THE IMPACT OF STUDENT-DIRECTED PROJECTS IN INTRODUCTORY STATISTICS

DIANNA J. SPENCE
University of North Georgia
djspence@ung.edu

BRAD BAILEY
University of North Georgia
bbailey@ung.edu

JULIA L. SHARP
Clemson University
jsharp@clemson.edu

ABSTRACT

A multi-year study investigated the impact of incorporating student-directed discovery projects into introductory statistics courses. Pilot instructors at institutions across the United States taught statistics implementing student-directed projects with the help of a common set of instructional materials designed to facilitate such projects. Researchers measured the impact of these projects on student learning and on students' attitudes and beliefs about statistics. Results of the quantitative analyses are shared, with subsequent discussion of their implications. Findings suggest that inclusion of student-directed research projects in introductory statistics can lead to greater statistics self-efficacy and improved statistical knowledge in specific domains. Additional analyses suggest that these student benefits may improve as their instructors gain more experience facilitating such projects.

Keywords: *Statistics education research; discovery projects; student-directed projects; statistics self-efficacy; linear regression; hypothesis testing*

1. INTRODUCTION

1.1. BACKGROUND AND RATIONALE

The use of authentic student-directed research projects in elementary statistics courses addresses some of the recommendations that have been made in the preceding two decades for improving statistics education. For instance, the Guidelines for Assessment and Instruction in Statistics Education [GAISE] College Report (Aliaga et al., 2005) promotes the use of real data and the fostering of active learning. In particular, the GAISE report suggests that statistics instructors leverage "alternatives such as projects" (Aliaga et al., p. 4). These recommendations resonate with those previously posed that statistics instruction is most effective when based on real data (Cobb & Moore, 1997) and that explorations with real data sets are more meaningful and instructive when the data are collected by the students themselves (Hogg, 1991). Researchers and educators also contend that statistics students should authentically experience the statistical research process for themselves (Bryce, 2005) and that statistics instruction should be student-centered (Roseth, Garfield, & Ben-Zvi, 2008). Landrum and Smith (2007) echo the call for project-based statistics instruction, stating that the "ideal situation would be [for students] to finish a complete project that included data collection and analysis" (p. 52).

The improved outcomes in statistics education sought by educators and researchers are not limited to statistical knowledge and skills. Student attitudes and beliefs about statistics are inextricably linked to their ability to learn statistics effectively (Gal & Ginsburg, 1994; Ramirez, Schau, & Emmioglu, 2012). Therefore, researchers have paid attention not only to statistics pedagogy that imparts knowledge and skills, but also to teaching techniques that foster positive dispositions. For instance, more constructivist approaches to statistics instruction have been found to improve student attitudes about statistics (Mvududu, 2003). Likewise, in one study, a real-world statistics project not only contributed more to student learning, but also increased students' motivation in the course (Yesilcay, 2000). The student-directed projects that are the focus of this report were intended to address two aspects of student disposition in particular: perceived usefulness of statistics and statistics self-efficacy.

Perceived usefulness refers to students' beliefs about the utility and relevance of statistics not only in the world, but also in their lives and personal endeavors. This construct corresponds approximately to one of four factors measured by the original 28-item Student Attitudes Toward Statistics [SATS] instrument (Schau, 2000). One of six factors in the more recently introduced 36-item version of the instrument, this construct is referred to as "Value" in SATS literature, representing the value that students place on the discipline of statistics (Ramirez et al., 2012; Tempelaar, Schim van der Loeff, & Gijselaers, 2007). Students' perceptions of the usefulness of statistics made an especially fitting target for student-directed projects, since students would choose for themselves the topic and design of their project, increasing the project's personal relevance to the student. Such personal relevance has been found an important contributing factor to students' statistical learning (Mvududu, 2003). This relationship mirrors the observation that students' perceptions of the relevance of statistical data increase their motivation to engage in statistical analysis of those data (Singer & Willett, 1990; Thompson, 1994).

Self-efficacy refers in general to a person's beliefs in their ability to perform in given contexts or to carry out certain tasks (Bandura, 1997). Thus, statistics self-efficacy refers to a student's belief in their statistical abilities (Hall & Vance, 2010). Many researchers have established that students with higher self-efficacy in a given domain tend to exhibit higher performance in that domain, especially in the quantitative disciplines (e.g., Pajares & Miller, 1994). Students with greater self-efficacy in associated domains also demonstrate more effective problem-solving and learning strategies (Cleary, 2006). Researchers have also observed stronger performance in statistics by students with greater statistics self-efficacy (e.g., Castro Sotos et. al., 2009; Hall & Vance, 2010). The instruments used to measure statistics self-efficacy in such studies has varied, depending on what aspect of statistical competence was being investigated. For instance, Finney and Schraw (2003) distinguished between students' confidence in their current statistical ability ("current statistics self-efficacy" or CSSE) and their confidence in their ability to learn certain statistical skills ("self-efficacy to learn statistics" or SELS). It is understandable that instruments for measuring statistics self-efficacy are highly customized, because self-efficacy, even within a given domain (such as statistics) is highly task-specific (Schunk & Pajares, 2002).

