu and Candace (2019) first discovered that variance components associated with the substantive constructs of students' statistics attitudes differed considerably between pretest and posttest. This discovery was only made possible when an additional common method factor was introduced into the regular six-factor model.

Xu and Candace (2019) drew on 1,865 and 1,562 introductory statistics students who completed pre- and post-versions of the SATS-36, respectively. Using common factor model in conjunction with ordinal data-analytical techniques, the researchers found that the *Interest* and *Value* constructs of statistics attitudes showed large method variance yet only at posttest. Therefore, this study unequivocally identified a need to revise the SATS-36 items, as method effects are typically associated with item word-directionality as well as general and/or specific wording (DiStefano & Motl, 2006), or more broadly with the distorted measurement process itself (Maul, 2013). Furthermore, these differential findings at pretest and posttest indicated that common method effects are item- and construct-specific as well as time-dependent. Specifically, students' interest in statistics shifted drastically over time, as did their appreciation of the practicality of statistics. It is unclear under what circumstances these findings can be applied to the design of course interventions. Yet, it is reasonable to posit that, for example, students are more concerned with learning statistics at the beginning of the courses while their interest shifts drastically toward using statistics at the end. As a result, statistics educators may capitalize on the differential aspects of students' statistics attitudes in their effort to enrich statistics curriculum materials.

Pertinent to course intervention, Carlson and Winquist (2011) conducted a study to determine if the workbook curriculum, or a collection of active learning approaches, could improve students' statistics attitudes. A sample of 59 students (13 males and 46 females) completed both pre- and post-versions of the SATS-36. To eliminate possible sources of confounding factors, the researchers also compared students who completed both tests to those who completed only pretests in terms of GPA as well as final exam grades. There were no significant differences in students' academic performance, indicating that student inclination for abstention from the posttest was unrelated to academic skills.

Carlson and Winquist (2011) analyzed average score differences using nonparametric rank test (i.e., Mann–Whitney U test). Post-attitude scores on three subscales – *Cognitive Competence*, *Affect* and *Difficulty* – increased among students who were exposed to the workbook curriculum. The results suggested that statistics courses in tandem with the workbook curriculum tended to build up students' confidence and affect for statistics. In addition, changes in scores were positively correlated with final exam scores. Taken together, active learning may hold great promise for improving both students' statistics attitudes and academic achievement. The findings of this study resonated with the call for the abandonment of lecture-centered statistics instruction in favor of constructivist approaches or a "collaborative mentor-apprentice model" of teaching as commonly seen in post-graduate education (Moore, 1997, p. 126).

Completely randomized experimental design provides unbiased estimates of treatment effects induced by instructional interventions (van Breukelen, 2013). This type of research design is very rare in the field of statistics education research. As of this writing, only one study has been identified (Lesser, Pearl, & Weber, 2016). The researchers conducted a student-randomized experiment to explore the degree to which fun items-embedded instructional strategies (e.g., a 20-second jingle about correlation) affected students' attitudes. Fifty-three students at a two-year college and 115 students at a midsized university in U.S. were asked to complete both pre- and post-versions of the SATS-36 along with the Statistics Anxiety Measure.

The findings included statistically insignificant difference in each of the six attitude subscales between the treatment and control groups. Similar to students' statistics attitudes, statistics anxiety also differed very little between the two groups. The researchers speculated that this statistical insignificance might be due to lack of statistical power as the result of small sample size, and hypothesized in the end for future directions that fun items might ameliorate statistics anxiety among introductory statistics students. However, a large-scale student-randomized study may be required to detect this hypothetical change.

Other uncommon statistics course outcomes include students' conceptions about statistics, a construct similar to statistical reasoning (delMas, Garfield, Ooms, & Chance, 2007). The relationship between this outcome and students' statistics attitudes has also been examined (Evans, 2007). A sample of 115 students in six statistics classes were randomly selected from three different departments at a large U.S. university. Students' statistics attitudes were found positively correlated with academic achievement. Moreover, a marginally significant correlation was found between students' statistics attitudes and their conceptions about statistics. In other words, students with positive attitudes toward statistics tended to have accurate conceptions about statistics. However, as cautioned by the author, this finding is largely suggestive rather than conclusive.

Evans (2007) further conducted interviews with statistics teachers. When asked about their views on interventions for eliminating misconceptions and thus improving attitudes, most teachers agreed that instruction strategies with strong links to real-world problems was among the most effective. For example, one instructor had students collect survey data and use those data to predict the results of an upcoming political election. Overall, the findings corroborated many ideas and philosophies behind social learning theory; that is, effective teaching comes primarily through engaging students in real-life situations.

Social learning theory is woven through many modern learning theories such as active learning, constructivism and learner-based learning, thus leading back to Carlson and Winquist (2011) in which active learning approaches were indeed found beneficial to development of students' statistics attitudes. In a recent study, Songsore and White (2018) provided more supporting evidence to the necessity of teaching the type of statistical concepts that are directly applicable to students' everyday lives and/or tightly related to their career goals. These two emerging themes tie to the *Interest* and *Value* components of students' statistics attitudes.

In this study, students enrolled in an online introductory statistics course offered by a Canadian university were asked to reflect about how much and in what ways statistics became relevant. Online discussion forums were used for data collection. The subsequent thematic analysis revealed two main themes – pragmatism and personal development. Pragmatism consisted of four subthemes – decision-making, entertainment, appraisal and evaluation of real-life situations while personal development included academic and professional development. Given these findings, the researchers suggested that introductory statistics instruction should at least be aligned with teaching strategies that improve *Interest* and *Value*.

There are compelling reasons to believe that introductory statistics students may benefit from high-quality teachers and teaching practices. Schau (2003) reported that, among course sections taught by 11 instructors, the instructor-instructor variability in average attitudes scores was greater at the end of the courses than at the beginning. More importantly, Schau further identified instructor and instructional characteristics as one determinant of students' statistics attitudes along with students' prior experiences with mathematics. This study provided indirect evidence on some associations between instructor and students' statistics attitudes, and thus identified the need to quantify instructor and instructional effects on students' statistics attitudes in a rigorous manner.

Teacher Effects on Students' Academic Attitudes and Behaviors

A substantial body of empirical studies have addressed the relationships between teachers' attributes and students' academic achievement. For example, Rowan et al. (2002) reported that teacher and teaching practice accounted for 4 to 16 percent of variation in students' academic achievement (mathematics and reading). However, there are a very limited number of studies with a focus on how teachers' attributes contribute to students' non-cognitive outcomes.

Dwelling on a nationally representative sample of young students, Downey and Pribesh (2004) examined the effects of student-teacher race match on teachers' evaluation of students' externalizing problem behaviors. The researchers found that gender mismatch significantly affected teachers' behavioral ratings of students. However, this study has limitations as being cross-sectional and therefore does not capture changes in teachers' rating across time.

To ameliorate this methodological problem, Egalite and Kisida (2018) used the Measure of Effective Teaching (MET) Longitudinal database to examine the effects of teacher-student demographic incongruence on students' non-cognitive outcomes. The MET project was funded by the Bill and Melinda Gates Foundation. It included over 1500 teachers from nearly 300 schools. Central to this project is that teachers were randomly assigned to classrooms of students.

Using the MET data, Egalite and Kisida (2018) identified teacher-student demographic similarities by leveraging contemporaneous evaluations of the same students by multiple teachers, a strategy commonly used to stratify samples based on teacher-student relations in educational policy research (e.g., Gershenson, 2016).

Stratifying samples into small blocks prior to regression analysis is necessary for achieving covariate balance, thus enabling the researchers to make causal inferences. Student-reported non-cognitive outcomes included measures of personal effort, happiness in class, feeling cared for and motivated by teachers, student-teacher social and emotional interaction, and college aspirations. The results suggested that students assigned to a demographically similar teacher had larger gains in academic attitudes and behaviors than those assigned to a demographically dissimilar one. Of multiple dimensions of teacher-student demographic congruence examined in this study, gender and racial/ethnic matches demonstrated the largest impact on students' attitudinal benefits. These findings have significant implications for educational policy-making in several key ways. For example, the teacher workforce in current K-12 education is still largely white and female but the student population is becoming increasingly diverse. This study has thus identified a need for diversifying the current teacher workforce, perhaps through exposing underrepresented high school and college students to the teaching profession.

Another commonly researched school/teacher quality is class size reduction (CSR). Class size or student-teacher ratio is presumed to impact student learning, particularly for low-SES students. Early evidence from the Tennessee Student Teacher Achievement Ratio (STAR) project, a random-assignment experiment conducted in the 1980s supported this claim (Mosteller, 1995). Over the years, many states have mandated or incentivized CSR. However, evidence presented in later studies is mixed on how reliable and valid the findings from the STAR project were. A large number of researchers hold vastly different views on the benefits that CSR brings upon student learning outcomes and argue that educational dollars should have been instead spent on teacher pay raise (e.g., Hanushek, 2011).

Similar to teacher demographics, research on class size has also focused predominantly on the extent to which smaller classes affect students' achievement on tests with only a few exceptions. Dee and West (2011) conducted a study to test the long-postulated hypothesis that smaller classes could improve students' non-cognitive competencies. The researchers used a nationally representative sample of students from the National Education Longitudinal Study of 1988 (NELS:88). Students' non-cognitive skills were measured by a handful of items that were subsequently factored into the behavioral and psychological engagements. Due to the observational nature of the NELS:88 data, the researchers split data into small strata and compared outcome measures from within these strata. This identification strategy enabled the researchers to estimate the causal effect of class size on student behavioral and psychological engagements with school.

The results indicated that CSR was associated with active school engagement. The effect sizes were small yet still statistically significant, ranging from 0.05 to 0.09. The effects of CSR decreased yet persisted 2 years later. Dee and West (2011) further conducted cost-benefit analyses of CSR based on the follow-up interviews with over 2,000 NELS:88 participants. By cross-examining both quantitative and qualitative results, the researchers contended that the causal effects of smaller classes on students' non-cognitive skills were true and the effect sizes might be possibly even understated.

If teachers really matter as much as the previous studies suggest in shaping students' non-cognitive outcomes, the central question becomes what teacher attributes predict those outcomes. Apart from teacher demographics and the unknown grouping mechanisms due to smaller student-teacher ratio, other teacher backgrounds such as years of teaching, levels of educational attainment, certification status may also contribute to teacher effects on students' non-cognitive competencies. Jennings and DiPrete

(2010) conducted a study to explore the relationships between teacher background and students' social and behavioral skills. Both skills have long been poised to play a central role in the production of human capital.

Using a teacher valued-added model and the data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) project, Jennings and DiPrete (2010) first examined the size of teacher effects on students' social and behavioral skills and reported that estimated effect size is 0.35 based on the estimation method used in Nye et al. (2004). Next, the researchers examined the correlations among teacher capabilities to improve reading, mathematics, and social and behavioral skills. The results suggested that there was a strong correlation between mathematics and reading while the correlation between non-cognitive skills and either mathematics or reading was marginal. This finding indicated that teachers' ability to improve social and behavioral skills might belong to a specific dimension of teaching practice. Lastly, the researchers used teacher background variables such as education, experience and certification to predict teacher effects. The results showed that these variables weakly predicted how effective teachers were at social and behavioral teaching. Taken together, these findings indicated that there existed additional unidentified dimensions of teaching practices that accounted for the variation in students' non-cognitive outcomes.

Based on motivation and goal-oriented theories, a host of social and behavioral skills may be associated with students' achievement goals. Therefore, a new hypothesis is that teachers could also affect students' achievement goals, which serve as a more direct predictor of educational success than social and behavioral skills. Ruzek et al. (2015) conducted a study to examine teacher's contributions to changes in students' academic achievement and motivational goals. The researchers employed teacher value-added methodology in tandem with Best Linear Unbiased Predictor (BLUP) estimator of random intercept, which is interpreted as teacher effects on student outcomes (Bates, Mächler, Bolker, & Walker, 2015; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). BLUP estimator, better-known as Empirical Bayes estimator in the teacher effectiveness literature (Guarino, Maxfield, Reckase, Thompson, & Wooldridge, 2015), is advantageous compared to other teacher effect estimators because it pulls estimated value toward zero when group sizes are small and/or differences in the outcome scores across teachers are negligible. This well-known "shrunken" property reduces the noise assumed to arise from small group size and insufficient between-level variability, both of which are very common in social sciences and education research.

Ruzek et al. (2015) measured students' achievement goals in terms of mastery-approach goal, performance-approach and performance-avoidance goal. Using data from 2,026 students nested within 35 seventh-grade teachers across seven schools, the researchers found that teacher effects on students' motivation were between 0.03 SD and 0.08 SD. These effect sizes are relatively small yet arguably meaningful. In addition, the correlation between teacher's contribution to students' academic achievement (measured using California Standard Test in mathematics) and teacher's contribution either to students' performance-approach or to performance-avoidance goal was small and statistically insignificant. Only teacher's contributions to students' mastery-approach goal and academic achievement were correlated ($r = 0.50$). The size of this correlation was moderate and therefore suggested that the impact of teachers on students' test scores and achievement goals was not contemporaneous. These findings substantively contributed to educational researchers' understanding of how teachers differentially impact students' non-cognitive competencies.

Blazar and Kraft (2017) extended the existing body of research by examining teachers' contributions to changes in additional dimensions of academic attitudes and behaviors in upper-elementary mathematics classes. The full sample included 10,575 students nested within 310 teachers across 52 schools. Students' academic achievements were measured by both test scores in mathematics on low-stakes assessment and test scores in mathematics and English language arts on high-stakes state assessments. To identify the additional attitudes, the researchers conducted exploratory factor analysis and determined three additional dimensions that consisted of *Behavior in Class*, *Self-Efficacy in Math*, and *Happiness in Class*.

The results from the subsequent application of teacher value-added methodology to students' non-cognitive outcomes substantiated the presence of teacher effects on various dimensions of academic attitudes and behaviors. Namely, estimated teacher effects on students' self-efficacy and behavior in class were 0.14 SD and 0.15 SD, respectively. Furthermore, teacher effects on students' happiness in class were twice as large (0.30 SD). Blazar and Kraft (2017) also found that the correlations were strong between the BLUP estimates of teacher effects within outcome type. In contrast, the correlations were weak between teacher effects across outcome type. Specifically, the correlations between standardized high-stakes mathematics test scores and either self-efficacy in mathematics (r = 0.16) or happiness (r = -0.09) or behavior in class (r = 0.10) was small and statistically insignificant. Kraft (2019) came to the same conclusion by drawing on the randomization data from the MET Project to estimate teacher effects on students' complex cognitive and non-cognitive skills. Together, these findings indicated that teachers who were effective at raising students' test scores were not necessarily effective at improving their academic attitudes and behaviors, thus adding to decades' worth theories of multidimensional teaching (Pianta & Hamre, 2009). However, the authors pointed out that simply correlating the "shrunken" BLUP estimates of teacher effects on different learning outcomes resulted in imprecise correlations. The appropriate way to estimate the stability/consistency of teacher effectiveness across learning outcomes, school subjects, student groups or time is to model multiple responses simultaneously using multivariate multilevel modeling approaches (Grilli, Pennoni, Rampichini, & Romeo, 2016; Leckie, 2018).

Blazar (2018) further validated teacher effects on students' academic attitudes and behaviors using a randomized experimental design. Drawing on data from the National Center for Teacher Effectiveness (NCTE), the author found that the magnitudes of teacher effects on students' *Behavior in Class*, *Self-Efficacy in Math*, and *Happiness in Class* were almost the same as those obtained in Blazar and Kraft (2017). As a result, the author argued that teacher value-added methodology was sufficient to produce valid measures of teacher who were effective at improving students' non-cognitive competencies, at least in the setting of upper-elementary mathematics, although this methodology would have been most useful had it been used in conjunction with classroom observations and student survey.

Despite the ample evidence regarding teacher effects on students' non-cognitive competencies, not enough is known on how teachers' non-cognitive skills come into play. This gap is primarily due to lack of valid measures of teachers' non-cognitive competencies. Cheng and Zamarro (2018) hypothesized that survey effort served as a proxy for respondents' non-cognitive skills. That is, a conscientious teacher is expected to expend much effort while completing a survey. The researchers used the MET project data to test this hypothesis. The results indicated a marginally significant association between measures of survey effort and other teacher quality measures. It also implied that survey effort measure might capture more dimensions of teacher quality compared to traditional teacher quality measures.

Cheng and Zamarro (2018) used an instrumental variable approach to estimate teacher effects on students' non-cognitive outcomes that were measured in terms of self-effort and self-grit. The results suggested that survey effort predicted students' self-effort and self-grit, albeit in a suggestive and inconclusive manner. In other words, teachers' conscientiousness, proxied by how much effort teachers expend in completing surveys, may also be an important dimension of teacher quality that holds promise for improving students' academic attitudes.

Summary of Findings

Student characteristics predict statistics attitudes. However, to date, evidence is mixed primarily due to vastly different contexts in which studies were conducted. Coetzee and van der Merwe (2010) found that age appeared to affect students' statistics attitudes; younger students perceived statistics to be more difficult. Cashin and Elmore (2005) reached the opposite conclusion on the direction in which age affected students' statistics attitudes. van Es and Weaver (2018) reported that female students tended to underscore on three attitude subscales – *Affect*, *Cognitive Competence*, and *Difficulty*. This finding was inconsistent with those from other studies conducted in U.S. (Carnell, 2008; Cashin & Elmore, 2005), both of which reported no differences in students' statistics attitudes between males and females. van Es and Weaver (2018) also found racial groups did not differ on students' statistics attitudes.

In addition, a general consensus has been reached regarding the strong predictive role that prior mathematics achievement plays. Multiple studies showed that students who had a strong quantitative background tended to have more positive attitudes toward statistics at both the beginning and the end of the courses (Cashin & Elmore, 2005; Chiesi & Primi, 2010; Dempster & McCorry, 2009; Hood et al., 2012; Paul & Cunnington, 2017). Wang et al. (2018) conducted a cross-country comparison study and showed the differences in attitudes toward statistics among college students from U.S. and Chinese enrolled in business statistics courses. Their study demonstrated the importance in contextualization of students' quantitative backgrounds when students' academic attitudes were the learning outcome of interest.

Students' statistics attitudes have emerged as an important learning outcome, as students with positive attitudes are more likely to use and/or re-learn statistics in the everyday lives after they leave introductory statistics courses. Tempelaar et al. (2007) showed that students' statistics attitudes measured at the end of the semester were strongly correlated with final examination scores. This finding was corroborated by a study utilizing meta-analytical approaches (Emmioğlu & Capa-Aydin, 2012). Furthermore, Vanhoof et al. (2006) found this was also the case for the association between students' attitudes and long-term results on statistics examination yet not on general mathematics examination. This finding indicated that students' attitudes may later materialize as factors that are essential to student future success in terms of becoming a staunch learner and user of statistics after they graduate.

Despite the strong predictive power that students' statistics attitudes have on academic achievement, introductory statistics courses in the U.S. did not improve students' attitudes toward statistics (Schau & Emmioğlu, 2012). Some attitude components even decreased at the end of courses. Kerby and Wroughton (2017) administered an additional measurement of students' statistics attitudes at mid-semester and found that attitudes did not remain stable throughout the semester nor decreased monotonically. Rather, many students' attitudes toward statistics increased over the first half of the course, suggesting that instructors should seize this period when attitudes are on the rise. Instead of focusing on attitudinal changes in mean scores, Xu and Schau (2019) examined the common method variance associated with the substantive constructs of students’ statistics attitudes. The *Interest* and *Value* constructs showed a large proportion of method variance only at the posttest, indicating the dynamic nature of students’ attitudes.

Other researchers examined course interventions that had potentials to improve students’ statistics attitudes. Carlson and Winquist (2011) implemented a workbook curriculum in an introductory statistics class and found students’ statistics attitudes increased at the end. In contrast, Lesser et al. (2016) failed to observe the similar trend through the use of fun items as a means of improving students’ statistics attitudes although their instructional approaches seemed to reduce statistics anxiety. Beside a successful replication of the quantitative findings in Tempelaar et al. (2007), Evans (2007) also conducted interviews with instructors and concluded that students’ exposure to real-world statistics problems might improve students’ statistics attitudes. Songsore and White (2018) further identified two attitudinal themes pertinent to real-life situations – pragmatism and personal development. Lastly, instructor can and do have large impact on students’ statistics attitudes, as preliminarily evidenced in Schau (2003) where the author observed a large instructor-level variability in average attitude scores at the end of the courses relative to the variability at the beginning.