Researchers have previously attempted to establish empirical evidence that student-directed projects have a positive impact on student learning and/or student dispositions in statistics. Perhaps one of the more accessible ways to obtain quantitative data on the impact of using projects is to survey students about their experiences with the projects or about their impressions of the extent to which the projects helped them to learn (e.g., Da Silva & Pinto, 2014). Yet such methodologies do not explicitly compare learning and disposition outcomes between students who have used projects to learn statistics and those who have not. Comparisons of this nature are more difficult; not surprisingly, when one such comparison was attempted, the impact of the student projects was found to vary substantially from one instructor to the next, highlighting the potentially confounding effect of different instructors' approaches to project implementation (Spence, Sharp, & Sinn, 2011). Hence, establishing an empirical basis from which to generalize about the projects' benefits has presented somewhat of a challenge. For a study to systematically measure the impact of a particular type of student project across multiple instructors, steps must be taken to ensure that participating instructors are all implementing the projects in essentially the same way.

1.2. PURPOSE OF STUDY AND DESCRIPTION OF STUDENT PROJECTS

The purpose of the present study is to examine the impact of student-directed projects on student learning, attitudes, and beliefs about statistics. Each of these outcome variables is defined explicitly in the next section, which describes research methods used for this study. For the sake of clarity, throughout this article the word "project" will be used only in reference to students' research projects, while the word "study" will refer to the investigation into the impact of these projects on the student outcomes of interest.

A brief summary follows of these student-directed projects, for which the authors have previously published detailed descriptions (see Bailey, Spence, & Sinn, 2013). There are two varieties of project-- one for linear regression analysis and another for a t-test (one sample, two independent samples, or two dependent samples). Students typically work in groups ranging in size from 2 to 4, though some students may work alone by choice. Each project team chooses their own research question, not from a predetermined list of topics, but by brainstorming about their own interests and formulating a question of interest to the team members. The team then determines what variables they need to answer this question and articulates how these variables should be operationalized, measured, calculated, or quantified. Using these variable definitions, students then gather their own project-related data. To accomplish this, some groups write and distribute surveys, while others find data on physical phenomena through direct observation and/or measurement; others locate data on the Internet. The students organize and analyze their data, applying the appropriate descriptive and inferential methods for the data set. The culmination of their project is a formal paper and an in-class presentation, in which they describe their research question, research design, and data collection techniques; summarize their data and the details of their analyses; and state their conclusions in the context of their research question, based on the results of their analyses. The inclusion of these types of projects is consistent with Recommendations 2 and 4 from the aforementioned GAISE College Report (Aliaga et al., 2005), which endorse the use of real data and to foster active learning in statistics classes. The GAISE College Report also suggests that instructors take steps to ensure the students are interested in these real data sets. The student-directed projects meet this recommendation to the greatest extent possible, since the students gather their own data for the purpose of answering questions they themselves have posed.

The curriculum materials used to facilitate these projects were instructor and student guides freely available online at http://faculty.ung.edu/DJSpence/NSF/materials.html.

These materials were written to help carry out student-directed projects in introductory statistics courses (see Bailey et al., 2013). The instructor guide contains recommended project timelines, project proposal formats, and assignments and prompts for students at various project stages. The student guide advises students on each phase of the project, starting with the process of selecting a topic and defining variables, and ending with the requirements for a comprehensive report on their research. Finally, multiple appendices provide project-related resources, including public domain surveys, data-rich websites, and scoring rubrics for evaluating the projects and presentations.

2. RESEARCH METHOD

2.1. PARTICIPANTS

Eight instructors were selected at institutions across the United States and were paid for their participation. Because funding was available for eight instructors, an invitation to participate was sent by e-mail to various contacts available to the researchers, with requests to forward to others. These contacts included college level instructors serving as readers for the Advanced Placement (AP) Statistics exam, alumni of Project NExT through the Mathematical Association of America (MAA), fellow research collaborators on other projects, former colleagues at other institutions, and professional contacts from conferences and workshops. Criteria for participation included prior experience teaching statistics; it was also stipulated that to be qualified to participate, instructors should not have previously used or currently be using projects of this type in their statistics courses.

From the pool of interested applicants, eight instructors were selected with the aim of obtaining the greatest possible diversity in settings, while selecting instructors with reasonable prior experience teaching statistics. The instructors selected had between 6 and 23 years of experience teaching statistics. Four of the instructors were female and four were male. The instructors had varied backgrounds as measured by the field of their terminal degree; two specialized in statistics, four in mathematics, and two in mathematics education. They were employed at three public and five private colleges and universities, located in California, Massachusetts, Minnesota, New York, Oregon, Pennsylvania, South Carolina, and West Virginia. These institutions had undergraduate enrollments ranging from under 1,000 to over 15,000 students, with minority enrollments between 14% and 39%. The introductory statistics class sizes in the study ranged from 18 to 60 students.

Because the goal of the study was to measure the effectiveness of the project-based approach and the supporting materials across a broad spectrum of geographic and academic settings, there are very few unifying characteristics of the students in the statistics classes studied. All were enrolled as undergraduates in an introductory non-calculus statistics course at a college or university in the U.S.

2.2. STUDY DESIGN

The eight instructors described above participated in a quasi-experimental study including both control and treatment. Each instructor first taught his or her statistics course as he or she usually would and without the inclusion of such student-directed research projects; this semester will be henceforth referred to as the control semester. All of these instructors were at that time experienced statistics instructors; however, it is worthwhile to note that they all reported not having implemented projects of this nature previously in their teaching. Thus, the instructors' usual approaches to teaching were unchanged for the control semester. After their control semester, the instructors attended a workshop in which they were given access to the curriculum materials described previously; they practiced completing mock student projects and became acquainted with the materials and requirements for incorporating student-directed projects into their own courses. In a subsequent semester, these same instructors taught their statistics courses using methods and materials to incorporate student-directed projects according to the guidelines provided in the workshop. Each instructor implemented both types of project (i.e., linear regression and t-test). Use of the prescribed instructional materials and adherence to the guidelines was confirmed by e-mail communication, telephone interviews, journal prompts, and classroom visits, which were conducted to collect data for a simultaneous qualitative investigation.

The semesters in which the instructors included student-directed projects in their courses are referred to as treatment semesters. During the initial planning phase of the study, only a single treatment semester was planned for each instructor. Hence, participating instructors only agreed to be part of the study through the end of their single treatment semester. However, as the initial treatment semester progressed, some concern arose surrounding the instructors' level of familiarity and comfort with the materials and with facilitating such projects. The instructors were implementing these projects for their very first time. This lack of experience with a new technique could easily have constrained an instructor's effectiveness, as well as the effectiveness of the projects themselves. By contrast, because each instructor had several years of experience teaching statistics, their control semesters likely reflected teaching strategies that the instructors had refined over a number of years. It was easy to imagine that any potential benefit gained by introducing these projects would be obscured by the instructors' learning curves. Therefore, although they had only committed to carry out a single treatment semester, instructors were invited to continue their involvement in the study beyond the initial treatment semester by conducting additional treatment semesters, during which their familiarity with the materials and the projects would likely increase. Therefore, the treatment semesters are designated as Treatment 1, Treatment 2, or Treatment 3, identifying the first, second, or third semester during which an instructor taught a treatment section.

2.3. DEPARTURES FROM PROTOCOL AND EXTENDED PARTICIPATION

Of the eight participating instructors, three departed from established protocols during the study. These breaches in protocol can be summarized as teaching a class during a treatment semester that was fundamentally different from that taught during the control semester. The validity of the study depended greatly on each instructor's treatment and control classes being as similar as possible in target audience, course content, and course format. In some instances, circumstances leading to the protocol exception were unanticipated events outside the instructor's control (e.g., one instructor's spouse was transferred to another geographical area; the only available options for that instructor's treatment class were a short summer session version of the course before the instructor moved away, or a treatment semester at a new institution that was not necessarily comparable to the one where the control semester was conducted.) In another instance, it was determined that one instructor taught a control semester which was not the expected elementary statistics course for non-majors, but a statistical methods course for students likely to major in statistics. Hence, the instructor's control and treatment classes were not comparable. Those instructors who were unable to maintain protocol were identified in the data set, and analyses were conducted both with and without their data. It is important to note that the decision was made in advance to conduct analyses both ways. Therefore, results are presented for two groups: "all instructors" or "instructors within protocol," the second group excluding those that broke protocol.

The five instructors who remained within protocol were invited to continue their participation in the study by conducting additional treatment semesters after their initial treatment semester ended. Although all five instructors indicated that they planned to use the projects again, the courses that certain instructors had arranged to teach during subsequent semesters were in some way no longer comparable to the course they had taught in their control and initial treatment semesters. One instructor was changing texts; another was teaching a new specialized version of the course specifically for the life sciences; another was piloting a new second course in statistics that did not qualify as an introductory course. For the same reason that the three instructors who did not maintain protocol were excluded, instructors were no longer asked to participate in another treatment class when they could not ensure that the class would be comparable in content and/or audience to their control class. Therefore, of the five who were approached about extending their participation, only three instructors were able to conduct an additional treatment class for a second semester, and only one was able to conduct a third treatment semester.

Although valuable data were collected through the inclusion of additional treatment semesters beyond the first, the authors acknowledge the limitation imposed by the fact that these additional treatment semesters were not part of the originally planned study. Had the initial plan entailed repeated treatment semesters, the study would probably have been much stronger. However, the value introduced by examining repeated treatment semesters was sufficient to warrant the unplanned extension. The additional treatment semesters were organized in the same manner as the initial treatment; the instructors used the curriculum materials to facilitate student-directed projects and then administered the survey and content knowledge assessments in their classes.