Teacher and teaching practices can and do affect a vast variety of students’ non-cognitive outcomes. Earlier studies focused on the impact of teacher-student demographic congruence on teachers’ ratings of students’ social and behavior skills. Findings suggested that racial mismatch led to poor behavioral ratings of black and/or low-SES students (Downey & Pribesh, 2004). Egalite and Kisida (2018) reported that students assigned to demographically similar teachers had larger gains in academic attitudes and behaviors than those to demographically dissimilar teachers. Apart from teacher demographics, student/teacher ratio or class size was found strongly correlated with students' active school engagement (Dee & West, 2011).

Current development in data storage, methodology and computing capacity has greatly speared ahead research effort in the field of teacher effectiveness research. Drawing on large data sets and teacher value-added methodology with student characteristics being controlled for, several studies found that teachers did have a large impact on multiple dimensions of attitudes and behaviors including social and behavioral skills (Jennings & DiPrete, 2010); motivation and achievement goals (Ruzek et al., 2015); self-efficacy in mathematics, behaviors and happiness in class (Blazar & Kraft, 2017). Blazar and Kraft (2017) further examined the correlations between teacher effects on both test scores and students' noncognitive competencies, and found teachers who raised students' test scores were not equally effective at improving students' academic attitudes and behaviors. This finding provided strong empirical support to the inconsistency of teacher effects across types of student outcomes (Loeb & Candelaria, 2012). Kraft (2019) corroborated this finding by exploiting the data on randomization of class rosters. Blazar (2018) further validated the measures of teacher and teaching quality that was effective at improving students' academic attitudes and behaviors. Moreover, Cheng and Zamarro (2018) found that teachers' conscientiousness contributed to improvements in students' self-grit and self-effort, clearly indicating the importance of teachers' attitudes toward teaching as one dimension of teaching competencies.

Theoretical Framework

This study is firmly grounded in Eccles' Expectancy-Value Theory (EVT), one of the most influential theories conceived to explain students' academic attitudes and behaviors. EVT provides a framework that has been frequently applied to a vast variety of educational contexts (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000), particularly those perceived to be relevant for American K-12 mathematics education. Crucial for this study is that this theory was used to operationalize students' statistics attitudes. In their conceptualization of the Model of Students' Attitudes Toward Statistics, Ramirez et al. (2012) *a posterior* aligned EVT with students' statistics attitudes. Specifically, Eccles' value super-construct corresponds to a combination of interest in statistics, enjoyment from doing statistics, importance of statistics attached to everyday activities, and diligence. Eccles' difficulty construct corresponds to students' perception about the levels of difficulty associated with statistics. Eccles' self-concept construct matches students' cognitive competency.

Eccles' EVT also provides theoretical links between teaching behaviors and students' underlying developmental processes such as their academic attitudes, behaviors, and emotional skills (Muenks, Wigfield, & Eccles, 2018). Some of the students who have large goals may exhibit disruptive attitudes and behaviors in classes where competition among students is emphasized and high achievement with low effort is valued. This is the place where teachers assume a pivotal role in properly organizing classrooms as well as encouraging students to develop strong non-cognitive skills that keep them working hard and away from disrupting the classroom.

The theoretical links provided by the EVT underpin the current study. Statistics instructors at the high end of the instructor effectiveness distribution are expected to be more likely to hold students' perspectives in regard, show sensitivity to students' frustration, and create a positive climate in statistics classrooms. For example, a positive classroom climate not only establishes a positive teacher-student relationship that fosters growth in self-efficacy by creating a safe learning environment free of ridicule and criticism; thus, motivating the students to exert greater effort and persistence making it more likely for them to be academically successful (Liu, Bridgeman, & Adler, 2012)

Conclusion

This chapter reviewed the three key components of the value-added model for this study: (a) student characteristics, (b) students' statistics attitudes, and (c) teacher and teaching attributes. All of the three components are essential for examining instructor and instructional effects on students' statistics attitudes. The following chapter will describe the quantitative methodology to be used for this study.

CHAPTER III:
METHODOLOGY

The purpose of this study was to examine instructor and instructional effects on students' statistics attitudes. This study included a thorough and in-depth analysis of a large-scale data set obtained from a purposeful sample of introductory statistics instructors and their respective students from multiple post-secondary institutions across the U.S. who participated in the SATS project. The data were analyzed using an array of statistical techniques, including descriptive statistics, paired *t*-test, correlation analysis, and hierarchical linear model (HLM)/multilevel model with covariate adjustment. This chapter presents an overview of the research problem, operationalization of theoretical constructs, the purpose of the study, research questions, research design, population and sample, instrumentation, data collection procedures, data analysis, privacy and ethical considerations, and limitations of this study.

Overview of the Research Problem

Recently, students' academic attitudes have received much attention among education researchers and policy makers (Blazar & Kraft, 2017). In addition to raising test scores, teachers are also expected to improve students' attitudes toward various school disciplines. Longitudinal studies have documented the beneficial effects of students' academic attitudes and behavior on their long-term outcomes such as educational attainment, adult earnings and health status (Chetty et al., 2011; Moffitt et al., 2011). Students' statistics attitudes have also emerged as a primary learning outcome (Ramirez et al., 2012). Despite the importance of academic attitudes, this construct remains much less studied than other course outcomes such as statistics achievement and statistical reasoning. As a result, not enough is known about what instructor and/or instructional attributes tend to improve students' statistics attitudes.

A long-standing hypothesis in the field of statistics education has posited that instructors can have a large impact on students' statistics attitudes (Schau, 2003). However, it is difficult to determine to what degree the extant empirical evidence support this claim. If instructors mattered as much in impacting students' statistics attitudes as many researchers and educators suggested, the critical question becomes "To what extent do instructors affect students' statistics attitudes?" To date, there exists very little, if any, evidence related to this issue. In addition, if instructors do have a large impact on students' statistics attitudes, then it is possible to identify certain dimensions of instructional practice that (at least partly) account for this impact. The first research problem is concerned with estimating instructor effects and the second one is with determining instructional effects (Blazar & Kraft, 2017).

Operationalization of Theoretical Constructs

This study consisted of the following constructs: (a) students' statistics attitudes, (b) instructors' attitudes toward teaching, (c) instructors' emphasis on data contexts, (d) instructors' emphasis on assessment, and (e) instructors' emphasis on conceptual understanding. Students' attitudes toward statistics were operationalized into six subscales – *Value*, *Cognitive Competency*, *Interest*, *Difficulty*, *Effort*, and *Affect* (Schau, 2003; Schau et al., 1995). Instructors' attitudes toward teaching were categorized into attitudes toward introductory statistics courses and attitudes toward teaching the specific course sections. Instructors' emphasis on assessment referred to the frequency in which instructors used assessment methods to evaluate and improve statistics instruction. Instructors' emphasis on data contexts referred to the extent to which instructors integrated data collection and analysis into statistics instruction. Instructors' emphasis on conceptual understanding referred to the extent to which instructors developed conceptual understanding in introductory statistics courses.

Research Purpose and Questions

The purpose of this study was to examine instructor and instructional effects on students' attitudes toward statistics. The following research questions guided this study:

1. Does taking introductory statistics courses change students' statistics attitudes?
2. Is there an unexplained variation across statistics instructors regarding students' statistics attitudes?
3. Do instructors' attitudes toward teaching predict students' statistics attitudes?
4. Does instructors' emphasis on data contexts predict students' statistics attitudes?
5. Does instructors' emphasis on assessment predict students' statistics attitudes?
6. Does instructors' emphasis on conceptual understanding predict students' statistics attitudes?
7. Do instructors who improve students' statistics attitudes also improve expected course grades?

Research Design

For the purpose of this study, an array of quantitative approaches was utilized to examine instructor and instructional effects on students' statistics attitudes. A purposeful sample of introductory statistics instructors and their respective students from multiple post-secondary institutions across the U.S. who participated in the SATS project were selected for inclusion in this study. The variables of interest were categorized into two levels – student and instructor. Student-level variables consisted of gender, age, prior experiences with mathematics, pre-attitude scores and post-attitude scores while instructor-level variables included instructors' attitudes toward introductory statistics courses and teaching their own classes as well as their emphasis on data contexts, assessment and conceptual understanding. Additionally, students were clustered under instructors. This feature gave rise to multilevel data (Burstein, 1980). Given the hierarchical nature of the data, HLM/multilevel model was deemed appropriate as opposed to aggregating data at a single level (Muthén, 1989, 1994).

Population and Sample

The population of interest in this study is all undergraduate students who enroll in college introductory statistics courses with no mathematics prerequisite or a college algebra-only prerequisite. The sample consisted of introductory statistics students from multiple post-secondary institutions across the U.S. Approximately 3,800 students provided responses to the survey administered. Observations missing over four item responses (10%) and/or key survey information were deleted. Moreover, the course sections with over 40% response rate were retained, a selection criterion adopted from Chance, Wong, and Tintle (2016). These selection criteria resulted in the final sample of 1,924 students within 90 classes taught by 23 instructors. A slight gender imbalance was observed: 40% male students and 60% female students. The median age of the students was 19.9 years. Race/ethnicity for both students and instructors was not recorded. Table 3.1 presents the sample demographics for the participating instructors and students.

The number of participating students varied considerably across course sections. That is, 80 sections contained 10-50 students; six sections included fewer than 10 students; and four sections included over 50 students. Of these 90 course sections, 46 sections were taught by 11 instructors at advanced degree-granting institutions. A purposeful sample of instructors who taught college introductory statistics courses with no mathematics prerequisite or a college algebra-only prerequisite (i.e., service courses) was utilized for this study.

Table 3.1

Demographic Information for Students and Teachers

	Students		Instructors	
	Frequency (*n*)	Percentage (%)	Frequency (*n*)	Percentage (%)
Total Sample Size	1,924	100.0	23	100.0
1. Gender				
Male	769	39.8	4	17.4
Female	1,155	60.2	19	82.6
2. Age				
15-20	1,068	55.5		
20-25	635	33.0		
25-30	203	10.6		
30-40	14	0.7		
40-50	2	0.1		
>50	2	0.1		
3. Rank				
Full Professor			4	17.1
Associate Professor			4	17.1
Assistant Professor			7	30.8
Adjunct (Full Time)			2	9.0
Adjunct (Part Time)			6	26.0
4. Highest Degree				
Doctorate			16	69.0
Master’s			7	31.0

Instrumentation

Survey of Attitudes Toward Statistics (SATS-36)

Both pre- and post-course attitude scores were measured by the SATS-36. Multiple psychometric aspects of the SATS-36 have been validated, including substantive validity, content validity, external validity, structural validity, predictive validity and convergent validity (Nolan, Beran, & Hecker, 2012). Sufficient evidence has accumulated on measurement invariance of the SATS-28 across administration time and between genders (Dauphinee, Schau, & Stevens, 1997; Hilton, Schau, & Olsen, 2004). The psychometric measurement of students' statistics attitudes using the SATS-36 also appears to be very stable across student populations from many countries other than the U.S. (e.g., Persson, Kraus, Hansson, & Wallentin, 2019; Vanhoof et al., 2011; Wang et al., 2018), although the invariant properties of the SATS-36 remains largely unclear.

In addition, the psychometric properties of the SATS-36 at individual item level have also been studied (Xu & Schau, 2019). Findings from these studies confirmed the six-factor structure of the SATS-36, although certain modifications of survey items appear to be worthy (and necessary). Several *Difficulty* items were found to have weak factor loadings. However, more replications are needed to come to a firm conclusion that survey revision is needed. In addition, there are noticeable proportions of common method variance, which were found to be associated with the *Difficulty* construct only at pretest and with the *Value* and *Interest* constructs only at posttest.

Appendix C provides an overview of the pre-version of the SATS-36. The pre and posttest versions contain identical items except for changes in tense. Students receive a mean score on each subscale. The SATS-36 contains 36 seven-point Likert scale items (1 = Strongly Disagree, 4 = Neutral/No opinion, 7 = Strongly Agree). Each of these items belongs to one of six attitude subscales – *Affect* (6-items), *Cognitive Competence* (6-items), *Value* (9-items), *Difficulty* (7-items), *Interest* (4-items) and *Effort* (4-items).

The composite scores for *Affect*, *Cognitive Competence*, *Value*, *Difficulty*, *Interest* and *Effort* are 6-42, 6-42, 9-63, 7-49, 4-28 and 4-28, respectively. The students who give higher numerical responses to any item have more positive attitudes than those who give lower responses. In addition, the *Difficulty* items are reversely coded. Thus, the students who have higher composite scores on the *Difficulty* component perceive statistics to be less difficult. Cronbach's coefficient alpha values for *Affect*, *Cognitive Competence*, *Value*, *Difficulty*, *Interest* and *Effort* in pre-version of the SATS-36 are 0.81, 0.84, 0.87, 0.76, 0.89 and 0.81, respectively. The corresponding alpha values in post-version are 0.85, 0.86, 0.90, 0.79, 0.91 and 0.77 (Schau & Emmioğlu, 2012). According to the Cronbach's coefficient alpha values for all of the six attitude subscales, both versions of the SATS-36 exhibits good to excellent internal consistencies.

Instructor-Specific Measure

The instructor-specific measure consists of 10 items. They were selected from a pool of over 20 items on the original questionnaire. Table 3.2 shows the items and the domains into which each item was classified. These items were selected based on a search for key instructional dimensions in the literature of statistics education as well as on the six recommendations that are made by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report (GAISE College Report ASA Revision Committee, 2016).

Two questions measure instructors' attitudes toward introductory statistics courses and attitudes toward teaching their classes, respectively, on seven-point Likert scale (1 = strongly dislike, 4 = neutral/no opinion, 7 = strongly like). The additional eight questions were created in line with the GAISE recommendations, on five-point Likert scale (1 = strongly dislike, 3 = neutral/no opinion, 5 = strongly like). The composite scores for *Concept* (4-items), *Data* (2-items), and *Assessment* (2-items) are 4-20, 2-10, and 2-10, respectively. The instructors who give higher numerical responses to any item expose students to the corresponding instructional practice more frequently than those who give lower responses. For the *Concept*, *Data*, and *Assessment* domains, final scores were created by averaging raw responses across the corresponding set of items.

Table 3.2

Descriptions of Instructor-Specific Items

Domain	Variable name	Question for instructors
1. *Concept*	StatLit	Emphasizing statistical literacy
	StatThinking	Emphasizing statistical thinking
	TechConcept	Using technology to develop conceptual understanding
	ActLearn	Fostering active learning
2. *Data*	TechData	Using technology for data analysis
	DataContext	Using data in a meaningful context
3. *Assessment*	EvalStuLearn	Using assessments to evaluate student learning
	ImprStuLearn	Using assessments to improve student learning
4. Gen_*Attitudes*	GenThisCourse	I like teaching introductory statistics course.
5. Spe_*Attitudes*	ThisCourse	I liked teaching this specific section of this course.

Data Collection Procedures

The research gained approval from the University of Houston-Clear Lake (UHCL)'s Committee for Protection of Human Subjects (CPHS) before any data were collected. The data used in this study are from students' responses obtained from the SATS Project data set. SurveyMonkey, a web-based data collection software program, was used to collect the data across three academic years from the 2007 fall term through the spring term of 2010. Instructors teaching U.S. statistics courses volunteered to ask their students to take the SATS-36. Students responded to the survey during or outside of class within two weeks of the beginning and of the end of their classes. The length between pre- and post-course administrations of the SATS-36 is approximately 12 weeks.

The archived data were initially stored in Microsoft Excel when they were received. The data are being secured on the researcher's computer hard drive and Google Drive at all times. Once the study is completed, the researcher will maintain the data for 5 years which is the required time set forth by CPHS and district guidelines. Once the deadline has passed, the researcher will permanently destroy all data files.

Data Analysis

First, the archived data obtained from the SATS Project were imported into a R data frame. Survey non-responses were imputed using the hot-deck method (Andridge & Little, 2010), which was implemented through R package *VIM* (Kowarik & Templ, 2016). Crucial for this study is that imputation did not induce biases nor diminish variability in the sample. An array of statistical techniques was employed to analyze the quantitative data, including descriptive statistics, paired *t*-test, correlation analysis and HLM with covariate adjustment. The R software was used throughout this study for data processing, management, and analysis (R Core Team, 2018).

To answer the first research question, a paired *t*-test was conducted to examine whether or not students changed their attitudes toward statistics from the beginning to the end of the semesters. To answer the second research question, unconditional models were estimated to assess the extent of unexplained variation across statistics instructors for each of the six attitude components. Chi-square difference test was used to examine the difference in fit caused by models with and without random effects. This simple test is straightforward and produces a likelihood ratio test statistic, which can be used to construct more sophisticated model-fitting indices (Hu & Bentler, 1999). Given the nested nature of the data and the existence of between-level variation, HLM (i.e., multilevel model) was deemed appropriate as opposed to single-level data aggregation in order to quantify the sizes of instructor effects on students' statistics attitudes (Muthén, 1989, 1994).

To answer research questions three through six, multilevel covariate adjustment models with instructor-specific covariates were employed to examine whether students' statistics attitudes at the end of the semesters were predicted by (a) instructors' attitudes toward teaching, (b) instructors' emphasis on assessment, (c) instructors' emphasis on data contexts, and (d) instructors' emphasis on conceptual understanding. Lastly, to answer research question seven, two sets of random effects were estimated; one set related to instructor-associated contributions to students' statistics attitudes and the other related to instructor-associated contributions to students' expected course grade. Pearson correlation was used to examine the relationships between the two sets of random effects.

Following conventions (e.g., Ballou, Sanders, & Wright, 2004; McCaffrey et al., 2004), a multilevel model with covariate adjustment was specified to estimate instructor effects on students' statistics attitudes for student i linked with instructor j:

$$Y_{ij\_post} = \alpha Y_{ij\_pre} + \pi X_{ij} + \varphi \bar{X}^{c} + \omega_{s} + \tau_{j} + \varepsilon_{ij} \qquad (1)$$

Here, Y_{ij_post} is the outcome of interest measured at the end of the courses. The model includes a vector of student characteristics X_{ij} (i.e., gender, age, self-reported previous mathematics achievement), the same X_{ij} averaged to course section level $\bar{X}^{c}$, and a school type fixed effect ω_{s}. Previous mathematics achievement is another determinant of students' statistics attitudes besides instructor and instruction characteristics while gender and age were frequently found to account for a fraction of the variation in students' statistics attitudes (Ramirez et al., 2012). In addition, the purposes of controlling for $\bar{X}^{c}$ and ω_{s} in the model were to remove peer effects across course sections and any effects due to different type of schools, respectively (Blazar & Kraft, 2017).

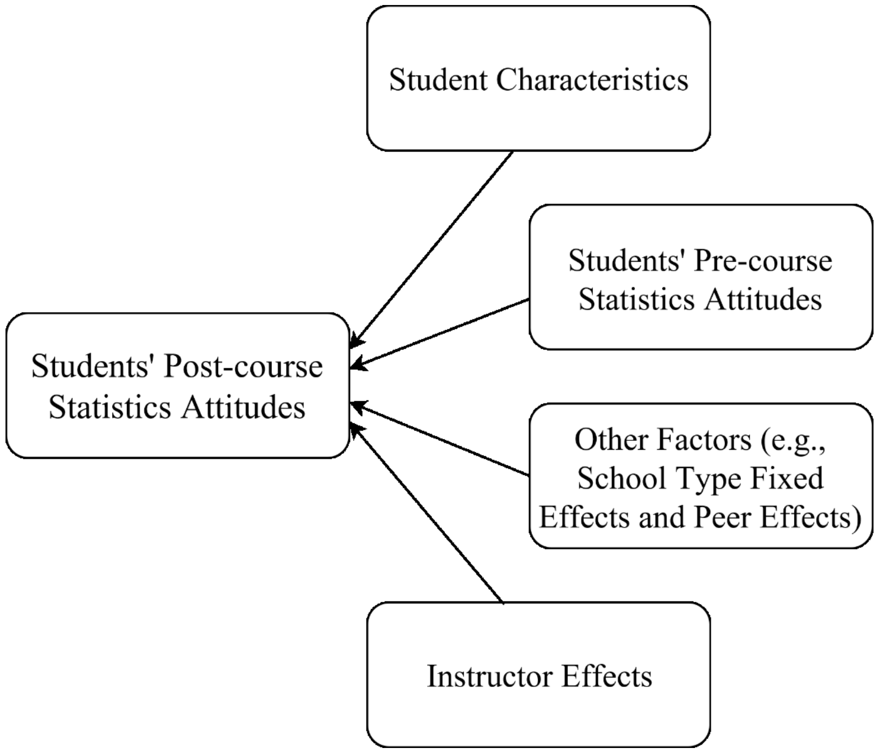

Figure 3.1 Conceptual Overview of the Value-Added Model Used in This Study

A conceptual overview of the value-added model used in this study is diagrammed in Figure 3.1. It visualizes all the components in Equation 1 expect the idiosyncratic student-specific error. The directions of the arrows illustrate the output nature of students' post-course statistics attitudes as well as the input nature of student characteristics, effects of peers in the same course sections, school type fixed effects, students' pre-course statistics attitudes, and instructor effects. This visual representation makes a clear case as to how value-added models take into consideration various factors that contribute to students' statistics attitudes at the end of the courses.