2.4. VARIABLES

The control and treatment designations outlined above comprised the primary explanatory variable in this study. An additional explanatory variable was the instructor's level of experience implementing the treatment (e.g., Treatment 2 indicates treatment in which the instructor had more experience with the projects than in Treatment 1, and so on.)

The three main outcome variables were students' content knowledge, statistics self-efficacy, and perceived usefulness of statistics. For all three outcome variables, careful attention was given to developing and revising instruments to align well with the anticipated improvements in student outcomes when the class included these projects. Before describing the process of developing the instruments, we first give the intended definition for each variable, both to reiterate the foci and expected outcomes of the projects, and to support the contention that existing instruments were not satisfactory to measure each variable as defined.

Content knowledge refers to a student's demonstrated knowledge and understanding of statistics learning objectives within the scope of the projects. Concepts that were addressed by carrying out the projects included four broad categories of knowledge: linear regression, hypothesis testing with t-tests, sampling, and recognizing which type of analysis was appropriate for a given scenario.

Likewise, statistics self-efficacy is defined as the student's belief in their ability to understand and use statistics, specifically within the scope of the intended outcomes of the projects. Thus, the construct encompasses student self-beliefs about statistical abilities that should have been reinforced by these projects. In particular, the tasks that were required of them as they conducted the projects included carrying out and interpreting linear regression analyses, carrying out and interpreting t-tests, planning and carrying out data collection, and more broadly, learning to apply new statistical ideas as they were introduced.

Finally, perceived usefulness describes a student's perception of how much they personally may benefit from understanding statistics, as well as their perception of the overall importance of statistics as a discipline. The articulation of perceived usefulness was based on the intention that students would come to appreciate the utility of statistics through their personal experience using it to investigate a topic of their own choosing. Therefore, perceived usefulness of statistics was defined to include students' sense of the relevance of statistics to their own pursuits and to the world outside the classroom. This personal relevance also included the students' beliefs about whether they would themselves benefit from understanding statistical ideas.

2.5. DEVELOPMENT OF INSTRUMENTS

The present study is the third stage of our investigation into the impact of these types of projects on student outcomes. These stages have included a preliminary study at a single institution, a local study at three institutions, and the current national study at eight institutions. The development of the instruments used in the present study was an iterative process that took place as the stages progressed; this evolution and the stages themselves have been described previously (see Bailey et al., 2013). In the regional study, the three outcomes of interest were identified explicitly as content knowledge, perceived utility, and statistics self-efficacy; a separate instrument was created for each, and all three instruments were validated prior to their use in the study. The development and validation of these original instruments has also been described previously (see Spence et al., 2011). After the regional study and prior to the national study, the instruments were reviewed, taking into consideration student data from each instrument, input from the pilot instructors involved in the regional study, and input from an advisory panel of professionals involved in some aspect of statistics education. This panel was convened to provide input into all aspects of the current study before it took place. Based on these reviews, all three instruments were revised before starting the present study; once again, all three instruments were also validated prior to the start of the present study. We discuss the development and revision process below, providing background from the regional study where needed to explain the revisions that were made to the instruments.

Content Knowledge Assessment. The content knowledge assessment was intended to measure student knowledge about t-tests and linear regression in particular, since these were the types of projects that students conducted. Therefore, no existing content knowledge assessments were sufficiently specific. For instance, the Statistics Concept Inventory addresses many other topics, such as descriptive statistics and probability (Allen, 2006). Therefore, an assessment targeting the desired content was constructed with multiple choice questions similar to items on the Statistics Concept Inventory and items on recent Advanced Placement Statistics exams. The assessment developed for the regional study contained 18 questions, with 12 questions in some way related to hypothesis testing with t-tests and 6 questions related to linear regression. The questions pertaining to t-tests were classified in one of two groups— those about "usage" (how a t-test is carried out) and those about "inference" (interpreting the results of the test).

When the assessment was reviewed after the regional study, two main areas for improvement were identified: 1) some of the assessment items were not well aligned with the content that students would have learned specifically by participating in the projects; and 2) some of the assessment items targeted more than one concept simultaneously, making it more difficult to interpret what students did and did not know. For instance, five of the "usage" questions emphasized the ability to distinguish in some way between the z and t statistics; four of those same questions also targeted students' ability to distinguish between dependent and independent samples scenarios. Another of the questions required students to correctly interpret the meaning of a given percentile value, a skill not directly promoted or required by the projects. When the test was revised, items were rewritten or removed if they did not address concepts explicitly applied during the execution of the projects; hence, the revised assessment did not target a student's ability to interpret percentiles or to distinguish between z and t statistics.

However, items were added to the assessment to measure potential benefits that were not measured in the first test. Because pilot instructors routinely noted that students seemed to gain a much better understanding of statistical sampling through their project work, items were added to measure the students' understanding of sampling. Likewise, items were added to assess a student's ability to distinguish not only between independent and dependent samples t-tests, but also between contexts suitable for t-tests and those suitable for linear regression.