Further, to address the possibility of biases in estimated instructor effects due to the observational nature of the SATS project data, a prior measure of the corresponding attitude outcome (Y_{ij_pre}) was included. Very few, if any, studies involving post-course students' statistics attitudes as outcomes adjusted for covariates such as pre-course attitudes. This is methodologically problematic since students' statistics attitudes measured at the beginning and the end of the courses are typically correlated to a considerable degree. Failure to adjust for prior scores on statistics attitudes is mathematically equivalent to misspecification in regression models (Millar, & Schau, 2010).

The empirical strategy of including pre-course statistics attitudes was intended to mitigate the biases in estimates of instructor effects due to nonrandom sorting of students across instructors. The core assumption underlying this strategy was that the student outcomes of interest observed at the beginning of the semesters is sufficient to summarize all the factors that have caused the disparities in those outcomes up to that point (Ballou et al., 2004). In practice, those factors are often unmeasurable and/or beyond instructors' control. For Equation 1, Y_{ij_post} and Y_{ij_pre} were standardized into z-scores so estimated instructor effects on students' attitudes were on the same scale as those obtained in other studies (Blazar, 2018; Blazar & Kraft, 2017; Kraft, 2019; Ruzek et al., 2015). This standardization procedure facilitated comparison across different studies.

Note that Hanushek and Rivkin (2010) reviewed a large body of the value-added literature and proposed that in practice the true between-school variation was so small as to be negligible. In addition, incorporating school fixed effects was likely to cause several problems such as difficulty in interpretation (p. 268). However, this proposal was not adopted by several recent value-added studies (Gershenson, 2016; Jennings & DiPrete, 2010; Ruzek et al., 2015). Technically, a separate yet stronger concern with inclusion of the school fixed effects is that this specification implicitly assumes that there is no school-level confounding; that is, school and teacher fixed effects are not subject to collinearity. It is assumed that there would be no dependency of instructor-level random effects on schools even when school fixed effects were introduced. Such dependency, if existing, would be rather pernicious since it leads to downward biases in teacher effectiveness estimates.

A preliminary examination of the SATS project data revealed a strong collinearity between fixed effects due to instructors and the institutions to which they belonged. This problem was exacerbated when the model in which *Difficulty* was the outcome of interest and school was included as a fixed effect resulted in *singular* fits. Technically, such fits stemmed from estimated variance-covariance matrices that are not of full rank (Bates, et al., 2015); that is, the variance estimate of the instructor-level random effect is (very close to) zero. For the other attitude components, the issue with collinearity was not as severe as that detected with *Difficulty*.

The collinearity was very likely due to sampling error as several institutions contained only single instructors. Therefore, the tentative models with school fixed effects being specified might produce a lower boundary of estimates of instructor effects on students' statistics attitudes. Given the suggestions offered in Hanushek and Rivkin (2010), the final models in this study included only school type fixed effects but not school fixed effects. However, Blazer and Kraft (2017) suggested that fixed school effects should be included in order to account for nonrandom sorting of students and teachers across schools and to reduce the possibility of reference bias in self-reported measures of students' academic attitudes and behaviors. In any cases, this assumption needs to be carefully checked in the stage of value-added model specification. In addition, whether or not to include school fixed effects in studies conducted in the higher education settings instigate further research to warrant the accuracy of instructor effect estimates.

As with any other studies in the teacher effectiveness literature (e.g., Ballou et al., 2004; Blazar & Kraft 2017; Guarino et al., 2015; McCaffrey et al., 2004; Rubin, Stuart, & Zanutto, 2004), only the intercept was allowed to vary across instructors in the present study while the slopes were treated as fixed. This specification of multilevel covariate adjustment models led to a statistical equivalent of linear mixed effects model with only random intercept being specified. Random slope models are also easy to implement. However, including random slopes would make the value-added results difficult to interpret (McCaffrey et al., 2004).

The error structure is of two levels – the instructor-specific random error τ_j and the idiosyncratic student-specific random error ε_{ij} with zero means. Both errors assume normal density. As per usual, two levels of errors were assumed to be uncorrelated with each other. Furthermore, the errors were assumed to be uncorrelated with school type fixed effects, and both student- and instructor-specific covariates. In practice, however, it is very difficult to tell based on the empirical evidence to what degree these assumptions are not violated. Violation of any of the above assumptions runs the risk of producing biased instructor and instructional effect estimates.

The sizes of instructor effects on students' statistics attitudes were obtained by quantifying the variance of the instructor-level random error τ_j (i.e., σ^2_τ) via restricted maximum likelihood (REML) estimation. Unlike maximum likelihood estimation that yields (downwardly) biased estimator of variance component, REML ascertains both the maximal efficiency and the consistency of σ^2_τ estimator by eliminating the undesired effects due to estimation of mean structure (Bates et al., 2015). Here, maximal efficiency refers to the least mean squared error upon the appropriate choice of an objective function and consistency refers to convergence of estimator to its true value with a larger sample size. In other words, instructor effects on students' statistics attitudes were estimated in some best possible manner using REML.

Instructor effects are formally defined as the square root of σ^2_τ (i.e., standard deviation of the instructor-level random error). Conditional intraclass correlation coefficient (ICCs) were also computed based on Equation 1 (i.e., a model with all covariates included as opposed to an unconditional model in which only random effects are included). Conditional ICCs serve as yet another useful metric that quantifies the instructor-to-instructor variability in student outcome data. Lastly, BLUP estimates were extracted to capture instructors' contribution to each outcome, followed by an analysis of correlations between the BLUP estimates. To be precise, BLUP predictor is defined as the conditional expection (or posterior mean) of the instructor-level random effect given students' post-course statistics attitudes or $E(\tau \mid Y_{post})$.

To examine instructional effects on students' statistics attitudes, the following multilevel model was specified:

$$Y_{ij\_post} = \beta T_j + \alpha Y_{ij\_pre} + \pi X_{ij} + \varphi \bar{X}^c + \omega_s + \tau_j + \varepsilon_{ij} \qquad (2)$$

Beyond and above the same components, Equation 1 was modified to further include a vector of the jth teacher's self-reported perceptions of teaching practices scored on the previously identified instructional domains (T_j). All but two instructors taught at least two course sections. The outcome scores were averaged across course sections taught by one individual instructor and thus served as instructor-specific variables. Consequently, the vector of regression coefficients β contained the parameters of primary interest (i.e., instructional effects on students' statistics attitudes). For Equation 2, Y_{ij_post} and Y_{ij_pre} were kept on the original scale to make the unstandardized estimates of regression coefficients more interpretable.

Multilevel models with covariate adjustments were estimated using R package *robustlmm* (Koller, 2016) and *lme4* ((Bates et al., 2015). The former package provides a series of robust methods that correct for the skewness and kurtosis in the outcome data, thereby producing adjusted parameter estimates and standard errors. Sample R code in relation to *robustlmm* was provided in Appendix B to facilitate replicability as well as encourage replication of the results from the present study. Note that, instead of fitting several univariate multilevel models to individual learning outcomes separately, this study would have been more methodologically rigorous had it been done using a multivariate multilevel modeling approach (e.g., Grilli et al., 2016; Leckie, 2018). Hadfield (2010) presents this type of models and the mathematics that is required to estimate these models, and also provides the sample R code for fitting these models via Markov Chain Monte Carlo algorithm.

Lastly, all dependent and independent variables except student gender and school type were treated as (approximately) continuous variables. Normal assumptions were checked using the Kolmogorov-Smirnov (KS) test. Results from the KS test indicated that normality appeared to be an appropriate approximation for all attitude scores except the *Effort* scores, which severely clustered toward the high end and exhibited very small between-instructor variability, a phenomenon commonly referred to as ceiling effect.

A significance level at 0.05 was used as an empirical cutoff for hypothesis testing throughout this study. Note that *p*-value alone is not sufficient in reporting of inferential results in most research contexts, according to the guideline set forward by the American Statistical Association (Wasserstein & Lazar, 2016). Particularly, due to the large sample size in this study, *p*-value was always small and thus invariably rejected null hypotheses. As a result, local effect sizes in terms of Cohen's f^2 were calculated in the multilevel setting along with *p*-value, using the formula as described in Selya, Rose, Dierker, Hedeker, and Mermelstein (2012).

Privacy and Ethical Considerations

The researcher gained approval from the University of Houston-Clear Lake's Committee for Protection of Human Subjects (CPHS) prior to any research activities. Permission to use the SATS-36 and the SATS project data was granted to the researcher by Candace Schau (see written consent in Appendix A), who created the SATS-36 and led the SATS project. The archived data are being secured on the researcher's computer hard drive and Google Drive at all times. Once the study has been completed, the researcher will maintain the data for 5 years which is the required time set forth by CPHS and district guidelines. Once the deadline has passed, the researcher will permanently destroy all data files.

Limitations of the Study

There were several limitations to this study. First, an important caveat in the design of this study was that, without videotaped observation of classroom activities, the degree of consistencies remained unknown between instructors' self-perceptions of teaching practices and their actual teaching performances. The inconsistencies could affect the validity of the approach used in this study. Second, the archived data might be subject to human error when it was created and reported. The researcher of this study had no means of detecting, rectifying and controlling for the possible inaccuracies. Third, the generalizability of the quantitative findings may be called into question for the following three reasons: (a) the absence of the demographic information on gender/ethnicity, (b) a limited number of participating post-secondary institutions, and (c) rapid demographic change in schools that may not be accurately captured by the archived data. Fourth, one additional concern about the current study was the issue of small instructor-level sample size (N = 23). The limited number of instructors might have caused insufficient statistical power and imprecise estimation when it comes to estimating both instructor and instructional effects on students' statistics attitudes. Jak, Oort, and Dolan (2014) suggested over 50 units at the second level. Lastly, the reliability of the instructor-specific measures used in this study remained unknown. This factor might have spuriously influenced the precision with which instructional effects were estimated on students' statistics attitudes.

Conclusion

The purpose of this study was to examine instructor and instructional effects on students' attitudes toward statistics. To test a long-standing hypothesis positing that instructors could have a large impact on students' students' statistics attitudes, a strictly quantitative approach was utilized. This chapter provides the details of the methods required to test the above hypothesis. In the next chapter, findings and results will be presented.

CHAPTER IV:
RESULTS

The purpose of this study was to examine instructor and instructional effects on students' statistics attitudes. This chapter presents the results from a multilevel analysis of the quantitative data. A detailed explanation of the participants' demographics and other important information is first presented, followed by the results for each of the seven research questions.

Participant Information

This study included 1,924 college students and 23 statistics instructors who consented to participate in the SATS project. Table 4.1 provides demographics and other important information for participating students and/or instructors. Of the 1,924 students, 1,155 (60%) were female and 769 (40%) were male. The median age of students was 19.9 years. Of the 23 instructors, seven (31%) had Master's degrees and 16 (69%) had doctoral degrees; six (26%) were part-time adjuncts, two (9%) were full-time adjuncts, seven (31%) were assistant professors, four (17%) were associate professors, and four (17%) were full professors. Participants were selected from three major types of U.S. institutions (five Baccalaureate Colleges, four Master's Colleges and Universities, and two Doctoral/Research Universities). Three (13.0%) instructors taught 235 (12.2%) students in doctoral/research universities; seven (30.4 %) instructors taught 536 (27.9%) students in master's college and universities; 13 (56.6%) instructors taught 1,153 (59.9%) students in Baccalaureate Colleges.

Besides age, students' previous achievement in mathematics was included in this study, as this variable was also found to strongly predict students' statistics attitudes. The median previous achievement in mathematics (premach) was 6.00 (measured on a 1-7 scale). Results suggested that the great majority of students perceived themselves to be very good at mathematics.

Table 4.1

Demographic Information for Students and Teachers

	Students		Teachers	
	Frequency (*n*)	Percentage (%)	Frequency (*n*)	Percentage (%)
Total Sample Size	1,924	100.0	23	100.0
1. Gender				
Male	769	39.8	4	17.4
Female	1,155	60.2	19	82.6
2. Age				
15-20	1,068	55.5		
20-25	635	33.0		
25-30	203	10.6		
30-40	14	0.7		
40-50	2	0.1		
>50	2	0.1		
3. Rank				
Full Professor			4	17.1
Associate Professor			4	17.1
Assistant Professor			7	30.8
Adjunct (Full Time)			2	9.0
Adjunct (Part Time)			6	26.0
4. Highest Degree				
Doctorate			16	69.0
Master’s			7	31.0
5. Institution Type				
Baccalaureate Colleges	1,153	59.9	12	52.3
Master’s Colleges	536	27.9	7	30.3
Doctoral Universities	235	12.2	4	17.4

Hierarchical Linear Model – Unconditional Model

The SATS project data were structured in a way that students were nested within course sections; course sections were nested within instructors; and instructors were nested within institutions. Preliminary examination revealed that section-to-section variability in students' statistics attitudes within instructors was very small so as to be negligible. In addition, institutions might not be viewed as exchangeable so their effects on students' statistics attitudes were not appropriate to be modeled as random. Thus, a two-level model with instructor being the higher level was optimal in terms of both model sophistication and parsimony.

The first step was to determine whether or not students taught by one specific instructor responded to the items on the SATS-36 in a similar manner, a process commonly observed in group contexts. If this were the case, then there exists a hierarchical structure that would give rise to the multilevel data. This can be achieved by comparing the one-way ANOVA model specified with random intercept (also known as the null or unconditional model) to another one-way ANOVA model with no random intercept (Peters, 2013), for each of the six attitude component. Chi-square difference test was used to determine the degree of unexplained variance in students' statistics attitudes between instructors. Large Chi-square value would indicate the presence of unexplained variance at the instructor level. In addition, the larger unexplained between-instructor variance is, the more similarly students taught by one instructor responded to the SATS-36 items. Single-level models overlook this structure of dependencies among students and are therefore deemed inappropriate when it comes to the analysis of the SATS project data. Neglecting the instructor-level variation would lead to biased estimates of regression parameters as well as erroneous standard errors.

Results from chi-square difference tests suggested that there was a considerable amount of unexplained variance between instructors for all attitude components except *Effort*. This finding indicated the need of removing *Effort* from the subsequent analysis. In subsequent subsections, the results were presented individually regarding the hypothesis tests of the presence of between-instructor variation in each attitudinal dimension.

Interest

Findings from Chi-square difference test indicated that unexplained variation existed in *Interest* between instructors ($\chi^2 = 212.3$, $df = 1$, $p < 0.001$). The unconditional ICC, or the ratio of between-instructor variance to total variance, was 0.196, indicating differences in instructors account for 19.6% of the total variation in *Interest*.

Affect

Findings from Chi-square difference test indicated that unexplained variation existed in *Affect* between instructors ($\chi^2 = 208.6$, $df = 1$, $p < 0.001$). The unconditional ICC, or the ratio of between-instructor variance to total variance, was 0.211, indicating differences in instructors account for 21.1% of the total variation in *Affect*.

Cognitive Competency

Findings from Chi-square difference test indicated that unexplained variation existed in *Cognitive Competency* between instructors ($\chi^2 = 101.1$, $df = 1$, $p < 0.001$). The unconditional ICC, or the ratio of between-instructor variance to total variance, was 0.093, indicating differences in instructors account for 9.3% of the total variance in *Cognitive Competency*.

Difficulty

Findings from Chi-square difference test indicated that unexplained variation existed in *Difficulty* between instructors ($\chi^2 = 96.3$, $df = 1$, $p < 0.001$). The unconditional

ICC, or the ratio of between-instructor variance to total variance, was 0.099, indicating differences in instructors account for 9.9% of the total variation in *Difficulty*.

Value

Findings from Chi-square difference test indicated that unexplained variation existed in *Value* between instructors ($\chi^2 = 223.5$, $df = 1$, $p < 0.001$). The unconditional ICC, or the ratio of between-instructor variance to total variance, was 0.251, indicating differences in instructors account for 25.1% of the total variation in *Value*.

Effort

Findings from Chi-square difference test indicated that unexplained variation existed in *Effort* between instructors ($\chi^2 = 52.1$, $df = 1$, $p < 0.001$). The unconditional ICC, or the ratio of between-instructor variance to total variance, was 0.034, indicating differences in instructors account for 3.4% of the total variation in *Effort*.

Statistics Attitude Scores across Instructors

To corroborate the previous findings regarding the presence of large instructor-level variability, scores on statistics attitude across instructors were visualized for each subscale. Figures 4.1 and 4.2 present the post-attitude scores and change scores across instructors. A visual examination of the boxplots immediately revealed that both post-attitudes and changes in attitudes varied greatly across instructors. It also made a clear case that the *Effort* scores exhibited very small variability and the ceiling effects. Naturally, one of the goals for employing multilevel covariate adjustment model in this study was to see what instructor-specific attributes might account for some of that instructor-to-instructor variability in students' statistics attitudes at the end of the courses. Results indicated that there was a considerable amount of variability in each post-course attitude scores (expect *Effort*) to be productively modeled as a function of instructor-specific attributes.

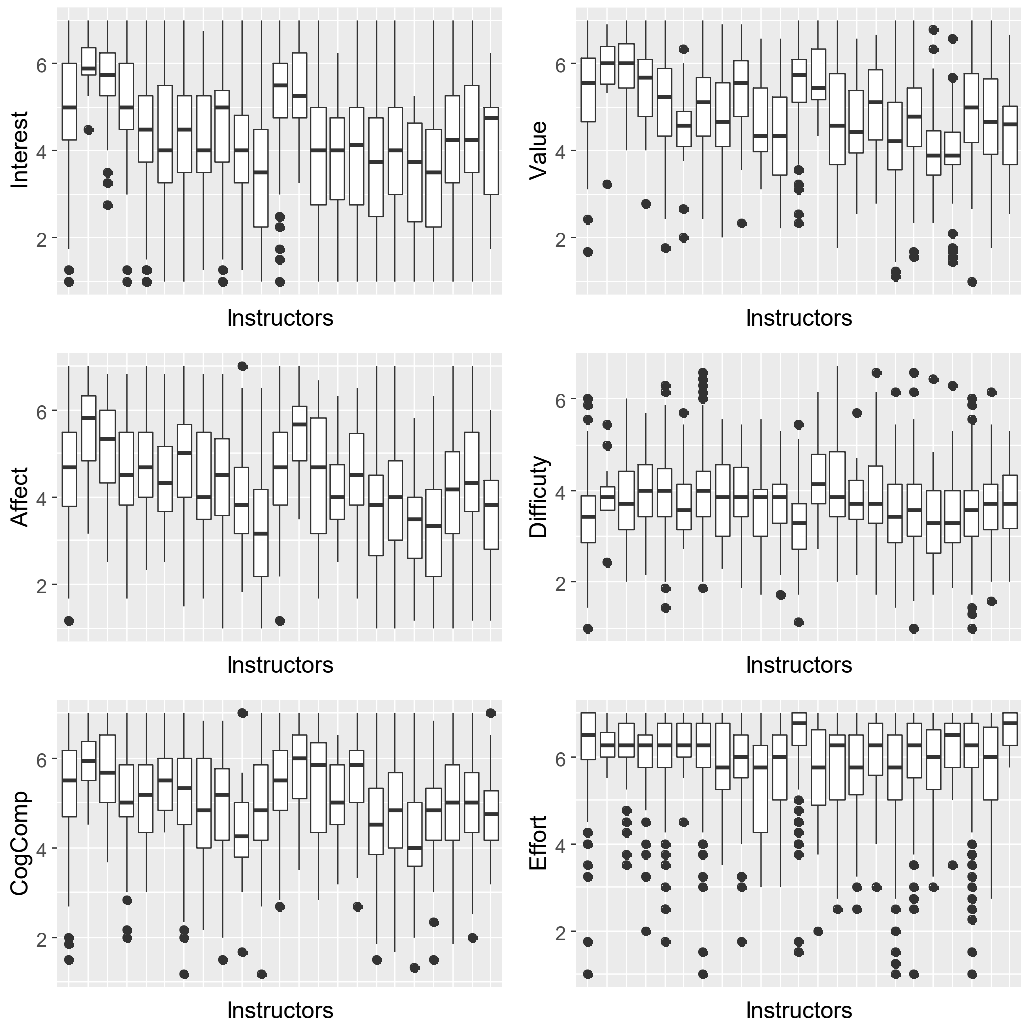

Figure 4.1 Post-Attitude Subscale Scores across Instructors

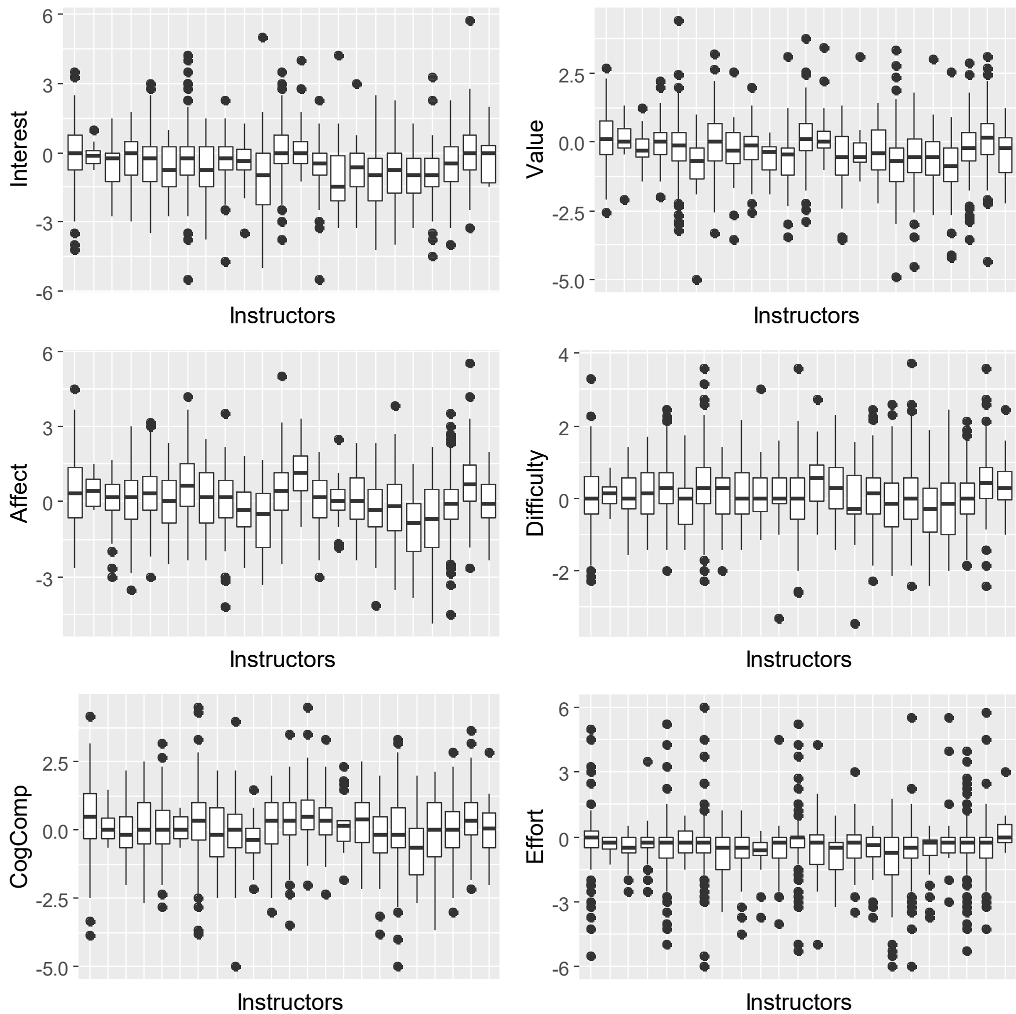

Figure 4.2 Changes in Attitude Subscale Scores across Instructors

Nonrandom Sorting across Institution Types

The final sample were obtained by pooling data across different type of institutions. To draw valid conclusions from the current models, one of the most standard and basic assumptions is that instructor-level random effects must not be conditional on institution types (McCaffrey et al., 2004). In other words, implicit in the multilevel covariate adjustment model is the core assumption that instructors must be exchangeable among schools. Empirically, this is a tenable assumption. The differences in instructor quality across the participating post-secondary institutions in the SATS project were considered negligible in the sense that an instructor who worked at one school was qualified to teach entry-level introductory statistics courses at any of the other schools (C. Schau, personal communication, September 20, 2018). This claim was essential to this study as it was related to systematic sorting based on students' prior statistics attitudes.

To provide additional evidence on the appropriateness of pooling data across different institution types, an assumption was checked against as to whether nonrandom sorting of students to different type of institutions had caused detectable associations between institutional types and students' prior statistics attitudes. Systematic sorting mechanism in relation to pre-attitude scores would biases the estimates of instructor and instructional effects on students' statistics attitudes if it occurs (McCaffrey et al., 2004).

Figure 4.3 presents the boxplots of pre-attitudes scores for each instructor by institution types, and shows that average pre-attitude scores were not systematically high or low for instructors at any of the three institution types. Specifically, the instructors at any of the three types of institutions did not always receive students with highest (or lowest) statistics attitudes in their classes. The results indicate that any biases due to nonrandom sorting based on pre-attitude scores may be limited for this study, thereby justifying the inclusion of a school type fixed effect ω_s in Equation 1.

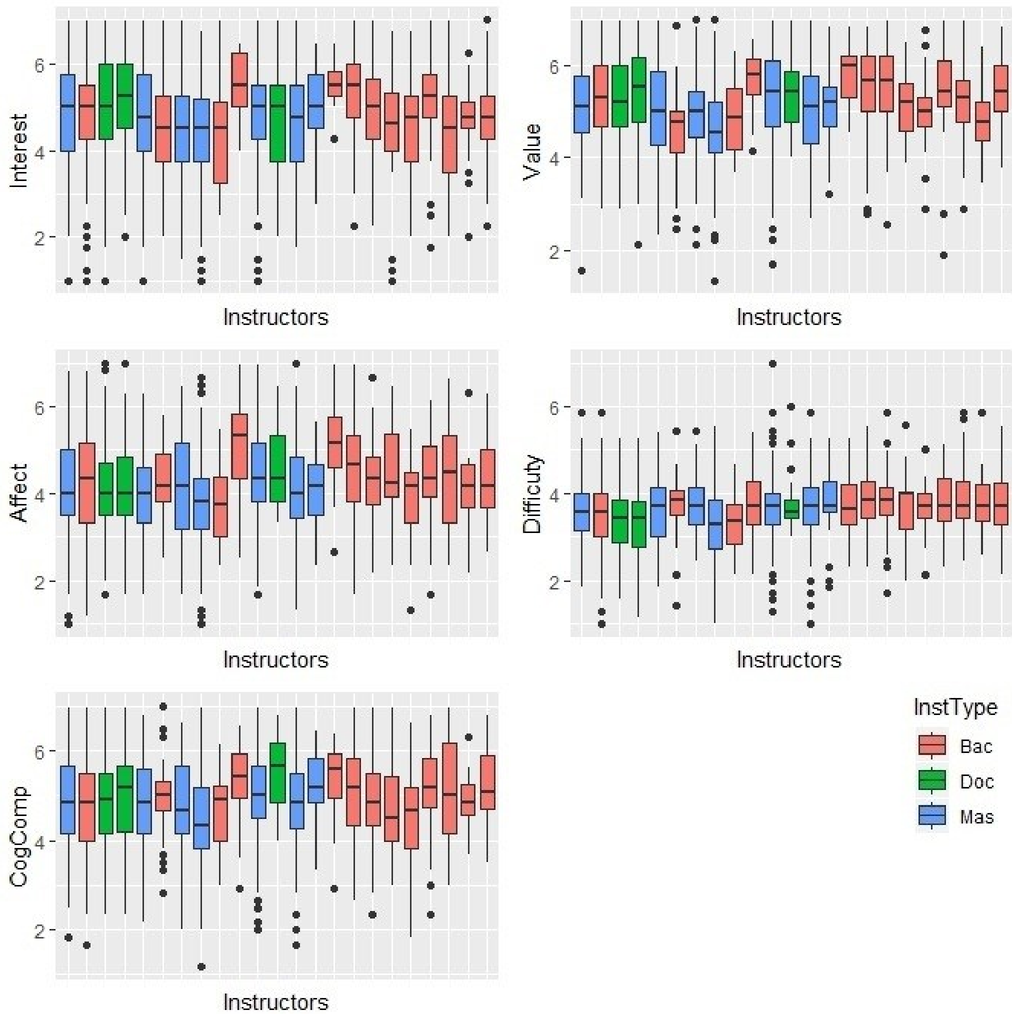

Figure 4.3 Pre-Attitude Subscale Scores for Each Instructor by Institution Types

Instructor-Specific Measure

The instructor-specific measure was created from the ten items on the instructor questionnaire (see Table 3.2). The two attitude items were kept separate from the rest of the items as well as from each other. The *Gen_Attitudes* item purported to measure instructors' attitudes toward introductory statistics course, as a reflection of their beliefs about what to teach in the introductory courses and how to teach these types of courses. On the other hand, the *Spe_Attitudes* item purported to measure instructor's attitudes toward teaching their own course sections. Including instructors' attitudes in this study as independent variables is based on the premise that effective teachers are expected to possess a wide range of teaching skills, both cognitive and non-cognitive, to such a degree that teachers continually enhance students' motivation to learn (Cheng & Zamarro, 2018). The two attitude items, by definition, measured different yet interrelated dimensions of teaching competencies. This claim was further corroborated by a moderate pairwise correlation between instructors' responses to the two items ($r = 0.42$, $p < 0.05$, $r^2 = 0.18$).

Table 4.2

Descriptive Statistics for Instructor-Specific Measures

	Univariate statistics		Pearson correlations				
	Median	MAD	1	2	3	4	5
1. *Gen_Attitudes*	6.75	0.36	1.00				
2. *Spe_Attitudes*	6.00	0.72	0.42*	1.00			
3. *Concept*	4.25	0.62	0.41*	0.28	1.00		
4. *Data*	4.74	0.38	0.47*	0.21	0.56*	1.00	
5. *Assessment*	3.71	1.05	-0.11	-0.19	0.21	0.15	1.00

Note. **$p < 0.01$, *$p < 0.05$. MAD stands for median absolute deviation.

The additional eight items were explored by first classifying them into three domains simply at face value. Of the eight items, two contained the key word “data” in the naming and purported to measure how well instructors integrated data collection and analysis into their classes; two contained the key word “evaluation” and were thus intended to measure how often instructors used assessment methods as a means of teaching and improving student learning; the other four measured the degree to which instructors deepened students’ conceptual understanding of statistics possibly through placing emphases on statistical thinking and/or implementing active learning approaches. The respective items were then combined to create a final score for each domain by averaging the raw responses across the respective set of items (see the Domain column in Table 3.2).

The instructor-specific measure has strong substantive and content validity as they align with the six recommendations put forward in the GAISE report (p. 6). Specifically, the *Concept* domain corresponds to the GAISE recommendation 1, 2, and 4; the *Data* domain mirrors the recommendation 3 and 5; the *Assessment* domain aligns with the recommendation 6. However, information on the factorial validity of the instructor-specific measures was unavailable because the factor analysis model failed to converge perhaps due to the small number of participating instructors. Table 4.2 presents summary statistics for the instructor-specific measurements. Due to extreme skewness and kurtosis, the scores were summarized using median and median absolute deviation (MAD), the robust equivalents of mean and standard deviation.

The pairwise correlation between the *Concept* and *Data* domains was 0.56 ($p < 0.01$, $r^2 = 0.314$), suggesting that *Concept* and *Data* may represent two distinct dimensions of instructional practice but were also intertwined to a great degree (Neumann et al., 2013). This finding can be viewed as a faithful reflection of the interactions among the GAISE recommendations for instructors. For example, the entry "Use technology to explore concepts and analyze data" evidently bifurcates into both the *Concept* and *Data* realms. So does the recommendation "Foster active learning." It advocates instructional strategies that actively engage students in data generating process, which is key to the understanding of statistical concepts. The rest of the pairwise correlations are weak to moderate (see Table 4.2). As a result, it was reasonable to posit that the instructor-specific measures have captured five unique dimensions of statistics instruction which were later used as instructor-specific variables in multilevel covariate adjustment models.

Research Question One

Research question one, *Does taking introductory statistics courses change students' statistics attitudes?*, was answered by comparing pre-attitude mean scores and post-attitude mean scores. Paired t-test was used to determine if there were statistically significant differences between mean scores in attitude components measured at pretest and posttest. Table 4.3 presents the results of mean pre-post responses and hypothesis testing.

Table 4.3

Mean Pre-Post Responses on Attitude Subscales (n = 1,924)

Subscale	Pretest		Posttest			
	M	*SD*	*M*	*SD*	t-value	p-value
1. *Effort*	6.39	0.93	5.96	1.06	13.128	< 0.001*
2. *CogComp*	4.85	1.03	4.94	1.13	-4.028	< 0.001*
3. *Affect*	4.17	1.10	4.28	1.33	-3.849	< 0.001*
4. *Difficulty*	3.62	0.77	3.75	0.93	-4.668	< 0.001*
5. *Value*	5.09	0.99	4.83	1.11	6.011	< 0.001*
6. *Interest*	4.69	1.22	4.18	1.42	10.128	< 0.001*

Note. *$p < 0.001$

Upon examining the actual changes cores, *Affect*, *Cognitive Competence*, and *Difficulty* remained about the same from pretest to posttest, despite that results from paired t-test showed statistically significant differences. This phenomenon is caused by the large number of participating students (n = 1,924). Paired t-test with large sample always rejects null hypothesis in terms of p-value in spite of how small the difference is. The *Value* scores show a decrease by 0.26 point ($p < 0.001$). The mean change scores for *Interest* and *Effort* decreased by about 0.5 point ($p < 0.001$), suggesting an important decrease that may be meaningful in evaluating attitudinal changes induced by instructional interventions.

Interest

Findings suggested that there was a statistically significant mean difference between the average *Interest* scores measured at the end of the courses and those measured at the beginning, $t(1923) = 10.128$, $p = < 0.001$, $d = 0.19$ (small effect size), $r^2 = 0.036$. Students' interest in statistics dropped significantly at the end of introductory courses (M = 4.18) compared to their interest in statistics at the beginning (M = 4.69). The r^2 value of 0.036 indicated that only approximately 3.6% of the variance in the decrease of students' interest in statistics can be attributed to taking introductory statistics courses.

Affect

Findings suggested that there was a statistically significant mean difference between the average *Affect* scores measured at the end of the courses and those measured at the beginning, $t(1923) = -3.849$, $p = < 0.001$, $d = 0.07$ (small effect size), $r^2 = 0.005$. Students have made significant gain in their feeling about statistics at the end of introductory statistics courses (M = 4.28) compared to their feeling at the beginning (M = 4.17). The r^2 value of 0.005 indicated that only approximately 0.5% of the variance in the gain of students' feeling about statistics can be attributed to taking introductory statistics courses.

Cognitive Competency

Findings suggested that there was a statistically significant mean difference between the average *Cognitive Competency* scores measured at the end of the courses and those measured at the beginning, $t(1923) = -4.028$, $p = < 0.001$, $d = 0.08$ (small effect size), $r^2 = 0.006$. Students have made significant gain in the confidence in their ability to cope with statistics at the end of introductory courses (M = 4.94) compared to their confidence at the beginning (M = 4.85). The r^2 value of 0.006 indicated that only approximately 0.6% of the variance in the gain of students' confidence in their ability to cope with statistics can be attributed to taking introductory statistics courses.

Difficulty

Findings suggested that there was a statistically significant mean difference between the average *Difficulty* scores measured at the end of the courses and those measured at the beginning, $t(1923) = -4.668$, $p = < 0.001$, $d = 0.09$ (small effect size), $r^2 = 0.008$. Students have made significant gain in the perception of how easy it is to learn statistics at the end of introductory courses (M = 3.75) compared to their perception at the beginning (M = 3.62). The r^2 value of 0.008 indicates that only approximately 0.8% of the variance in the gain of students' perception of how easy it is to learn statistics can be attributed to taking introductory statistics courses.

Value

Findings suggested that there was a statistically significant mean difference between the average *Value* scores measured at the end of the courses and those measured at the beginning, $t(1923) = 6.011$, $p = < 0.001$, $d = 0.12$ (small effect size), $r^2 = 0.015$. Students' perception of the utilitarian value in statistics dropped significantly at the end of introductory courses (M = 4.83) compared to their perception at the beginning (M = 5.09). The r^2 value of 0.015 indicates that only approximately 1.5% of the variance in the decrease of students' perception of the utilitarian value in statistics can be attributed to taking introductory statistics courses.

Effort

Findings suggested that there was a statistically significant mean difference between the average *Effort* scores measured at the end of the courses and those measured at the beginning, $t(1923) = 13.128$, $p = < 0.001$, $d = 0.29$ (small effect size), $r^2 = 0.089$. Students' willingness to expend effort to study statistics dropped significantly at the end of introductory courses (M = 5.96) compared to their willingness at the beginning (M = 6.39). The r^2 value of 0.089 indicates that only approximately 8.9% of the variance in the decrease of students' willingness to expend much effort to study statistics can be attributed to taking introductory statistics courses.

Research Question Two

Research question two, *Is there an unexplained variation across statistics instructors regarding students' statistics attitudes?*, was answered by using multilevel models. Table 4.4 reports three metrics – instructor-level sample variability for each subscale, conditional ICCs and the standard deviations of instructor-level variance components (i.e., instructor effects on each component of students' statistics attitudes). The latter two metrics were produced directly by HLM.

The results suggested that there was a moderate to considerable amount of instructor-level variability in post-attitude scores, ranging between 0.24 and 0.63 *SD*s. The lowest variability was seen in *Effort* while the highest was in *Interest.* Notably, the instructor-to-instructor variability was substantially greater in each of the six attitude subscale scores at the posttest than at the pretest. Thus, it is possible to obtain estimates of instructor effects on students' statistics attitudes through harnessing the substantive variability in post-attitudes scores.

Table 4.4

Unexplained Variability in Instructors' Contributions to Students' Statistics Attitudes

Subscale	Pre instructor-level mean score and *SD*	Post instructor-level mean score and *SD*	*SD* of instructor-level variance	Conditional ICC
1. *Interest*	4.81 (0.34)	4.37 (0.63)	0.30	0.14
2. *Affect*	4.30 (0.35)	4.32 (0.61)	0.36	0.16
3. *CogComp*	4.96 (0.25)	5.05 (0.39)	0.19	0.06
4. *Difficulty*	3.65 (0.19)	3.76 (0.25)	0.18	0.06
5. *Value*	5.21 (0.32)	4.91 (0.54)	0.34	0.20
6. *Effort*	6.38 (0.19)	5.91 (0.24)	0.09	0.02
7. *exgrade*	3.35 (0.08)	2.93 (0.28)	0.28	0.12

Note. Statistically significant ($p < 0.05$)

Conditional ICCs followed a similar pattern. The ICC value for *Effort* is very low with it slightly under 0.02. A low ICC smaller than conventionally acceptable values (0.05) indicated that there were not enough systematic differences between instructors in their contributions to the *Effort* outcome. Technically, the small differences might not be productively modeled as a function of instructor-level variables. As a result, the *Effort* outcome data were removed from the subsequent analysis concerning instructional effects on students' statistics attitudes (and also omitted from Figure 4.3). Moreover, the results from instructor effect estimates suggested that instructors who were 1 *SD* above the average on the distribution of instructor effectiveness increased *Interest* by 0.30 *SD*, *Affect* by 0.36 *SD*, *Difficulty* by 0.18 *SD*, *CogComp* by 0.19 *SD*, and *Value* by 0.34 *SD* over one introductory statistics course. Given the sizable instructor effects observed on students' statistics attitudes, the next step is thus to seek possible explanations as to what instructional dimensions potentially account for these effects.

Lastly, results showed that students' expected course grades dropped from 3.35 at the beginning of the courses to 2.93 toward the end. However, the variability in instructor-associated contributions to expected course grades increased drastically toward the end of the courses, from 0.08 to 0.28. These findings provided sufficient evidence to substantiate a large degree of instructor-associated changes in expected course grades. Both standard deviation of the instructor-level variance (0.28) and ICC (0.12) suggested that instructors also had a large impact on students' expectation of the final grades they were to receive.

Research Question Three

Research question three, *Do instructors' attitudes toward teaching predict students' statistics attitudes at the end of introductory statistics course?*, was answered by employing multilevel models. Specifically, instructors' attitudes toward introductory statistics courses (*Gen_Attitudes*) and attitudes toward teaching their specific course sections (*Spe_Attitudes*) served as the instructor-specific variables. Table 4.5 presents both the standardized and unstandardized coefficient estimates.