All of the modifications yielded a 17-item content knowledge assessment; the revisions substantially improved the alignment of the test with the expected benefits of the projects, so that these benefits could be more reliably measured. In addition, the revisions reflected an explicit focus on four major categories of content knowledge that were perceived as primary understandings fostered by the projects. These four categories were therefore defined as subscales when the instrument was revised, prior to its use in the present study (the national pilot test). These four subscales were linear regression, hypothesis testing with t-tests, recognition of the appropriate analysis, and sampling.

Self-efficacy Instrument. Much like the content knowledge assessment, the instrument for measuring statistical self-efficacy was intended to target those student self-beliefs that may have been strengthened by participating in the projects. Many statistics self-efficacy scales can be found in the literature (e.g., Castro Sotos et al., 2009; Finney & Schraw, 2003; Hall & Vance, 2010). However, each of these scales is specific to some particular domain(s) of statistical competency, reflecting the well-established finding that self-efficacy constructs are highly domain-specific and task-specific (Schunk & Pajares, 2002). Correspondingly, the present study required its own task-specific instrument to measure the appropriate self-beliefs in context.

For the regional study, a 14-item scale was constructed using 6-point Likert style scoring. The scale was developed following criteria and sample items provided in the *Guide for Constructing Self-Efficacy Scales* (Bandura, 2006). Four items addressed the student's self-beliefs about their ability to understand statistics (sample item: "I am confident that I understand basic statistics concepts.") Five items were devoted to self-beliefs about tasks specific to linear regression (sample item: "I am confident that I can identify outliers and influential points in a scatterplot and predict their influence on the correlation coefficient *r*.") The remaining five items targeted self-beliefs about tasks specific to hypothesis testing with t-tests (sample item: "I am confident that I can construct a t-test to compare the means of two populations using data collected from two independent samples.")

As with the content knowledge assessment, a review of this instrument after the fact revealed an opportunity to improve its alignment with the expected outcomes of the projects. For instance, although the project provided students hands-on experience collecting data, the original instrument had no items addressing a student's self-beliefs about data collection. It was also noted that for overall statistical self-beliefs, it was likely more appropriate to address the student's belief in their ability both to *use* statistics and to *learn* statistical ideas than to understand statistics concepts in general. This observation speaks both to the task specificity of the student's self-belief and to the notion that by participating in a project, students should experience learning some set of statistical ideas, an experience they could then be more confident in repeating in the future. These observations led to the revision of the self-efficacy instrument to a 16-item survey with a focus on four areas, corresponding in part to the four themes identified for the content knowledge instrument. These areas, which were identified as subscales when the instrument was revised, targeted student self-beliefs regarding linear regression, hypothesis testing with t-tests, data collection, and the ability to learn and use statistics.

Perceived Usefulness Instrument. To measure student perceptions of the usefulness of statistics, a third instrument was needed. As noted previously, the Survey of Attitudes Toward Statistics (SATS) scales measure a related construct called "value" in addition to several other attitudinal factors that were beyond the focus of this study. Likewise, the variable of interest was related to only one of multiple constructs measured by a number of other instruments; similar constructs were described with terms such as "present utility", "professional utility", "perceived worth", and "usefulness" on various instruments measuring four or more constructs (Ramirez, Schau et al., 2012). Other attitude instruments were even less applicable; for instance, the Attitudes Toward Statistics (ATS) scale addressed not only students' beliefs about statistics itself, but also their attitudes about their statistics courses, including their affective reactions to studying statistics (Wise, 1985). Hence, although some overlap was evident between factors measured by these instruments and our construct of perceived usefulness, the existing instruments were not a good match for our purposes.

For the regional study, a 12-item instrument was developed using 12 Likert-style items on a 6-point scale. The instrument included a focus on the utility of statistics in students' future careers, reflecting the fact that in the regional study, students were encouraged to select career-specific project topics. At the conclusion of this study, it was observed that many of the students did not yet have a clear sense of the career they wished to pursue. Analysis of the data from this instrument seemed consistent with this observation. Therefore, in the revised 11-item instrument, the career-specific focus was replaced with a focus on how statistics are used and encountered in the world, in the media, and in daily life. Items were also added to address more explicitly the student's belief that they would personally benefit from understanding statistical concepts.

2.6. PROCEDURE AND INSTRUMENTS ADMINISTERED

For the present study, content knowledge is defined as the student's score on the revised 17-item multiple choice assessment described above. Each content knowledge question was recorded as answered either correctly (1) or incorrectly (0). Thus, the score on the assessment was the number of questions a student answered correctly, with possible scores ranging from 0 to 17. The KR-20 values obtained in the study for the main scale and all 4 subscales are given in Table 1; the use of the KR-20 as a reliability measure emphasizes the dichotomous nature of the data from this instrument, as no partial credit was available on these items. Given the small number of items, the low KR-20 values obtained are not surprising (Tucker, 1949). These measures primarily reflect the fact that the instrument and its subscales were quite brief. In particular, for a 17-item instrument, the KR-20 of 0.654 for the content knowledge assessment overall was deemed satisfactory.