The results showed two statistically significant relationships. Instructors' attitudes toward teaching their own sections (*Spe_Attitudes*) were found to be positively associated with students' interest in statistics ($\beta = 0.32$, $p < 0.05$) and their feeling toward statistics ($\beta = 0.28$, $p < 0.05$), yet not with other attitude components (here β is unstandardized). Specifically, for an instructor who had low attitudes toward teaching their own classes, one unit increase in their attitudes would lead to a 0.32 or 0.28 unit increase in average *Interest* or *Affect* score among students taught by this instructor. In terms of standardized coefficient estimates, a 1 *SD* increase in instructors' attitudes toward teaching classes was associated with a 0.18 *SD* increase in students' interest in statistics and a 0.16 *SD* increase in students' feeling about statistics. Note that unstandardized coefficients made more sense than their standardized counterparts in this study. If data on the independent variables are highly skewed, standardization procedure might cause loss of important information.

Table 4.5

Effects of Instructors' Attitudes on Students' Statistics Attitudes

	Interest	*Difficulty*	*Value*	*CogComp*	*Affect*
Panel A: Unstandardized β					
Gen_Attitudes	-0.25	-0.10	-0.17	0.01	-0.16
	(0.16)	(0.08)	(0.16)	(0.09)	(0.18)
Spe_Attitudes	0.32*	0.06	0.23	0.08	0.28*
	(0.11)	(0.06)	(0.11)	(0.07)	(0.13)
Panel B: Standardized β					
Gen_Attitudes	-0.11	-0.06	-0.09	-0.01	-0.07
	(0.05)	(0.05)	(0.09)	(0.06)	(0.09)
Spe_Attitudes	0.18*	0.05	0.15	0.03	0.16*
	(0.05)	(0.05)	(0.07)	(0.05)	(0.07)

Note. *$p < 0.05$

On the other hand, there were no statistically significant relationships between instructors' attitudes toward teaching classes and the *Value*, *Difficulty* and *Cognitive Competency* components ($p > 0.05$). In other words, there was insufficient evidence to suggest that instructors' attitudes toward teaching their own classes had any impact on any of the three statistics attitude components. As a point of comparison, instructors' attitudes toward introductory statistics courses (*Gen_Attitudes*) were not significantly associated with any of the five attitude components.

Provided that *p*-value alone is not sufficient to glean inferential information from statistical analysis, Cohen's f^2 was also computed to examine the local effect sizes of individual instructor-specific variables. Cohen's f^2 for the effects of instructors' attitudes toward teaching their course sections on *Affect* and *Interest* were 0.03 and 0.02, respectively. These values corresponded to small size of effect (i.e., $f^2 = 0.02$) according to Cohen's conventions on effect sizes.

Research Question Four

Research question four, Does instructors' emphasis on data contexts predict students' statistics attitudes at the end of introductory statistics course?, was answered by employing multilevel models. Specifically, instructors' emphasis on data contexts (data) served as the instructor-specific variable. Table 4.6 presents both the standardized and unstandardized coefficient estimates.

Table 4.6

Effects of Instructors' Emphasis on Data Contexts on Students' Statistics Attitudes

	Interest	*Difficulty*	*Value*	*CogComp*	*Affect*
Panel A: Unstandardized β					
Data	0.27	0.14	0.32*	0.15	0.40*
	(0.14)	(0.07)	(0.15)	(0.08)	(0.17)
Panel B: Standardized β					
Data	0.15	0.13	0.24*	0.06	0.25*
	(0.08)	(0.06)	(0.11)	(0.07)	0.11)

Note. *$p < 0.05$

The extent of instructors' emphasis on data collection and analysis in teaching introductory statistics was positively associated with *Value* ($\beta = 0.32$, $p < 0.05$) and *Affect* ($\beta = 0.40$, $p < 0.05$). Specifically, after experiencing instructors who emphasized data collection and analysis with meaningful contexts, students on average tended to have more positive feeling toward statistics and were also more likely to appreciate the relevance of statistics to both their personal and professional lives. For individual instructors who had little emphasis on data contexts, one unit increase in their emphasis would lead to a 0.32 or 0.40 unit increase in average *Value* or *Affect* score among students taught by this instructor. In terms of standardized coefficient estimates, a 1 SD increase in instructors' emphasis on data contexts is associated with a 0.24 SD increase in students' appreciating the worth of statistics and a 0.25 SD increase in students' positive feeling about statistics. In contrast, there was no statistically significant relationships between instructors' emphasis on data contexts and *Interest*, *Difficulty* or *Cognitive Competency* ($p > 0.05$). In other words, there was insufficient evidence to suggest that instructors' emphasis on data contexts had any impact on any of the three attitude components. Cohen's f^2 for the effects of the *Data* domain on *Affect* and *Value* are both about 0.02. These values correspond to small size of effect (i.e., $f^2 = 0.02$).

Research Question Five

Research question five, *Does instructors' emphasis on assessment predict students' statistics attitudes at the end of introductory statistics course?*, was answered by employing multilevel models. Specifically, instructors' emphasis on assessment methods (*assessment*) served as the instructor-specific variable. Table 4.7 presents both the standardized and unstandardized coefficient estimates.

Table 4.7

Effects of Instructors' Emphasis on Assessment on Students' Statistics Attitudes

	Interest	*Difficulty*	*Value*	*CogComp*	*Affect*
Panel A: Unstandardized β					
Assessment	0.16	0.07	0.11	0.04	0.26
	(0.13)	(0.07)	(0.14)	(0.08)	(0.16)
Panel B: Standardized β					
Assessment	0.07	0.05	0.07	0.01	0.13
	(0.06)	(0.05)	(0.08)	(0.05)	(0.08)

Note. *$p < 0.05$

None of the five regression coefficients were statistically significant. Findings suggested that there was no statistically significant relationship between instructors' emphasis on assessment methods and any of the five statistics attitudes component ($p > 0.05$). In other words, there was insufficient evidence to suggest that instructors' emphasis on assessment methods had any impact on any of the statistics attitude component.

Research Question Six

Research question six, Does instructors' emphasis on conceptual understanding predict students' statistics attitudes at the end of introductory statistics course?, was answered by employing multilevel models. Specifically, instructors' emphasis on conceptual understanding of statistics (concept) served as the instructor-specific variable. Table 4.8 presents both the standardized and unstandardized estimates of regression coefficients.

Table 4.8

Effects of Instructors' Emphasis on Conceptual Understanding on Statistics Attitudes

	Interest	*Difficulty*	*Value*	*CogComp*	*Affect*
Panel A: Unstandardized β					
Concept	-0.07	0.02	-0.14	0.04	-0.15
	(0.14)	(0.08)	(0.25)	(0.09)	(0.17)
Panel B: Standardized β					
Concept	-0.03	0.02	0.10	0.11	0.10
	(0.08)	(0.07)	(0.11)	(0.07)	(0.11)

Note. $*p < 0.05$

None of the five regression coefficients were statistically significant. Findings suggested that there was no statistically significant relationship between instructors' emphasis on conceptual understanding and any of the five statistics attitudes components

($p > 0.05$). In other words, there was insufficient evidence to suggest that instructors' emphasis on conceptual understanding had any impact on any of the five attitude components. These findings have significant implications for instructional and curricular development in current statistics education.

Research Question Seven

Research question seven, *Do instructors who improve students' statistics attitudes also improve expected course grade?*, was answered by the correlation analysis. BLUP estimates of instructor-level random effects were extracted to capture individual instructors' contributions to students' statistics attitudes and their contributions to students' expected course grades. The correlation was then examined between BLUP estimate derived for each of the five attitude components and that derived for expected course grades. Scatterplots in Figure 4.4 illustrates the extent to which instructor-associated contributions to different student outcomes co-varied. In subsequent subsections, the correlation results were presented individually with respect to the hypothesis testing of the presence of correlations. The findings strongly suggested that most statistics instructors produced concurrent increase/decrease in students' expected grade and statistics attitudes.

Interest

Results indicated that there was a statistically significant positive relationship between instructor-associated contributions to expected course grades and their contributions to students' interest in statistics ($r = 0.641$, $p < 0.001$, $r^2 = 0.410$). Therefore, instructors who were effective at improving students' interest in statistics also effectively improved expected course grades. Approximately 41.0% of variation in instructor-associated contributions to expected grades can be attributed to the instructors' contributions to students' development of their interest in statistics.

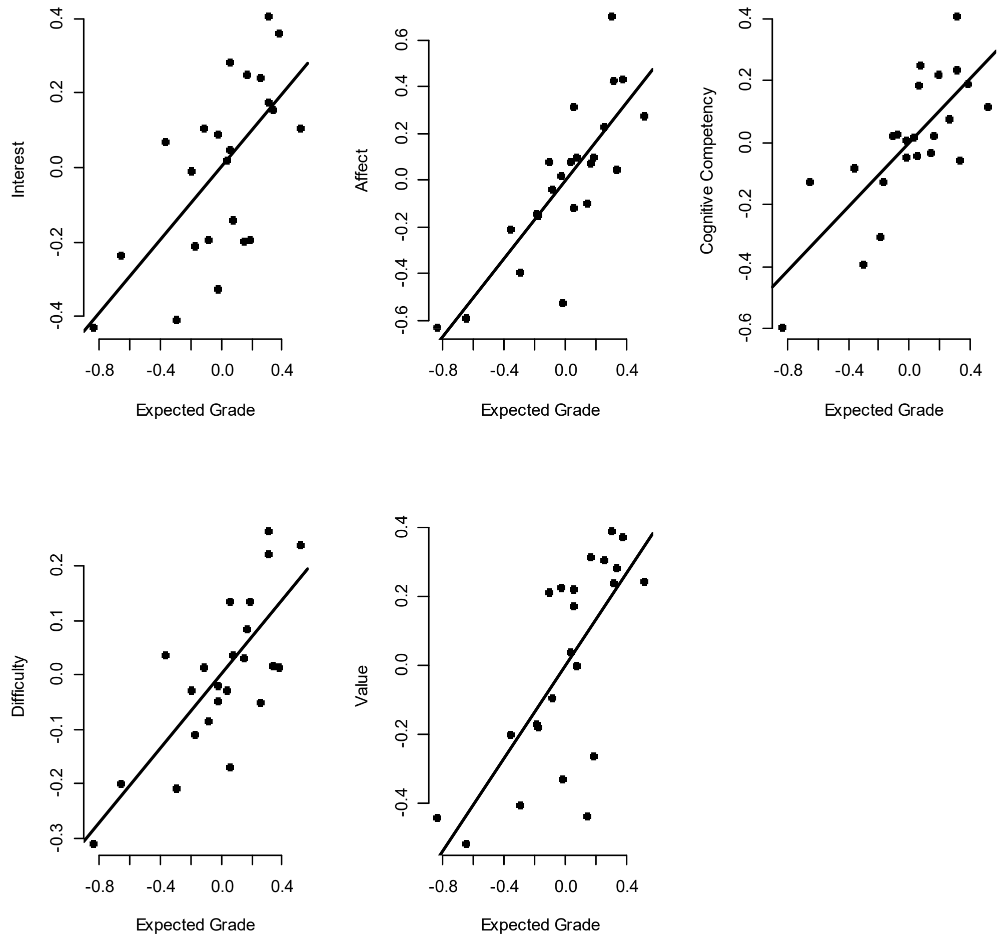

Figure 4.4 Instructor Effects on Statistics Attitudes and Expected Grade

Cognitive Competency

Results indicated that there was a statistically significant positive relationship between instructor-associated contributions to expected course grades and their contributions to students' confidence in learning about statistics ($r = 0.788$, $p < 0.001$, $r^2 = 0.621$). Therefore, instructors who were effective at improving students' confidence in learning and handling statistics also effectively improved their expected course grades. Approximately 62.1% of variation in instructor-associated contributions to expected grades can be attributed to the instructors' contributions to students' development of their confidence in learning and handling statistics.

Difficulty

Results indicated that there was a statistically significant positive relationship between instructor-associated contributions to expected course grades and their contributions to students' perception of the difficulty level of statistics ($r = 0.812$, $p < 0.001$, $r^2 = 0.659$). Therefore, instructors who were effective at improving students' perception of the difficulty level of statistics also effectively improved their expected course grades. Approximately 65.9% of variation in instructor-associated contributions to expected grades can be attributed to the instructors' ability to improve students' perception of the difficulty level of statistics.

Value

Results indicated that there was a statistically significant positive relationship between instructor-associated contributions to expected course grades and their contributions to students' perception of the relevance of statistics to both personal and professional lives ($r = 0.812$, $p < 0.001$, $r^2 = 0.659$). Therefore, instructors who were effective at improving students' perception of the utilitarian value in statistics also effectively improved their expected course grades. Approximately 65.9% of variation in instructor-associated contributions to expected grades can be attributed to the instructors' ability to improve students' perception of the relevance of statistics to both personal and professional lives.

Affect

Results indicated that there was a statistically significant positive relationship between instructor-associated contributions to expected course grades and their contributions to students' feeling about statistics ($r = 0.691$, $p < 0.001$, $r^2 = 0.478$).

Therefore, instructors who were effective at improving students' feeling about statistics also effectively improved expected course grades. Approximately 47.8% of variation in instructor-associated contributions to expected grades can be attributed to the instructors' contributions to students' development of their feeling about statistics.

Conclusion

This chapter presented the results from the multilevel analysis of the quantitative data collected using the SATS-36 along with an instructor survey consisting of ten questions. In the next chapter, the findings will be summarized. Moreover, the implications of the findings for statistics education and educational policy will be discussed. Lastly, recommendation for future studies will be provided.

CHAPTER V:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

The purpose of this study was to examine instructor and instructional effects on students' statistics attitudes. A long-standing hypothesis in the field of statistics education has posited that instructors could have a large impact on students' statistics attitudes (Schau, 2003). Schau showed that the instructor-level variability in average attitude scores was considerably larger at the end of the courses than at the beginning, and further identified instructor and instructional attributes as one determinant of students' statistics attitudes. If instructors mattered as much in impacting students' statistics attitudes as this study suggested, the critical question then becomes "to what extent do instructors affect students' statistics attitudes?" To date, there existed very little, if any, evidence related to this issue. In addition, if instructors do have a large impact on students' statistics attitudes, then it is possible to identify certain dimensions of instructional practice that (at least partly) account for this impact.

The present study was intended to close this gap. To quantify instructor and instructional effects on students' statistics attitudes in a rigorous manner, this study drew on a rich set of data collected from 1,924 students enrolled in introductory statistics courses across 11 U.S. post-secondary institutions. The students completed the SATS-36, a well-studied measure of students' statistics attitudes. That 1,924 students were clustered within 23 instructors gave rise to a multilevel structure of the SATS project data, which in turn prompted the researcher to use multilevel modeling approach in testing the foregoing hypothesis.

Additionally, 23 instructors completed a pool of questions on a separate questionnaire. Ten questions were selected to form an instructor-specific measure of teaching competencies in statistics classrooms. Eight of the ten questions were created in line with three domains that were deemed essential since the early 1990s and had held up over the years (GAISE College Report ASA Revision Committee, 2016). The three domains consist of conceptual understanding, data collection and analysis within proper contexts, and assessment methods. The other two questions depict instructors' attitudes toward introductory statistics courses and instructors' attitudes toward teaching specific course sections, respectively. Results from the correlation analysis of the five instructional domains suggested that instructors' responses to the ten questions might have captured five unique yet interrelated dimensions of instructional practice specifically in statistics classrooms.

In this chapter, the findings of this study were summarized in the context of the current state of knowledge in statistics education. Implications for statistics instruction and educational policy in higher education were discussed. Several recommendations were provided for future research. Issues, challenges and opportunities were outlined on the horizon for statistics education research.

Summary

Research Question One

The pre-attitude scores varied considerably across attitude components. Interpretations of the results suggested that, at the beginning of the introductory statistics courses, students had neutral feelings about statistics (*Affect*); they perceived statistics as a slightly challenging subject (*Difficulty*); they anticipated investing much effort to study statistics (*Effort*); they were interested in statistics yet only to a certain degree (*Interest*); they somewhat appreciated the utilitarian value of statistics in their personal and professional lives (*Value*); and they tended to have confidence in their ability to learn and handle statistics (*Cognitive Competency*).

As anticipated from the post-attitude scores, the change scores too varied across attitude components. Results from paired *t*-test indicated that there were statistically significant differences between students' statistics attitudes measured at the beginning and the end of the courses. However, this phenomenon was due to the large number of students included in this study (n = 1,924). The effect sizes of the actual changes in average scores were very small. Thus, not all of the differences are practically significant.

Schau and Emmioğlu (2012) deemed a one-half point increase/decrease as practically important in their attempt to guide evaluation of the changes in students' statistics attitudes induced by instructional interventions. Only the decreases in the *Effort* and *Interest* components met this identifying criterion. Both of them decreased by just about one-half point. The interpretation of these findings suggested that lecture-based introductory statistics courses did not improve students' statistics attitudes. In some cases, students' statistics attitudes even decreased considerably. As a result, there is a need to identify and validate instructional approaches that hold promises for improving students' statistics attitudes.

The findings on pre-post differences in attitude scores largely aligned with theories of motivation stating that students' academic attitudes, as a psychological trait, are recalcitrant to change over time unless educational interventions are timely, targeted and tailored (Cohen et al., 2017). A substantial body of evidence in the literature of statistics education also suggests that students' statistics attitudes are resistant to change even when moderate to intensive course interventions were employed (Lesser et al., 2016; Ramirez et al., 2012). Interestingly, Kerby and Wroughton (2017) found that course interventions might be more effective if they were applied to a period when students' statistics attitudes were on the rise, which often occurred during the first half of introductory statistics courses.

Research Question Two

Social sciences and education researchers are well aware that much of what goes on in educational research takes place in group contexts. A group can and does influence the thoughts and behaviors of its members through different sorts of collective mechanisms (Burstein, 1980). In the context of research on statistics attitudes, students linked with the same individual instructor shared very similar classroom experiences and therefore were more likely to develop highly similar statistics attitudes than those taught by different instructors (Schau, 2003). As anticipated, results from chi-square difference tests showed that there was a considerable amount of unexplained variation in students' statistics attitudes (expect *Effort*) across instructors. Accordingly, results from multilevel analysis suggested that instructors had a large impact on students' statistics attitudes as being manifested in three different metrics – instructor-to-instructor sample variability, conditional ICCs and standard deviation of the instructor-level random effect (i.e., instructor effects). In addition, the impact varied considerably across different attitude components. The largest impact was found on *Value*, *Interest*, and *Affect* whereas the smallest one was on *Effort*.

These findings are in line with Schau (2003) who first observed a large variability in average attitude scores across instructors at the end of the courses. That average post-course scores varied greatly could be viewed as an implicit indication of the presence of instructor effects on students' statistics attitudes. Waters et al. (1988) also hinted at the same possibility by showing that students linked with different instructors on average changed their statistics attitudes in either direction. However, it is hard to judge based on these two studies to what degree the empirical evidence supports the presence of instructor effects on students' statistics attitudes.

This study provided the first direct evidence that instructors varied considerably in their abilities to improve students' statistics attitudes. The magnitudes of instructor effects, ranging between 0.18 and 0.36 *SD*s, are generally in agreement with those from Blazar and Kraft (2017) where the researchers observed a sizable impact the upper-elementary school teachers had on students' academic attitudes and behaviors (0.14 to 0.31 *SD*s). That is, instructors with higher value-added estimates, *ceteris paribus*, improved students' statistics attitudes. Interpretation of the findings suggested that statistics attitudes (*Effort* excluded) for a student assigned to an average instructor might have been between 0.18 and 0.36 *SD*s higher, had he/she been assigned to an instructor 1.0 *SD* higher on the teacher effectiveness distribution. Such counterfactual estimations can be made only by using value-added models as opposed to other measurement models found in teacher accountability systems.

One of the two fundamental issues related to teacher effectiveness measurements concerns the consistency of teacher effects across school subjects (Leckie, 2018; Loeb, Soland, & Fox, 2014). Inconsistent teacher effects across subject areas indicate the necessity of studying those effects on individual subjects rather than the overall achievement. Blazar and Kraft (2017) reported that teachers differentially impacted students' mathematics and reading achievements, albeit to a lesser extent. This finding suggested the inconsistent teacher effectiveness across student learning outcomes. Thus, it spoke to the multidimensional nature of teaching skills. Note that instructor effects on students' achievement in statistics were not examined in this study. Unlike students' statistics attitudes and expected course grades that were measured twice in each semester, final grades for introductory statistics courses in the SATS project data were rather sub-optimally constructed for use in the current multilevel model.

The other fundamental issue concerns the stability of teacher effects across time for consecutive cohorts (Loeb & Candelaria, 2012). The structure of the SATS project data did not permit the researcher to address this issue either. Nonetheless, these two issues need further examination. Large year-to-year instability and/or subject-to-subject inconsistency in instructor effectiveness measurements would call for caution when decisions are made solely based on value-added measures.

In principle, a teacher value-added methodology can identify causal quantities only if students are as good as randomly assigned to their teachers. When data are observational, value-added models may still identify causal relationships yet only in hopes that some extreme assumptions are not violated (Rubin et al., 2004). Several researchers have laid out a list of statistical assumptions central to teacher value-added models (Harris, 2009; Reardon & Raudenbush, 2009). Scherrer (2011) evaluated those assumptions in the context of America's public schools and found that they were mostly not tenable in practice. However, it should be noted that estimations of teacher effects via modern statistical techniques are often very robust to certain amount of deviations from the standard assumptions underlying various value-added models. Therefore, the analysis conducted in Scherrer (2011) concerning the validity of value-added models was at best one-legged in that it was based solely on substantive matters and apparently bereft of any statistical considerations.

Given the observational nature of the SATS project data, the approaches used in this study relied exclusively on covariate adjustment to reduce potential biases assumed to arise from nonrandom sorting of students to different schools and instructors. Specifically, a set of independent variables (i.e., student-specific covariates, average covariates per section, school type fixed effects and pre-attitude scores) were added to the multilevel model in an attempt to account for potential sources of biases, an empirical strategy described in Blazar and Kraft (2017). Student socioeconomic characteristics were not being controlled for since they are unavailable in the SATS project data. Note that debates are ongoing as to whether many other predictors should be additionally adjusted for in multilevel models, depending ultimately on the particular types of those models that are of interest.

Biases due to student-teacher assignments may be of much less concern in this study than in studies conducted in the 12 school settings. Such assignments in colleges are often not as complex as the corresponding processes that actually take place in many K-12 schools. The latter often involves assessments of students' needs, parental requests, teacher specializations, and many other factors. Indeed, a preliminary examination of the estimates of instructor effects on students' statistics attitudes across institutions did not reveal that the instructors who were situated at the higher end of the instructor effectiveness distribution tended to teach at more affluent and advantaged institutions (e.g., private college).

Several recent studies provided strong empirical support to both the reliability and the validity of the present study. First, teacher effect estimates appear to be rather robust to many external factors related to students' behaviors, such as students' low test-taking effort in terms of their tendency for rapidly guessing on test items (Jensen, Rice, & Soland, 2018). The insusceptibility to low student effort in item responses is considered to be an important attribute of value-added models when it comes to evaluating the present study. In the stage of data collection, it would not strain the credibility to assume that a large number of students have responded to the SATS-36 in a rather effortless manner by quickly, if not randomly, selecting item responses. Therefore, the findings in Jensen et al. (2018) largely attest to the reliability of this study by eliminating the potential method effects assumed to arise from to rapidly guessed item responses.

Most critically, recent quasi-experimental and randomized analyses provided strong empirical support for the unbiasedness of teacher effect estimates even in the presence of systematic sorting mechanisms (Blazar, 2018; Chetty, Friedman, & Rockoff, 2014; Kane et al., 2013; Kraft, 2019). Put another way, multilevel covariate adjustment models can produce teacher effect estimates that are similar in magnitude to those obtained by randomization experiments. Findings from these studies attest to the validity of the present study to a great degree. However, it should be noted that randomized experiments in education research were often not conducted on sufficient scale. As Kane et al. (2013) pointed out, an important caveat in their randomization study was that many principals showed strong reluctance to assign students to teachers in a complete random fashion. As a result, the insufficient randomization might still not allow the researchers to estimate the teacher effectiveness distribution with desired clearness and precision.

Despite signs of optimism in a negligible impact exerted by nonrandom sorting mechanisms, the instructor effect estimates obtained in this study should be put under scrutiny in future studies. Measuring teacher effectiveness based on multilevel covariate adjustment models (i.e., value-added models) are highly controversial and often nettlesome. There is a contentious debate on the general validity of teacher value-added methodology concerning whether this methodology can accurately capture true teacher effects (e.g., Guarino, Reckase, & Wooldridge, 2015; Rothstein, 2009). Implicit in the core idea of teacher value-added methodology is that a comparison can be made between the average performance of a teacher’s students and the predictions of these students’ performance had they been assigned to an average teacher. If predictions are not accurate, nor are teacher effect estimates. Guarino et al. (2015) found none of those commonly used value-added estimation strategies could successfully recover true teacher effects from simulated achievement data. This issue points to the need of supplementing teacher effectiveness measures with other evaluation tools (Koedel, Mihaly, & Rockoff, 2015; Steinberg & Donaldson, 2016).

Multilevel covariate adjustment modeling is only one of several methodological tools commonly used to tease out teacher effects on various student outcomes. Other statistical approaches that can be used to estimate instructor effects on students' statistics attitudes for future studies include propensity scores and instrumental variables (Angrist, Imbens, & Rubin, 1996; Austin, 2011). It is noted that the gist of all the three foregoing approaches ties to estimation of causal qualities. In the context of the current study, students' statistics attitudes can be viewed as one potential outcome, as being well defined in the framework of the Rubin causal model (Imbens & Rubin, 2015). In this sense, instructor/teacher effects are equivalent to treatment effects (Rubin et al., 2004).

Research Question Three

By having identified instructors' attitudes toward teaching their course sections as a bellwether of students' statistics attitudes, this work helped further arguments for making instructor and instructional quality a central theme in higher education (Chingos, 2016; Umbach & Wawrzynski, 2005). Only very recently have researchers begun to understand what makes teachers differentially contribute to student learning (Blazar & Kraft, 2017), and learning of statistics in particular (Chance et al., 2016). Specifically, Blazar and Kraft (2017) found that teachers who were more willing to provide content-specific support produced students with higher levels of self-efficacy in mathematics and happiness in mathematics classrooms. In addition, Chance et al. (2016) found that the levels of experience that individual instructors have with computer simulation-based statistics curriculum seemed to predict students' conceptual understanding of statistics.

Central to statistics education are instructor and instructional practices (GAISE College Report ASA Revision Committee, 2016). Statistics can be a hard course, as it takes root in probability. Counter-intuitive phenomena are common when it comes to statistics (Cobb & Moore, 1997). Students may develop frustration over learning statistics when they observe these phenomena. When this occurs, statistics instructors are expected not only to deliver concepts with greater precision but also to provide more motivational support to students who struggle to learn statistics. In this regard, this study represents a unique contribution to the field of statistics education as it has identified the associations of instructors' attitudes toward teaching their own classes with students' interest in statistics and their feelings about statistics.

Several factors have been identified as sources of support for student learning in classroom. These factors consist of classroom management skills, classroom climate, and teachers' non-cognitive skills including their attitudes toward teaching (Korpershoek, Harms, de Boer, van Kuijk, & Doolaard, 2016). The connections among this triad provide great insights into the interpretations of the current findings. Instructors' attitudes toward teaching may be viewed as one of the pillars of strong classroom management. Skills related to this critical area have been shown to enhance students' motivations to learn as well as improve their academic behaviors (Emmer & Stough, 2001). In addition, Duckworth, Quinn, and Seligman (2009) found that teachers' other non-cognitive competencies such as life satisfaction and grit strongly predict teacher effectiveness in terms of students' achievement gains.

Instructors who have higher attitudes toward teaching their classes are more likely to hold students' perspectives in regard, show sensitivity to students' frustration, and create a positive climate in statistics classrooms. These instructional characteristics tend to promote more appropriate students' academic attitudes and behaviors, as evidenced in the literature of K-12 mathematics education (Pianta & Hamre, 2009). A positive classroom climate not only establishes a positive teacher-student relationship that fosters growth in self-efficacy by creating a safe learning environment free of ridicule and criticism, it also creates opportunities for academic challenge and student success in the classroom. In this environment, students perceive their teachers as caring and supportive; thus, motivating the students to exert greater effort and persistence making it more likely for them to be academically successful (Peters, 2013). For example, instructors at the high end of the teacher effectiveness distribution may be more willing to take the time and break a recalcitrant problem into smaller pieces, thereby providing more support to ensure student success in learning statistics.

Interestingly, Clunies-Ross, Little, and Kienhuis (2008) found that teachers' self-reported practices were rather consistent with their actual practices of classroom management strategies. Those management skills are closely related to teachers' attitudes toward teaching (Emmer & Stough, 2001). Together, these two findings attest to the internal validity of this study to a certain degree. More specifically, it is argued that threat to the substantive validity of the measures of instructors' attitudes toward teaching their classes may be limited.

Research Question Four

Sufficient evidence was found to substantiate a positive relationship between instructors' emphasis on data contexts and the *Value* or *Affect* components. These findings are largely in agreement with several qualitative observations (Neumann, Hood, & Neumann, 2013; Songsore & White, 2018); that is, using data-centered approaches may foster student learning of statistics via links to students' increasingly appreciating the utilitarian value of statistics in their daily lives. The instructor-associated changes in students' statistics attitudes may be viewed as an outcome of group processes directed by the instructors who frequently had students engaged in classroom activities involving data collection and analysis. Carnell (2008) suggested that instructors might need to do this if their goal was to positively impact students' statistics attitudes in that the author found limiting students to one data collection project per course was evidently insufficient to improve students' statistics attitudes.

Caution must be exercised against placing too much emphasis on the causal interpretation of the findings on instructional effects in answering this and the previous question. Aside from the issue related to nonrandom sorting of students and instructors, another important caveat here is that instructional practices are not randomly assigned to instructors. In other words, instructors chose their own instructional tools. This type of nonrandom assignment processes may lead to dependency of random effects on instructor-specific covariates. In statistics parlance, dependency of random effects on higher-level covariates could be a potential source of biases in estimation of regression coefficients at the corresponding level (Bates et al., 2015).

However, it is virtually impossible to determine to what degree the empirical evidence reflects such dependency in practice. One might seek to ameliorate this problem by incorporating as many dimensions of instructional practice into multilevel models. However, it would not help much. No matter how well observed classroom quality is being controlled for in teacher value-added models, it would still be virtually impossible to isolate one dimension of instructional practice from all others. After all, classroom instruction is an interactive system. As a result, the findings of this study provided only suggestive rather than conclusive evidence on the underlying causal relationships between instructor-associated changes in students' statistics attitudes and certain dimensions of teaching practices. In other words, it is not possible for the researcher to assert with great certainty that the frequent use of data or instructors' positive attitudes toward teaching their classes was the true cause that underlay instructor-associated changes in students' statistics attitudes. In the very same spirit, Blazar and Kraft (2017) also cautioned against the causal interpretation of their findings on teaching effects on students' academic and behaviors.

Furthermore, it is noted that the sizes of instructional effects (i.e., Cohen's f^2) are all rather small (i.e., approximately 0.02). Cohen's f^2 takes only into consideration individual residual variance at the student level so it virtually attenuates effect size of an instructor-specific variable. Additional ways to assess effect sizes of higher-level variables include estimation of the proportional change in variance (PCV) at the corresponding level upon inclusion of these variables. PCV, useful as a measure of local effect size, may allow for a closer examination of instructional effects. This approach should be used in future studies when examining instructional effects on students' statistics attitudes.

Lastly, one major factor that may spuriously influence effect size is the reliability of the instructor-specific measures. According to classical measurement theory (Raykov, Dimitrov, Marcoulides, & Harrison, 2019), any observed score is considered to consist of an underlying "true" score and a measurement error score. In this sense, instructors' item responses and therefore composite scores observed in this study are subject to their own intrinsic measurement errors. Large measurement errors bias the estimates of instructional effects downwardly; that is, the larger the error score is, the smaller the size of instructional effects becomes on students' statistics attitudes.

The reliability score was, unfortunately, not estimable for each instructor-specific construct due to non-convergent issue of confirmatory factor analysis model. Not knowing the reliability, it is virtually impossible to correct for the measurement error for estimating instructional effects through the use of advanced statistical techniques (Gillard, 2010; Lockwood & McCaffrey, 2014; Millar, White, & Romo, 2013). This non-convergent issue was likely caused by the small instructor-level sample size ($N = 23$);

that is, the ratio of factor model parameters to sample size is forbiddingly large (Myers, Ahn, & Jin, 2011). As a result, the current research design provided no means of determining how reliable the instructor-specific measures were and whether the reliability of measurements had spuriously influenced the sizes of instructional effects on students' statistics attitudes.

Research Question Five

Insufficient evidence was found to substantiate associations between instructor-associated changes in students' statistics attitudes and instructors' emphasis on conceptual understanding of statistics. This finding is not surprising. First, students' statistics attitudes are known for their recalcitrance to change over a short span of time even when instructors used moderate to intensive course interventions (Schau & Emmioğlu, 2012). Second, Tempelaar et al. (2007) have shown that students' statistics attitudes were only weakly correlated with their ability to correctly reason about statistics before students entered the courses. Their findings provided indirect evidence that instructors who improved students' statistics attitudes did not necessarily improve statistical reasoning and/or statistical thinking.

However, results from this study should by no means be interpreted as an indication that instructional approaches with emphasis on conceptual understanding cannot benefit statistics students in terms of their attitudinal development. First, there is sporadic evidence that students exposed to active learning curriculum tended to have positive statistics attitudes in comparison to those who were not (Carlson & Winquist, 2011). In a randomization study, Alan, Boneva, and Ertac (2019) provided direct evidence on the malleability of noncognitive competencies (i.e., grit) and showed that educational interventions, when timely and tailored, could and did specifically target students' non-cognitive skills for the improvement. Second, even if there is no immediate impact on students' statistics attitudes over such a short course, effects attributable to instructional approaches may be latent and later materialize as positive changes in other outcomes that are essential to students' future educational success (Ruzek et al., 2015). Lastly, the mere addition of innovative approaches to traditional lecture-based instruction may not be enough to foster a practically significant increase in statistics learning outcomes, including students' statistics attitudes. As a result, the insignificant associations between the instructor-specific *concept* domain and students' statistics attitudes resonated with the recent call for a need of new statistics curricula from the ground up (Cobb, 2015).

Research Question Six

Insufficient evidence was found to substantiate associations between instructor-associated changes in students' statistics attitudes and instructors' emphasis on assessment methods. Few studies, if any, have thus far directly examined the relationships between students' statistics attitudes and the use of assessment methods in statistics classrooms although, indirectly, Posner (2011) found that allowing students to resubmit their statistics assignments through a proficiency-based assessment system positively impacted all the six statistics attitude components. However, it is unclear as to how this finding helps inform the links between assessment of learning statistics and development of attitudes toward statistics. As with course interventions emphasizing conceptual understanding of statistics, it is reasonable to postulate that many interventions involving assessment methods may not be strong enough to induce drastic changes in students' statistics attitudes.

Assessments are frequently used to collect important information on students' learning so teachers can provide corrective instruction accordingly. In the traditional statistics classrooms, assessments have been used predominantly to give students feedback (in terms of grades) on how well they learn statistics. In the older frameworks, knowledge of statistical concepts was primarily assessed in isolation and the interconnections among those concepts were overlooked (Schau & Mattern, 1997). A new framework was recently developed toward assessing students' conceptual understanding of statistics, such as the degree to which students correctly reason about statistics (Cobb, 2015; Garfield, 2003). Major instruments were designed to assess several learning goals in statistics classrooms, including statistical reasoning, statistical thinking, and statistical literacy (delMas et al., 2007; Tobías-Lara & Gómez-Blancarte, 2019). As a result, the insignificant associations found in answering this research question are not surprising in the sense that the negative results are in line with those upon examination of the previous research question. That is, current instructional approaches and assessment methods with the focuses on conceptual understanding of statistics do not seem to improve students' attitudes toward statistics.

Moreover, it is not clear whether and to what extent the instructors selected an appropriate assessment method for a particular purpose. It should be cautioned that the majority of statistics instructors are not well-versed in education research (J. Witmer, personal communication, July 27, 2019). So were most of those who were participating in the SATS project. Therefore, there might be some misalignments between the purposes for administering a student assessment and how this assessment method was actually administered. Levels of misalignments might have spuriously affected the estimates of the associations between the *Assessment* domain and instructor-associated changes in students' statistics attitudes. In practice, the precise alignments are crucial, as an appropriate assessment method for a particular purpose helps students learn how to improve their learning outcomes as well as better informs the instructors of how student learning of statistics is scaffolded and progresses.

Research Question Seven

Students' expected course grades are a performance measure used in almost all student evaluations of teachers (SET). Average expected grades by a given course section is commonly viewed as one determinant of a SET score. For example, McPherson (2006) reported that, with an increase of 1 point in average expected course grades, a SET score increased by 0.34 point. The findings of this study suggested that instructors who were effective at improving students' statistics attitudes were equally effective at improving expected grade. Thus, higher teaching evaluation scores may be achieved when instructors perform a better job of teaching through improving various dimensions of students' academic attitudes. In addition, the instructor effectiveness measure developed in this study may be used as a way to orient instructors on development of pedagogical skills centered on students' statistics attitudes.

Although not enough is known on the exact links between instructors' contributions to students' statistics attitudes and their contributions to expected grades, the unobserved effects of instructors' ability to improve students' academic attitudes may have as strong an influence on a typical instructor's SET scores as all other effects that have been identified in the literature. However, a concern arose concurrently as to whether instructional approaches might become prescriptive to such a degree that instructors could "inadvertently buy" positive statistics attitudes or academic attitudes in general (and thus good SET scores) through a policy of easy grading or grade inflation (McPherson, 2006). Additional studies are needed to distinguish between these two possibilities – instructor-associated changes in students' statistics attitudes due to a good job of teaching and changes due to instructors' proclivity for easy grading.

It should be noted that instructors' contributions to expected course grades in relation to the *Effort* construct was not examined in this study. The *Effort* scores exhibited very little instructor-to-instructor variability, which would have certainly distorted the estimation of the correlation, were thus omitted from multilevel analysis. However, Tippin, Lafreniere, and Page (2012) have pointed out that the characteristics that underlie students' perception of grading might be better understood through looking closely at how much effort students expend to learn materials in classrooms. In addition, professors were found more likely to be show fairness in grading to students who expend greater effort and achieve better results. The relationship seems to be very intriguing between instructor-associated changes in expected grade and students' willingness to invest effort to study statistics and thus may form a basis for additional studies.

To sound a note of caution, the correlations based on BLUP estimates are imperfect measures of the true relationships between instructor effects on students' statistics attitudes and their effects on expected course grades (Guarino et al., 2015). Both sets of estimates to be correlated are measured with considerable error. That is, the precision in BLUP estimates decreases rapidly as the proportion of between-instructor variation (i.e., ICC) becomes smaller. The smaller between-instructor variation is, the more BLUP estimates put weights on zero. If this occurs, there would be very little instructor-level information to be learned about from the data. Numerous studies have highlighed this issue (e.g., Loeb & Candelaria, 2012). Desipte the awareness, equally numerous studies on teacher effectiveness still apply this approach unreservedly (e.g., Blazar & Kraft, 2017; Kraft, 2019; Loeb, Soland, & Fox, 2014; Ruzek et al., 2015), probably because correlating estimates from separate models is easier to implement.

A possible way to ameliorate this problem is through estimating the joint density of the outcome data via model-based approaches (Fraley & Raftery, 2002). For example, Leckie (2018) fit a joint value-added multilevel model simultaneously to all responses in the outcome data. This approach is preferred because it produces unbiased correlations between estimates of teacher effects (Leckie, 2018). However, it poses great methodological challenges to statistical modeling (i.e., specification, estimation, evaluation, and modification), as common methods for estimating HLM are incompatible with multivariate outcomes (Ledermann & Kenny, 2012). More flexible methods based on (Bayesian) generalized linear mixed models were proposed to tackle this methodological difficulty (Casella & George, 1992; Chib & Greenberg, 1995; Gelman et al., 2013; Gelman & Shalizi, 2013; Hadfield, 2010; Muthén & Asparouhov, 2012; Rabe-Hesketh, Skrondal, & Pickles, 2004).

Implications

This study provided the first direct evidence related to instructor and instructional effects on students' statistics attitudes. Consequently, substantive implications have emerged for instructional practice in statistics classrooms as well as for higher education policy. On one hand, content in introductory statistics courses is changing rapidly, as is the nature of statistics instruction. This study revealed the need to provide a full range of instructional practice in statistics education relevant to enhancement of students' cognitive and non-cognitive competencies. Namely, instructors should be expected not only to raise test scores but also to improve students' academic attitudes (Blazar & Kraft, 2017; Pianta & Hamre, 2009).

The differences in college instructors may not be as important to student learning as the differences in school teachers (Hoffmann & Oreopoulos, 2009). Nonetheless, the findings of this study may also help inform higher education policy in general since the influence of instructor and instructional practices on students' academic attitudes and behaviors permeates some fundamental issues concerning quality assessment in education (Amrein-Beardsley & Holloway, 2019; Blazar & Kraft, 2017). Additionally, this study was conducted primarily among statistics instructors who taught at lessselective four-year colleges (and a couple of community colleges). Therefore, it sets the stage for future research that aims to reform instructor recruitment and retention policies for this type of colleges, as they typically have relatively flexible staffing policies that allow the departments to employ a large number of part-time instructors and adjunct faculty. To sound a note of caution, while the findings of this study can certainly help inform evaluation of the effectiveness of instructional approaches and educational policies, this line of research alone is never enough to determine the most optimal approaches and policies for individual teachers. The latter is largely a matter of judgment.