Table 1. Content Knowledge, subscales, descriptions and KR-20 values

Variable	Description	Items	KR-20
CK-Total	Main scale: All content knowledge ítems	17	0.654
CK-LR	Subscale: Linear regression	7	0.583
CK-HT	Subscale: Hypothesis testing (t-tests)	5	0.343
CK-Rec	Subscale: Recognizing appropriate type of analysis	3	0.331
CK-Sam	Subscale: Sampling	2	0.218

Statistics self-efficacy is defined for the present study as the student’s score on the revised self-efficacy instrument, which consisted of 16 Likert-style items on a 6-point scale, with 1 representing “strongly disagree” and 6 representing “strongly agree”. The statistics self-efficacy score is the sum of the responses on all 16 items; thus, possible scores range from 16 to 96. Sample items from the instrument are shown in the appendix, including at least one sample item representing each subscale. A Cronbach alpha of 0.926 was obtained for this scale. The Cronbach alphas for all subscales of this instrument exceeded 0.7, an accepted threshold for good consistency (Kline, 2000.)

Perceived usefulness is defined for the present study as the student’s score on the revised perceived usefulness scale, which contained 11 Likert-style items on a 6-point scale. Thus, possible scores range from 11 to 66. Four of these items were reverse scored. A sample item is “Statistics can be used to answer important questions.” A sample reverse-scored item is “Once I finish this course in statistics, I probably won’t use any of the concepts I learned in the class.” A Cronbach alpha of 0.918 was obtained for the scale.

Both the self-efficacy and perceived usefulness scales were combined into a 27-item attitudes and beliefs survey. Both constructs and the self-efficacy subscales are shown in Table 2, along with the corresponding reliability coefficients.

Table 2. Attitude and belief survey scales and subscales with reliability coefficients

Variable	Description	Items	Cronbach alpha
SE-Total	Main scale: All statistics self-efficacy ítems	16	0.926
SE-LR	Subscale: Linear regression	5	0.852
SE-HT	Subscale: Hypothesis testing (t-tests)	5	0.872
SE-DC	Subscale: Data collection	4	0.798
SE-Gen	Subscale: Ability to learn and use statistics in general	2	0.867
PU	Perceived usefulness	11	0.918

At the end of every semester, including the control semester, the students were asked to complete the attitudes and beliefs survey, followed by the content knowledge assessment. The attitudes and beliefs survey was administered first because it measured students’ self-beliefs about their statistical knowledge, which was itself then measured with the content knowledge assessment.

2.7. METHODS OF ANALYSIS

Descriptive statistics include mean and standard deviation of each outcome variable (i.e., students' content knowledge, statistics self-efficacy, and perceived usefulness of statistics) in both control and treatment settings. Correlation coefficients among the outcome variables were also computed within each setting, including correlations among subscales of each main scale.

A linear mixed effects model was used to analyze each of the outcome variables with treatment (Control, First Semester, Second Semester or Third Semester) as the fixed effect and instructor and participant within instructor by treatment as random effects. A follow-up analysis to a significant overall test was conducted to compare the mean for all students in treatment classes to the mean for those in control classes. SAS v9.3 was used for these analyses, which were carried out for all instructors and again for only those who remained within protocol.

The mixed effects model was selected to examine all participating students as a group, taking into consideration effects introduced by different instructors. This analysis was preferred because relatively small individual class sizes yielded insufficient power to detect results for the effect sizes anticipated. However, for those interested in individual instructor outcomes, t-tests were also used to compare control and treatment means by instructor.

For those instructors who were available to complete treatment classes for a second and/or third semester, additional analyses were conducted to further explore another predictor of the student outcomes of interest-- the length of time the instructor had been using the treatment. Pairwise t-tests were conducted to compare outcomes among the control semester and those of Treatment 1, Treatment 2, and Treatment 3.

3. RESULTS

3.1. DESCRIPTIVE STATISTICS

Table 3 shows the mean and standard deviation for each outcome under control and treatment conditions for all instructors and for those within protocol only. Because the variables are measured on different scales, the table includes an additional column showing the mean as a percentage of the maximum possible score for each variable or subscale.

The reader may observe that within the "All Instructors" category, control group means are consistently slightly higher than the corresponding control group means for instructors who maintained protocol. A likely factor contributing to this pattern is the instructor previously noted whose control course was not the expected introductory statistics for non-majors. This observation supports the notion that conducting analyses with only those instructors who maintained protocol is more likely to address the intended research questions. Nevertheless, analyses were still conducted for all instructors, as they were deemed to have some benefit.

Among the in-protocol instructors, the greatest gains in content knowledge from control to treatment (as mean percentages of the highest possible score) appear in the areas of linear regression (9 percentage point gain) and recognition of the appropriate analysis for a given data set (13 point gain). Curiously, within the same in-protocol group, the largest gain in self-efficacy was in the area of hypothesis testing (7 point gain), an area which saw very little increase in content knowledge (1 point gain).