Implications for Statistics Instruction

Statistics education is becoming an essential component of higher education, partly due to an increased awareness for the utilitarian value of statistics in everyday lives (Cobb, 2015). In addition, the job market now demands more perspective graduates with at least some knowledge of statistics and data science (De Veaux et al., 2017). However, most students take only one introductory statistics course when they are in college. This limited exposure to formal statistics instruction poses a great challenge to instructors if their ultimate goal is to enhance student learning of statistics, including students' statistics attitudes (Ramirez et al., 2012; Schau, 2003).

Statistics is doubtlessly gaining prominence in higher education sector. That the popularity of statistics is rapidly growing among college students with a wide range of majors prompts a rapid change in the nature of introductory statistics content. Development of new pedagogy needs to be in synergy with delivery of the emerging content. Given the findings of this study, three trends for statistics pedagogy are becoming evident: (a) active learning of modern data science concepts; (b) interactive tools for data collection and analysis; and (c) training and support for statistics instructors.

First, the findings of this study suggested the need to redesign introductory statistics courses to include data-intensive instructions, as this type of instructional approaches appeared to hold great promises for improving students' statistics attitudes. Instructional approaches involving data collection and analysis permeate deeply into every corner of the current statistics classrooms. However, many of those activities are not based explicitly on evidence and findings from educational research (Schau & Emmioğlu, 2012). Therefore, not enough is known concerning the effectiveness and efficacy of those data-facilitating activities. This study provided the first direct evidence related to the amplified role that data-intensive instruction played in improving students' statistics attitudes, and thus signified the need to continually enrich statistics curriculum materials by including modern data science concepts and other adaptive materials related to data collection and analysis (De Veaux et al., 2017).

The Change Agents for Teaching and Learning Statistics (CATALST) curriculum exemplifies such an effort in this direction (Garfield, delMas, & Zieffler, 2012). CATALST, originating at the University of Minnesota, Twin Cities, is a drastically redesigned statistics curriculum with the focus on teaching statistics inference and hypothesis testing, two of the most challenging concepts in the current statistical curricula at all levels. Students exposed to the CATALST curriculum dwell exclusively on *in situ* simulation and randomization to learn statistics inference and hypothesis testing. In-class activities were carefully designed so as to help students develop correct reasoning about statistical concepts such as sample and sampling.

Central to the idea of CATALST is that students learn efficiently when they interact with computers and are thereby exposed to mathematical/statistical models that well approximate complicated real-world problems. Once students have learned how to set up appropriate models, they use the mathematical principles underlying the models to generate data on computer (i.e., simulation). The simulated data follow certain distributions that allow students to summarize important aspects of data and make statistical inference about those aspects without using any explicit form of statistical formula. More importantly, difficulty levels of CATALST modules cater to students' learning trajectories that are used as the basis for instructional scaffolding, a teaching philosophy with noticeable connections to learning trajectory-based instruction (Sztajn, Confrey, Wilson, & Edgington, 2012).

Studies have shown that randomization-based introductory statistics curriculum helped students retain their conceptual understanding of statistical concepts (Chance et al., 2016; delMas et al., 2007; Tintle et al., 2018), although it remains unknown whether this type of curricula specifically targets weaker students with lower pre-course attitudes toward statistics. Pertinent to this study, Garfield and her colleagues showed that CATALST helped students develop great appreciation for the practicality of statistics. Not only must students still learn to acquire computational skills because much of the statistics and mathematics they encounter in their daily lives requires computational thinking (Weintrop et al., 2016), they also learn to evaluate numbers in real-world situations such as those presented in various news outlets. Thus, the drastically redesigned CATALST curriculum seems to hold great promises for improving students' statistics attitudes since it fosters ways to think statistically perhaps via links to students' reduced levels of difficulties in learning statistical inference (Case & Jacobbe, 2018).

Second, the findings of this study strongly suggested the need to devise interactive tools for students to collect and analyze data both in and outside the statistics classrooms, as new data science content reflects the learner-centered and hands-on nature of statistical practice. Interactive tools, by definition, enable students to interact with statistics problems that are genuinely inquiry-led (Biehler, Ben-Zvi, Bakker, & Makar, 2012). Teaching practices supplemented with such tools fit well in the pedagogic framework known as inquiry-informed teaching and learning, whose core tenet dwells on constructivism psychological theories stating that students must be actively engaged in the inquiry process itself so as to construct their own knowledge.

Notably, there are several initiatives in this direction that are currently helping educators rethink statistics education. These initiatives are embracing drastic changes in statistics instruction and curriculum. For example, the Mobilize project, originating at the University of California, Los Angeles, is gradually reshaping the landscape of statistics education (Gould, 2017). This project was selected as an exemplar because this model has created many opportunities for students to adapt themselves to data-centered instruction and to draw upon active learning strategies.

The Mobilize project makes use of participatory sensing, a data collection paradigm that encourages students to collect and share data on their mobile devices whenever possible. This platform is highly related to students' everyday life since the purpose of its implementation is that, by sharing data, students can address common concerns collectively through multiple cycles of investigation as articulated in Wild and Pfannkuch (1999). Participatory sensing paradigm invites discussions and communications among students and has been shown to play a lively role in statistics and data science classrooms (Gould, 2017). Statistics classes as such may well be fostering positive statistics attitudes among students. This type of teaching models requires instructors to establish a classroom environment where learning activities are scaffolded and well-organized. Yet, many statistics instructors lack classroom management and organization skills in this critical area.

Third, the findings of this study suggested the need to provide professional training and support for statistics instructors. In reality, the hiring of new university faculty members by the Mathematics and Statistics Department are typically research-oriented. As a result, it is not unusual that new faculty are hired without formal knowledge of inquiry pedagogy. Lacking this set of skills may negatively impact student learning of modern statistics concepts since inquiry process lies at the heart of statistical practice of professions and therefore should be effectively encultured into the ways in which students think and reason about statistics (Cobb, 2015). In reality, opportunities to procure knowledge and skills of inquiry pedagogy are hardly provided to statistics instructors and professors as a part of professional development.

This is very problematic in the sense that a drastic shift in pedagogy to learner-centered focus appears to be inevitable in college education. Yet, most, if not all, statistics professors still have strong proclivity for lecture-centered approaches since this type of approaches worked very well for them when they were students. However, what works for statistics instructors does not necessarily work for students (Moore, 1997). This conflict is exacerbated by the fact that colleges are almost always lagging behind schools in terms of educational reform. Students may have already become accustomed to learner-centered environment before they take college introductory statistics courses where instructors may not be as sensitive to students' frustration and problems as their school teachers used to be.

Moreover, college instructors may be concerned less with creating positive classroom climate than with directly transferring knowledge. A relapse into lecture-centered classrooms in college may have a negative impact on students' learning outcomes, perhaps particularly on those perceived to be relevant for the *Affect* and *Interest* components of students' statistics attitudes. One way to ameliorate this problem is to motivate instructors to develop formal knowledge of pedagogic theory and skills in ways that enhance student learning, including attitudes toward statistics. Thus, the findings of this study could serve as a strong signal that professional development should be valued in both school and college settings, as emphasized in Goldhaber (2019). Specifically, college instructors and professors may receive training with focus on development of classroom management skills whereas school mathematics teachers may take classes with the goal to procure statistics content-related skills.

Implications for Educational Policy

This study utilized the core idea stemming from teacher value-added methodologies, the application of which has gained much momentum since around 1970s and is now permeating many current policy ideas (Amrein-Beardsley & Holloway, 2019; Hanushek & Rivkin, 2010). The original concept that underlay the term "teacher value-added" is nevertheless a century-old. This term was first proposed to measure the impact of teaching performances on students' test scores. Its usage had been primarily confined to academic use until later it gained prominence by finding its way into education policy-making.

The incorporation of value-added methodology into teacher accountability policy culminated in President Barack Obama's administration with it being viewed as an "objective" alternative to other subjective measures of teacher effectiveness such as principal assessments through classroom observations. Since the inception of the Race to the Top initiatives in 2009, there is a rapidly growing interest in incorporating measures of student outcomes in areas beyond core academic skills into teacher and school accountability systems (Amrein-Beardsley & Holloway, 2019). In K-12 education, the Every Student Succeeds Act (ESSA) was signed into law in 2015. Many viewed this enactment as a revival of American long and deep tradition that local government should have a tenacious grip on education policy-making (Heise, 2017). This federal law offers states with great latitude to design and implement student assessment systems with a few stipulations. For example, ESSA requires states to assess student success in at least one area beyond their academic performance. This nonacademic indicator could be students' non-cognitive competencies such as academic attitudes, behaviors, or emotional skills.

While the standards college students are held to are not enacted by the legislature, efforts directed at improving students' academic attitudes could serve as a strong signal that colleges and policymakers should value both types of academic skills – cognitive and non-cognitive (Rogaten et al., 2019). Findings from this study insinuated the possibility of incorporating instructor and instructional effects on students' statistics attitudes into the evaluation of statistics instructors. The measures of instructor effectiveness developed in this study can be used to orient the statistics and mathematics departments on the enhancement of the motivation to devise a systematized process by which to evaluate statistics instructors.

Yet more broadly, evaluation of teacher quality that improves students' multifaceted competencies has been of great significance to institutional research in higher education. The instructor effectiveness measures can also be applied to evaluating other college instructors than those teaching statistics. The instructor-specific measures used in this study captured various instructional domains that are not commonly collected as part of the instructor evaluation systems in colleges. If measures of the instructor effectiveness at improving students' academic attitudes coincide with measures of the instructor effectiveness at improving expected grade, as has been demonstrated in this study, it is then possible to evaluate instructors based on the extent to which they improve students' academic attitudes in higher education settings.

However, caution must be exercised against using either a single weighted composite of different measures of instructor effectiveness or a single performance category such as the recognition of instructor' contributions to students' academic attitudes and behaviors (Mihaly, McCaffrey, Staiger, & Lockwood, 2013). It is necessary to leverage several measures for evaluation of teaching quality. Simple composite scores would oversimplify the full complexity of the actual teaching performances. For instance, an instructor who is skilled at improving students' academic attitudes may not necessarily excel at raising final examination scores. On the other hand, a different instructor may be mediocre and thus middling across these two measures of teaching effectiveness. A single overall evaluation score only looking at composite scores would suggest these two instructors are situated at the similar places on the distribution of teaching effectiveness. In this sense, a more systematized process that recognizes the full complexity of instructional practice is needed, as it not only supports ongoing teacher development but also better informs important decisions involved in instructor evaluation such as whether or not to grant an instructor higher pay, bonus, compensation, retention, and/or tenure (Blazar & Kraft, 2017; Rothstein, 2015).

This study also helps inform K-12 education policy. Statistics content is rapidly filtering down to school levels and has been utilized across multiple school subjects, as evidenced in the Common Core State Standards Initiative (Franklin, 2013). Many K-12 educators have situated their works in redesign of mathematics curriculum materials to provide the best opportunities for students in K-12 to learn statistics (Usiskin & Hall, 2015). In this spirit, Hommik and Luik (2017) revised and adapted the SATS-36 for measuring attitudes toward statistics among secondary school students. If this trend continues (and it most likely will), the instructor effectiveness measure developed in this study may be carried over to evaluating teacher effects on students' statistics attitudes on much larger scale at all school levels, supplementing other measures of teacher effectiveness such as student survey and classroom observations (Steinberg & Donaldson, 2016).

Despite the nationwide push to incorporate value-added measures of teachers' ability to improve students' academic attitudes and behaviors into accountability policy, many researchers have voiced their strong concerns over a rush to apply these measures to high-stakes decisions (Amrein-Beardsley & Holloway, 2019). The basis for this admonition is understandable; value-added models fall short of a panacea in evaluating teaching performances. The prediction quality of value-added models is susceptible to numerous factors. Technically, the quality of the predictions is determined by how well covariates are being controlled for in value-added models.

In practice, the list of student-specific variables is always short. Given the limited number of predictors, a large portion of variation in student preparedness might not be captured by teacher value-added models. Prediction models as such run a great many risks of misidentifying effective teachers and thereby diluting the usefulness of teacher effectiveness measures. Matters are even worse when biases in value-added models could misguide teachers to devote much of their efforts toward gaming the accountability system rather than actually doing a good job on teaching. The worst-case scenario is that teachers might navigate the performance measurement process to their advantage, possibly through deliberately manipulating educational inputs at their disposal.

Besides potential biases associated with the prediction quality of value-added estimates arising from multilevel covariate adjustment models, an additional concern is that the science of measuring attitudes, behaviors, and emotional skills is still in its infancy (Duckworth & Yeager, 2015; West et al., 2016). So is the science of measuring students' statistics attitudes (Xu & Schau, 2019). Although the SATS-36 is a widely used and well-studied measure of students' statistics attitudes (Nolan et al., 2012), it is only valid and reliable when it is used in rather small-scale research settings rather than in large-scale cross-institutional assessments with repeated administration.

In this regard, substantive research in educational psychology is getting way ahead of the curve in developing tools that both methodologists and statisticians are able to provide. Nonetheless, central to the issue related to large-scale repeated administration of the SATS-36 is the property of measurement invariance grounded in item response theory, particularly at the ordinal level (Jak & Jorgensen, 2017; Muthén, 1984; Millsap, 2011; Millsap & Yun-Tein, 2004; Yang-Wallentin, Jöreskog, & Luo, 2010). Problematic items on the SATS-36 that cause violations of measurement invariance need to be removed and/or revised, before this instrument can be used to better inform the departments about important decisions based on measures of the teacher effectiveness at improving students' attitudes toward statistics. As Xu and Schau (2019) have shown, the violation of measurement invariance may be caused by a considerable amount of method variance (Spector, 2006), which is temporally differential in several statistics attitude constructs.

Recommendations for Future Research

Promising as the findings of this study are to instructional design and innovation in statistics education, it is important to note that the use of the instructor-specific measures may be still too exploratory to accurately capture certain instructional dimensions in statistics classrooms. As a result, the findings are necessarily tentative regarding instructional effects on students' statistics attitudes. Recommendations for future research will help further arguments for making instructor and instructional quality a central theme of instructional design and curriculum development in statistics education, as advocated in the GAISE report. The following recommendations are provided consistent with many of the current advances in research on K-12 mathematics education (e.g., Alan et al., 2019; Blazar & Kraft, 2017; Korpershoek et al., 2016; Pianta & Hamre, 2009) and include, but are not limited to, (a) evaluation of the interactions between instructors' teaching experiences and students' pre-course statistics attitudes; (b) examination of the consistency of instructor effects on multiple outcomes of learning statistics as well as the stability of instructor effects across time; (c) validation of the measures of the three instructional domains – *Data*, *Concept*, and *Assessment*; (d) development and validation of measures of instructors' attitudes toward teaching statistics; and (e) observations of statistics classrooms.

First, the effect of one independent variable often changes when its interactions with other variables are considered. Such conditional relations can be either modelled as multiplicative interactions in simple regression settings or as moderations in more sophisticated settings such as structural equation modeling. Pertinent to this study, the interactions between students' pre-course statistics attitudes and instructors' experiences in teaching (i.e., high, medium, low) are of great interest as the former can be viewed as a function of the latter. In practice, the pre-version of the SATS-36 is typically administered two weeks into the semester. Thus, instructors are expected to have already had a potentially non-negligible impact on students' statistics attitudes measured for the first time. Hypotheses concerning such conditional relations need to be explicitly tested in future studies.

Second, two fundamental issues related to value-added measures are the consistency and stability of teacher-value added measures (Leckie, 2018; Loeb & Candelaria, 2012). In other words, the instructor-level random effects for different learning outcomes or for different years need to be correlated. Pertinent to the present study, the consistency of value-added measures may be used to describe whether an instructor who improves students' statistics attitudes is likely to also improve other important learning outcomes such as statistics achievement and complex cognitive skills (e.g., conceptual understanding of statistics). Furthermore, knowing the stability of value-added measures helps inform the department of whether an instructor with a high value-added estimate in one semester is likely to be equally effective the next. Unfortunately, the SATS project data do not permit this line of research since it contains only the outcome data on students' statistics attitudes. Future studies should engender abundant opportunities for reconsideration of the status quo in statistics education research by collecting large-scale data on multiple outcomes of learning statistics.

Third, the participating instructors self-reported their teaching practices by answering an instructor survey. It remains unclear to what degree the instructor-specific instrument is reliable and/or valid. Future studies should address these qualities, as validity and reliability always remain the central pieces of psychological measurement and instrument development (Clark & Watson, 1995). The instructor-specific measure was not validated by confirmatory factor analysis in the present study although it appeared to have strong substantive validity and theoretical underpinnings, as eight of ten questions on the instructor questionnaire were developed in agreement with the recommendations made by the GAISE report. To fully appreciate the complexity of construct validity, the first step may thus be to test the structural and factorial validity of the instructor-specific measures developed in this study (i.e., the measures of the *Data*, *Concept*, and *Assessment* constructs). An instrument that is both reliable and valid will contribute substantially to the precision with which instructional effects on students' statistics attitudes are measured.

Fourth, it is of both theoretical and practical interest to explore the ways in which instructors' attitudes affect students' interest in statistics and feelings about this discipline. Research on non-cognitive dimensions of teaching competencies is noticeably lacking in the literature of statistics education, as evidenced by the fact that these dimensions are not being recognized in the recommendations made in the GAISE report.

This line of research will have significant implications for statistics instruction but also will pose an even greater challenge to statistics educators and researchers. Namely, it demands practical and reliable measures of instructors' attitudes. The use of individual items in this study should be viewed as a tentative approach to capturing either instructors' attitudes toward teaching their course sections or their attitudes toward introductory statistics courses.

Future studies should utilize mixed methods approaches to capture various dimensions of instructors' attitudes in classrooms that hold promises for improving students' statistics attitudes. For example, students may be interviewed to acquire useful information on what dimensions of instructors' attitudes they believe affect their attitudes toward statistics. The qualitative data can then be managed to build a set of refined items that constitute instructor-specific questionnaires to be used for measuring multidimensional constructs of instructors' attitudes in statistics classrooms. Ideally, such statistics-specific observation instruments should capture a multitude of instructional dimensions that encompass instructors' attitudes toward teaching such as their willingness to (a) incorporate research-informed and evidence-based approaches into classroom activities, (b) provide emotional support to those students whom have developed frustration over learning statistics, (c) manage classroom behaviors, (d) organize classroom environment, (e) improve the precision with which instructors deliver statistical content, and (f) promote student critical thinking. Such measures will enable the researchers to estimate instructors' non-cognitive contributions to student outcomes in a more precise manner.

Lastly, future research will also need to focus on audio- and/or video-recordings of classroom activities as a means of identifying and validating various dimensions of instructional practice in statistics classrooms. This line of research utilizing classroom observations is noticeably lacking in the field of statistics education research. Yet, it is worthy and necessary since audio-and video-recordings of classroom activities remain the centerpiece of many studies in research on mathematics education when the purpose is to create measures of instructor-specific constructs. For instance, Bacher-Hicks, Chin, Kane, and Staiger (2019) found that classroom observation scores derived from video- and audio-recordings of mathematics lessons, though less precise as a measure of teacher effectiveness than value-added scores, strongly predict teachers' performances. This is not surprising. Classroom observations, when done properly, provide a panoramic picture of the classroom environment by focusing on both individual students and teachers, and thereby capture the dynamics of what takes place between students and teachers in the classroom. In addition, observations of classroom activities help the researchers better interpret their findings by taking into consideration the actual size of the consistencies between teachers' self-reported teaching practices and their actual teaching practice.

The other major benefit of reconciling teacher self-reported measures with videotaped classroom teaching is that this approach minimizes the threat of reference bias (Duckworth & Yeager, 2015; West et al., 2016). Differences in school-wide norms by which students interact with teachers may change the implicit standard of comparison that teachers use to judge their own attitudes toward teaching practices. This phenomenon, often referred to as attitude-achievement anomaly in the literature of educational psychology (West et al., 2016), could cast a shadow over the general validity of the studies where comparisons of students' attitude outcomes are of primary interest across different educational settings. For example, comparing academic attitudes among students can be particularly nettlesome when they are from schools in which culture and climate are vastly different from each other. As a result, reference bias would also cause biases in outcome measurements related to teacher effectiveness at improving students' attitudes.

Conclusion

A substantial body of studies have examined the effects of individual teachers and various teaching practices on student learning outcomes. Findings from multiple studies suggested that teacher quality served as a more important factor contributing to student outcomes than any other educational inputs (Nye et al., 2004; Odden et al., 2004). Additional evidence indicated that teacher effects were inconsistent on student achievement outcomes across school subject areas (Leckie, 2018; Loeb & Candelaria, 2012). However, teacher and teaching effects on students' academic attitudes and behaviors have been much less researched than their effects on achievement outcomes such as standardized test scores. Several recent studies in K-12 education have shown that teacher-associated changes in students' academic attitudes and behaviors were substantially larger than teacher-associated changes in their scores on both low-stakes classroom assessments and high-stakes state tests (Blazar, 2018; Blazar & Kraft, 2017; Chetty et al., 2011; Kraft, 2019).

A mountain of empirical evidence shows that students' statistics attitudes are determined by a slew of factors that are beyond teachers' control, such as age, and prior experiences with mathematics and other quantitative disciplines (Schau, 2003; Ramirez et al., 2012). Yet, still very little evidence existed to support the presence of instructor and instructional effects on students' statistics attitudes although it was implied yet unsubstantiated in at least two earlier studies (Schau, 2003; Waters et al. 1988). If these effects were genuine, the question then became "To what extent instructor and instructional practices affect students' statistics attitudes?" This study has closed this gap in the literature and therefore represents a unique contribution to the field of statistics education research.

The present study utilized multilevel models with covariate adjustment to analyze the rich set of SATS project data. Results showed that instructors varied considerably in their abilities to improve students' statistics attitudes. This varied capability could partly be accounted for by how much instructors liked teaching their classes as well as by the extent to which instructors used data collection and analysis to engage students in learning statistics. Together, these two findings supported a data-intensive introduction to statistics for non-majors and also shed light on the importance of instructors' non-cognitive competencies in improving learning outcomes. Moreover, instructors who were effective at improving students' statistics attitudes were found to be equally effective at improving expected course grades. This finding indicated that instructors may achieve higher scores on teaching evaluation through a better job of teaching that improves students' statistics attitudes. Due to the connections that value-added models have to policy-making, substantive implications have also emerged for higher education policy.

In the U.S. and many other countries around the world, the number of college students enrolled in introductory statistics courses is growing rapidly. This trend will certainly continue in the age of information. The growth in enrollment is largely driven by the undercurrent that statistics has become indispensable for many academic disciplines. Some of these disciplines, such as biology, medicine, and psychology, were traditionally considered to be relatively statistics-free. Today, they are becoming increasingly statistics- and data-oriented. Despite the limited exposure to formal statistics instruction, students see clearly the need to become statistically literate in order to succeed in their fields of study as well as in their future endeavors. This trend demands high-quality statistics instruction in the teaching of Statistics 101 due to the vast differences in students' academic backgrounds and levels of mathematics preparations, as introductory statistics courses are now becoming required for many degree programs. In reality, many introductory statistics students are still not good at generalizing correctly what they have learned because these students do not have a basic accurate cognitive scaffold in place that allows them to do so.

For decades in statistics education, research efforts have focused predominantly on enhancement of instructors' abilities to raise test scores on statistics and to develop statistical reasoning. This work has expanded this focus to instructors' contributions as well as instructional characteristics related to desired statistics attitude outcomes. The findings of the present study, in conjunction with those from many other studies with respect to a variety of statistics learning outcomes, herald an exciting opportunity to expose students to a full range of instructional skills in the modern statistics classrooms.

REFERENCES

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics, 134*(3), 1121-1162.

Amrein-Beardsley, A., & Holloway, J. (2019). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy, 33*(3), 516-542.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40-64.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association, 91*(434), 444-455.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review, 73*, 1-15.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1).

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*(2), 135-167.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2012). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643-689). New York, NY: Springer.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, 13(3), 281-309.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146-170.

Brophy, S., Klein, S., Portsmore, M., & Rogers, C. (2008). Advancing engineering education in P-12 classrooms. *Journal of Engineering Education, 97*(3), 369-387.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. C. Berliner (Ed.), *Review of research in education* (pp. 158-223). Washington, DC: American Educational Research Association.

Carlson, K. A., & Winquist, J. R. (2011). Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach. *Journal of Statistics Education, 19*(1).

Carnell, L. J. (2008). The effect of a student-designed data collection project on attitudes toward statistics. *Journal of Statistics Education, 16*(1).

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409-432.

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal, 17*(2), 9-29.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174.

Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics scale: A construct validity study. *Educational and Psychological Measurement, 65*(3), 509-524.

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education, 24*(3), 114-126.

Cheng, A., & Zamarro, G. (2018). Measuring teacher non-cognitive skills and its impact on students: Insight from the Measures of Effective Teaching Longitudinal Database. *Economics of Education Review, 64*, 251-260.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*(4), 1593-1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The Aamerican Statistician, 49*(4), 327-335.

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal, 9*(1), 6-26.

Chingos, M. M. (2016). Instructional quality and student learning in higher education: Evidence from developmental algebra courses. *The Journal of Higher Education, 87*(1), 84-114.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational Psychology, 28*(6), 693-710.

Cobb, G. W. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician, 69*(4), 266-282.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801-823.

Coetzee, S., & van der Merwe, P. (2010). Industrial psychology students' attitudes towards statistics. *SA Journal of Industrial Psychology, 36*(1), 1-8.

Cohen, G. L., Garcia, J., & Goyer, J. P. (2017). Turning point: Targeted, tailored, andtimely psychological intervention. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (pp. 657-686). NewYork, NY: Guilford Press.

Curran, P. J. (2003). Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research, 38*(4), 529-569.

Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of Attitudes Toward Statistics: Factor structure and factorial invariance for women and men. *Structural Equation Modeling, 4*(2), 129-141.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Kim, A. Y. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and its Application, 4*, 15-30.

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis, 33*(1), 23-46.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Dempster, M., & McCorry, N. K. (2009). The role of previous experience and attitudes toward statistics in statistics assessment outcomes among undergraduate psychology students. *Journal of Statistics Education, 17*(2).

Deng, X., Liu, H., Shao, Y., Rayner, S., & Yang, R. (2008). The epidemic origin and molecular properties of B′: A founder strain of the HIV-1 transmission in Asia. *AIDS, 22*(14), 1851-1858.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440-464.

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education, 77*(4), 267-282.

Duckworth, A. L., Quinn, P. D., & Seligman, M. E. (2009). Positive predictors of teacher effectiveness. *The Journal of Positive Psychology, 4*(6), 540-547.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237-251.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109-132.

Egalite, A. J., & Kisida, B. (2018). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis, 40*(1), 59-81.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, 36(2), 103-112.

Emmioğlu, E., & Capa-Aydin, Y. (2012). Attitudes and achievement in statistics: A meta-analysis study. Statistics Education Research Journal, 11(2), 95-103.

Evans, B. (2007). Student attitudes, conceptions, and achievement in introductory undergraduate college statistics. *The Mathematics Educator, 17*(2), 24-30.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association, 97*(458), 611-631.

Franklin, C. (2013). Common core state standards and the future of teacher preparation in statistics. *The Mathematics Educator, 22*(2), 3-10.

GAISE College Report ASA Revision Committee (2016). *The GAISE (Guidelines for Assessment and Instruction in Statistical Education) College Report*. Alexandria, VA: The American Statistical Association.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education, 53*(3), 353-374.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22-38.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education, 44*(7), 883-898.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education, 10*(2).

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72-91.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd Ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8-38.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy, 11*(2), 125-149.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor?. *Psychology Science Quarterly, 50*(1), 21-43.

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *REVSTAT–Statistical Journal, 8*(1), 57-80.

Goldhaber, D. (2019). Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education, 70*(2), 90-101.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22-25.

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics, 10*(4), 2405-2426.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics, 40*(2), 190-222.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted?. *Education Finance and Policy, 10*(1), 117-156.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*(2).

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal, 113*(485), F64-F98.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-71.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350.

Heise, M., (2017). From No Child Left Behind to Every Student Succeeds: Back to a future for education federalism. *Columbia Law Review, 117*, 1859–1896.

Hilton, S. C., Schau, C., & Olsen, J. A. (2004). Survey of Attitudes Toward Statistics: Factor structure invariance by gender and by administration time. *Structural Equation Modeling, 11*(1), 92-109.

Hoffmann, F., & Oreopoulos, P. (2009). Professor qualities and student achievement. The *Review of Economics and Statistics, 91*(1), 83-92.

Hommik, C., & Luik, P. (2017). Adapting the Survey of Attitudes towards Statistics (SATS-36) for Estonian secondary school students. *Statistics Education Research Journal, 16*(1), 228-239.

Hood, M., Creed, P. A., & Neumann, D. L. (2012). Using the expectancy value model of motivation to understand the relationship between student attitudes and achievement in statistics. *Statistics Education Research Journal, 11*(2), 72-85.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Imbens, G., & Rubin, D. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge, MA: Cambridge University Press.

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology, 8*, 1640.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling, 21*(1), 31-39.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education, 83*(2), 135-159.

Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 40*(2), 267-284.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kerby, A. T., & Wroughton, J. R. (2017). When do students' attitudes change? Investigating student attitudes at midterm. *Statistics Education Research Journal, 16*(2), 476-486.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195.

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6).

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research, 86*(3), 643-680.

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software, 74*(7).

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources, 54*(1), 1-36.

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy, 12*(2), 241-279.

Leckie, G. (2018). Avoiding bias when estimating the consistency and stability of value-added school effects. *Journal of Educational and Behavioral Statistics, 43*(4), 440-468.

Ledermann, T., & Kenny, D. A. (2012). The common fate model for dyadic data: Variations of a theoretically important but underutilized model. *Journal of Family Psychology, 26*(1), 140-148.

Lesser, L. M., Pearl, D. K., & Weber, J. J. (2016). Assessing fun items' effectiveness in increasing learning of college introductory statistics students: Results of a randomized experiment. *Journal of Statistics Education, 24*(2), 54-62.

Liu, O. L. (2011). Value-added assessment in higher education: A comparison of two methods. *Higher Education, 61*(4), 445-461.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*(9), 352-362.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22-52.

Loeb, S., & Candelaria, C. A. (2012). *How stable are value-added estimates across years, subjects and student groups? What we know series: Valueadded methods and applications* (Knowledge Brief 3). Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis, 36*(4), 457-475.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*, 169.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(1), 3-20.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259-284.

Mihaly, K., McCaffrey, D.F., Staiger, D.O., & Lockwood, J.R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Millar, A. M., & Schau, C. (2010). *Assessing students' attitudes: The good, the bad, and the ugly.* Paper presented at the the Joint Statistical Meetings, Vancouver, British Columbia.

Millar, M., White, B., & Romo, R. (2013). *Comparing apples with apples: Assessing student attitudes in the presence of regression to the mean*. Paper presented at the the Joint Statistical Meetings, Montréal, Québec.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor & Francis Group.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., ... & Sears, M. R. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, 108*(7), 2693-2698.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. International *Statistical Review, 65*(2), 123-137.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113-127.

Muenks, K., Wigfield, A., & Eccles, J. S. (2018). I can do this! The development and calibration of children's expectations for success and competence beliefs. *Developmental Review, 48*, 24-39.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313-335.

Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport, 82*(3), 412-423.

Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal, 12*(2), 59-70.

Nolan, M. M., Beran, T., & Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal, 11*(2), 103-123.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

Ograjenšek, I., & Gal, I. (2016). Enhancing statistics education by including qualitative research. *International Statistical Review, 84*(2), 165-178.

Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assessment & Evaluation in Higher Education, 29*(1), 3-19.

Paul, W., & Cunnington, R. C. (2017). An exploration of student attitudes and satisfaction in a GAISE-influenced introductory statistics course. *Statistics Education Research Journal, 16*(2), 487-510.

Persson, I., Kraus, K., Hansson, L., & Wallentin, F. (2019). Confirming the structure ofthe Survey of Attitudes Toward Statistics (SATS-36) by Swedish students. *Statistics Education Research Journal, 18*(1), 83-93.

Peters, M. L. (2013). Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: A multi-level analysis. *International Journal of Science and Mathematics Education, 11*(2), 459-480.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Posner, M. A. (2011). The impact of a proficiency-based assessment and reassessment of learning outcomes system on student achievement and attitudes. *Statistics Education Research Journal, 10*(1), 3-15.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190.

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal, 11*(2), 57-71.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & Harrison, M. (2019). On the connections between item response theory and classical test theory: A note on true score evaluation for polytomous items via item response modeling. *Educational and Psychological Measurement, 79*(6), 1198–1209.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150.

Rogaten, J., Rienties, B., Sharpe, R., Cross, S., Whitelock, D., Lygo-Baker, S., & Littlejohn, A. (2019). Reviewing affective, behavioural and cognitive learning gains in higher education. *Assessment & Evaluation in Higher Education, 44*(3), 321-337.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2).

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, 105(1), 100-130.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence, 35*(5-6), 852-882.

Schau, C. (2003). *Students' attitudes: The "other" important outcome in statistics education.* Paper presented at the the Joint Statistical Meetings, San Francisco, CA.

Schau, C., & Emmioğlu, E. (2012). Do introductory statistics courses in the United States improve students' attitudes? *Statistics Education Research Journal, 11*(2), 86-95.

Schau, C., & Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *The American Statistician, 51*(2), 171-175.

Schau, C., Stevens, J., Dauphinee, T. L., & Vecchio, A. D. (1995). The development and validation of the survey of attitudes toward statistics. *Educational and Psychological Measurement, 55*(5), 868-875.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*(2), 122-140.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology, 3*, 111.

Shavelson, R. J., Domingue, B. W., Mariño, J. P., Molina Mantilla, A., Morales Forero, A., & Wiley, E. E. (2016). On the practices and challenges of measuring higher education value added: The case of Colombia. *Assessment & Evaluation in Higher Education, 41*(5), 695-720.

Songsore, E., & White, B. J. (2018). Students' perceptions of the future relevance of statistics afer completing an online introductory statistics course. *Statistics Education Research Journal, 17*(2), 120-140.

Spector, P. E. (2006). Method variance in organizational research: Truth or urban legend?. *Organizational Research Methods, 9*(2), 221-232.

Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2019). A new perspective on method variance: A measure-centric approach. *Journal of Management, 45*(3), 855-880.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481-520.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359.

Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637-652.

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction: Toward a theory of teaching. *Educational Researcher, 41*(5), 147-156.

Tempelaar, D. T., Schim van der Loeff, S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal,, 6*(2), 78-102.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., ... & VanderStoep, J. (2018). Assessing the association between pre-course metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation based inference. *Journal of Statistics Education, 26*(2), 103-109.

Tippin, G. K., Lafreniere, K. D., & Page, S. (2012). Student perception of academic grading: Personality, academic orientation, and effort. *Active Learning in Higher Education, 13*(1), 51-61.

Tobías-Lara, M. G., & Gómez-Blancarte, A. L. (2019). Assessment of informal and formal inferential reasoning: A critical research review. *Statistics Education Research Journal,, 18*(1), 8-25

Umbach, P. D., & Wawrzynski, M. R. (2005). Faculty do matter: The role of college faculty in student learning and engagement. *Research in Higher Education, 46*(2), 153-184.

Usiskin, Z., & Hall, K. (2015). The relationships between statistics and other subjects in the K-12 curriculum. *Chance, 28*(3), 4-18.

van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research, 48*(6), 895-922.

van Es, C., & Weaver, M. M. (2018). Race, sex, and their influences on introductory statistics education. *Journal of Statistics Education, 26*(1), 48-54.

Vanhoof, S., Castro Sotos, A. E., Onghena, P., Verschaffel, L., Van Dooren, W., & Van den Noortgate, W. (2006). Attitudes toward statistics and their relationship with short-and long-term exam results. *Journal of Statistics Education, 14*(3).

Vanhoof, S., Kuppens, S., Sotos, A. E. C., Verschaffel, L., & Onghena, P. (2011). Measuring statistics attitudes: Structure of the Survey of Attitudes toward Statistics (SATS-36). *Statistics Education Research Journal, 10*(1), 35-51.

Wang, P., Palocsay, S. W., Shi, J., & White, M. M. (2018). Examining undergraduate students' attitudes toward business statistics in the United States and China. *Decision Sciences Journal of Innovative Education, 16*(3), 197-216.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133.

Waters, L., Martelli, T. A., Zakrajsek, T., & Popovich, P. M. (1988). Attitudes toward statistics: An evaluation of multiple measures. *Educational and Psychological Measurement, 48*(2), 513-516.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127-147.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38*(1), 148-170.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68-81.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Xu, C., & Schau, C. (2019). Exploring method effects in the six-factor structure of the Survey of Attitudes Toward Statistics (SATS-36). *Statistics Education Research Journal, 18*(2), 39-53.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*(3), 392-423.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education, 16*(2).

APPENDIX A:

DATA PERMISSION

Dear Chao Xu,

I am happy to grant you permission to use the SATS project data in your dissertation. Please let me know if you have any questions.

My best with your work.

Candace

Candace Schau, PhD

CS Consultants, LLC

[redacted]

www.evaluationandstatistics.com

From: Chao Xu <[redacted]>
Sent: Friday, August 31, 2018 5:19 PM
To: Candace <[redacted]>
Subject: Request for using the SATS project data in my dissertation

Good evening Dr. Schau. I am a doctoral candidate at the University of Houston - Clear Lake. I found the structure of the SATS project data you kindly shared with me fits perfectly well with the purpose of my research. That is, I proposed to examine instructor and instructional effects on students' attitudes toward statistics. May I have your permission to use the SATS project data in my research for the dissertation and possible publication? Your contributions will certainly be acknowledged therein. Thanks in advance.

Best regards,
Chao Xu

APPENDIX B:

SAMPLE R CODE

```
> ##Only the Value construct is illustrated
> #Set working directory, read in data and load robustlmm package
> setwd("C:/Users/Chaohzxuchao/Desktop/Dissertation Research/Data")
> data <- read.table("JSE.dat", header = TRUE, sep = "\t", fill = TRUE)
> library(robustlmm)

> #Fit the model presented in Equation 1 to data
> fit_value <- rlmer(PostValue.z ~ PreValue.z + gender + Age + premach + InstType +
> +                  Age.class + premach.class + gender.class + (1 | Instructor),
> +                  REML = TRUE, data)

> #Extract instructor effects on Value
> attr(VarCorr(fit_value)$Instructor, "stddev")

> #Fit the model presented in Equation 2 to data
> fit_Value <- rlmer(PostValue ~ PreValue + gender + Age + premach + InstType +
> +                  Age.class + premach.class + gender.class +
> +                  ThisCourse.avg + GenThisCourse.avg + data.comp + lit.comp +
> +                  assess.comp + (1 | Instructor),
> +                  REML = TRUE, data)

> #Extract instructional effects on Value
> summary(fit_Value)$coefficients[11:14, ]

> #Extract BLUP estimates of instructor effects on Value and expected grade
> BLUP_value <- (unlist(ranef(fit_value)$Instructor))
> BLUP_exgrade <- (unlist(ranef(fit_exgrade)$Instructor))

> #The correlation between instructors' contributions to Value and expected grade
> cor.test(BLUP_value, BLUP_exgrade, method = "pearson")
```

APPENDIX C:

SURVEY OF ATTITUDES TOWARD STATISTICS (SATS-36)

DIRECTIONS: Only items from the pre-version of the SATS-36 are listed herein for the purpose of illustration. Candace Schau has a proprietary right to the SATS-36. Interested readers who are to use this instrument shall contact Candace Schau for further information at https://www.evaluationandstatistics.com.

I plan to complete all of my statistics assignments.

I plan to work hard in my statistics course.

I will like statistics.

I will feel insecure when I have to do statistics problems.

I will have trouble understanding statistics because of how I think.

Statistics formulas are easy to understand.

Statistics is worthless.

Statistics is a complicated subject.

Statistics should be a required part of my professional training.

Statistical skills will make me more employable.

I will have no idea of what's going on in this statistics course.

I am interested in being able to communicate statistics information to others.

Statistics is not useful to the typical professional.

I plan to study hard for every statistics test.

I will get frustrated going over statistics tests in class.

Statistical thinking is not applicable in my life outside my job.

I use statistics in my everyday life.

I will be under stress during statistics class.

I will enjoy taking statistics courses.

I am interested in using statistics.

Statistics conclusions are rarely presented in everyday life.

Statistics is a subject quickly learned by most people.

I am interested in understanding statistics information.

Learning statistics requires a great deal of discipline.

I will have no application for statistics in my profession.

I will make a lot of math errors in statistics.

I plan to attend every statistics class session.

I am scared by statistics.

I am interested in learning statistics.

Statistics involves massive computations.

I can learn statistics.

I will understand statistics equations.

Statistics is irrelevant in my life.

Statistics is highly technical.

I will find it difficult to understand statistical concepts.

Most people have to learn a new way of thinking to do statistics.