Table 3. Means and Standard Deviations by Setting for All Instructors and for In-Protocol Instructors Only

		All Instructors N=353 Control N=441 Treatment			Within Protocol Only N=198 Control N=344 Treatment		
Variable	Setting	Mean	SD	Mean as % of max	Mean	SD	Mean as % of max
CK-Total	Control	8.35	2.97	49%	7.52	2.83	44%
	Treatment	8.66	3.20	51%	8.63	3.33	51%
CK-LR	Control	3.17	1.83	45%	2.50	1.61	36%
	Treatment	3.26	1.79	47%	3.12	1.83	45%
CK-HT	Control	2.29	1.22	46%	2.26	1.21	45%
	Treatment	2.25	1.29	45%	2.30	1.31	46%
CK-Rec	Control	1.33	.89	44%	1.23	.87	41%
	Treatment	1.54	.99	51%	1.61	1.01	54%
CK-Sam	Control	1.55	.61	78%	1.53	.64	77%
	Treatment	1.61	.59	81%	1.60	.59	80%
SE-Total	Control	78.51	11.60	82%	76.00	12.98	79%
	Treatment	80.20	11.62	84%	80.13	11.82	83%
SE-LR	Control	25.75	3.68	86%	25.05	3.95	84%
	Treatment	25.70	4.08	86%	25.64	4.15	85%
SE-HT	Control	23.41	5.34	78%	22.08	6.01	74%
	Treatment	24.43	4.37	81%	24.38	4.47	81%
SE-DC	Control	19.12	3.29	80%	18.86	3.65	79%
	Treatment	19.98	3.06	83%	20.04	3.05	84%
SE-Gen	Control	10.23	1.73	85%	10.01	1.89	83%
	Treatment	10.09	1.79	84%	10.06	1.82	84%
PU	Control	50.86	9.50	77%	50.19	10.47	76%
	Treatment	51.43	9.47	78%	51.55	9.79	78%

Also worth noting is that in both control and treatment settings, all attitude and self-belief outcomes were markedly higher than the content knowledge outcomes, when considered as percentages of the highest possible score. All affective outcomes ranged from 70 to 85% on average, whereas most content knowledge outcomes hovered around an average of 50%.

Table 4 shows correlations between the primary outcome variables in the study by treatment group (control and treatment) across all instructors in the study. Included are the content knowledge (CK) and self-efficacy (SE) subscales; a box surrounds each variable group (CK and SE), identifying all correlations among subscales in that group.

Correlations among self-efficacy (SE) subscales were fairly strong in both settings, whereas correlations among content knowledge (CK) subscales were notably weaker. In both settings, SE correlated more strongly with PU than did CK.

3.2. PRIMARY ANALYSIS

For the linear mixed effects model, all significant results favored the treatment, and included multiple content knowledge and self-efficacy measures. There were no significant differences for the perceived usefulness scale. Table 5 below summarizes the significant results across all instructors.

Table 4. Correlations between Outcome Variables by Setting

	1	2	3	4	5	6	7	8	9	10	11
1 CK-Total						0.36*	0.39*	0.33*	0.22*	0.30*	0.39*
2 CK-LR			0.43*	0.19*	0.28*	0.35*	0.39*	0.31*	0.19*	0.31*	0.33*
3 CK-HT		0.23*		0.12	0.25*	0.24*	0.24*	0.25*	0.14^	0.18$^+$	0.28*
4 CK-Rec		0.21*	0.04		0.11	0.13^	0.14^	0.10	0.09	0.12	0.17$^+$
5 CK-Sam		0.27*	0.24*	0.08		0.16$^+$	0.16$^+$	0.14^	0.13^	0.10	0.22*
6 SE-Total	0.39*	0.37*	0.20$^+$	0.09	0.26*						0.51*
7 SE-LR	0.42*	0.43*	0.17^	0.10	0.29*			0.69*	0.67*	0.73*	0.44*
8 SE-HT	0.32*	0.29*	0.21*	0.07	0.17^		0.52*		0.66*	0.65*	0.44*
9 SE-DC	0.23*	0.22*	0.10	0.05	0.20$^+$		0.60*	0.51*		0.60*	0.44*
10 SE-Gen	0.29*	0.25*	0.16^	0.08	0.21*		0.62*	0.61*	0.60*		0.48*
11 PU	0.30*	0.27*	0.18$^+$	0.02	0.26*	0.60*	0.49*	0.43*	0.59*	0.54*	

Below diagonal: Correlations for Control classes (N=353)
Above diagonal: Correlations for Treatment classes (N=441)

* $p < .0001$
$^+ p < .001$
^ $p < .01$

Table 5. Significant Results for All Instructors

Variable	Treatment-Control Mean (SE)	F(df1,df2),p	t(df),p	Cohen's d
CK-Total	0.74 (0.34)	F(3,763)=3.52, p=0.0148	t(763)=2.16, p=0.0307	d=0.0998
CK-LR	0.51 (0.17)	F(3,763)=5.20, p=0.0015	t(763)=2.92, p=0.0036	d=0.0488
CK-Rec	0.26 (0.09)	F(3,763)=3.04, p=0.0283	t(763)=2.77, p=0.0057	d=0.2202
SE-HT	1.86 (0.62)	F(3,763)=5.22, p=0.0014	t(763)=2.98, p=0.0030	d=0.2127
SE-DC	0.79 (0.36)	F(3,763)=4.69, p=0.0030	t(763)=2.18, p=0.0292	d=0.2713

These results indicate that treatment students had overall content knowledge significantly higher than that of students in the control semester, on average. More specifically, for knowledge of linear regression and ability to choose the appropriate form of analysis, students in the treatment semesters significantly outperformed their counterparts in the control semester. Students in all the instructors' treatment classes showed significantly higher average self-efficacy for hypothesis testing and data collection than students in the control classes.

Although the results for all instructors comparing treatment and control groups (Table 5) are statistically significant, it is worth noting that even the largest of the effect sizes (measured by Cohen's d) can only be interpreted as a small effect, and those effect sizes less than 0.2 are below the typical threshold for even a small effect (Cohen, 1988; Olejnik & Algina, 2000). The researchers were aware that any weakness evident in the effects could be due in part to the breaches in protocol by a few instructors, as described above. Correspondingly, the data were analyzed only for the instructors who remained within protocol during the study (Table 6).

Table 6. Significant Results for Instructors within Protocol

Variable	Treatment-Control Mean (SE)	F(df1,df2), p	t,p	Cohen's d
CK-Total	0.97 (0.39)	F(3,520)=3.66, p=0.0124	t(520)=2.49, p=0.0132	d=0.3521
CK-LR	0.63 (0.19)	F(3,520)=5.66, p=0.0008	t(520)=3.25, p=0.0012	d=0.3539
CK-Rec	0.39 (0.09)	F(3,520)=6.74, p=0.0002	t(520)=4.16, p<0.001	d=0.4019
SE-Total	2.57 (1.13)	F(3,520)=1.88, p=0.1315	t(520)=2.27, p=0.0239	d=0.3372
SE-HT	2.64 (0.76)	F(3,520)=6.09, p=0.0004	t(520)=3.46, p=0.0006	d=0.4530
SE-DC	0.79 (0.33)	F(3,520)=3.52, p=0.0149	t(520)=2.43, p=0.0153	d=0.3594

The results for the in-protocol instructors are similar to those shown previously for all instructors. The in-protocol treatment semester students also reported significantly higher overall self-efficacy than did their counterparts in the control. Further, these significant results are all accompanied by effect sizes between 0.3 and 0.5, closer to accepted thresholds for medium effect sizes (Cohen, 1988; Olejnik & Algina, 2000).

3.3. COMPARISON OF OUTCOMES BY INSTRUCTOR

To allow the reader to examine individual instructor outcomes, Table 7 shows the comparison of control and treatment groups for each outcome by instructor. For each comparison, the difference between treatment and control is reported as an effect size (using Cohen's *d*), and the 2-tailed significance (*p*) is given. Equal variances were not assumed. Positive effect sizes favor treatment, whereas negative effect sizes favor control. For those instructors who were not within protocol, results are shown, but those columns are shaded to remind the reader that the control and treatment groups were unlikely to represent comparable populations, and hence, a control vs. treatment comparison between the two groups is probably confounded by other differences between the groups.

Table 7. Treatment Effects on Student Outcomes by Instructor

	Instructor							
	#1	#2	#3	#4	#5	#6	#7	#8
In protocol	YES	YES	YES	YES	NO	NO	YES	NO
Sample size								
Control	42	33	58	18	56	36	47	63
Treatment	82	53	55	21	31	16	133	50

Treatment Semesters		2	2	1	1	1	1	3	1
CK-Total	*d*	0.090	0.320	0.075	0.498	– 0.117	0.076	0.123	– 0.659
	p	0.615	0.141	0.693	0.131	0.602	0.810	0.466	**0.001
CK-LR	*d*	0.194	0.337	– 0.126	0.717	– 0.137	0.228	0.115	– 0.551
	p	0.299	0.123	0.535	*0.032	0.537	0.480	0.469	**0.005
CK-HT	*d*	– 0.201	0.077	– 0.250	– 0.421	– 0.008	– 0.483	– 0.187	– 0.443
	p	0.262	0.736	0.195	0.202	0.959	0.094	0.297	*0.022
CK-Rec	*d*	0.409	0.168	0.501	0.298	– 0.328	0.125	0.454	– 0.231
	p	*0.030	0.460	**0.009	0.363	0.134	0.698	**0.009	0.227
CK-Sam	*d*	– 0.346	0.187	0.106	0.816	0.262	0.754	– 0.308	– 0.273
	p	0.053	0.443	0.587	*0.020	0.243	**0.009	0.051	0.164
SE-Total	*d*	0.218	0.414	0.262	– 0.184	0.818	– 0.156	0.059	– 0.617
	p	0.254	0.058	0.166	0.572	**0.000	0.624	0.712	**0.002
SE-LR	*d*	– 0.103	– 0.371	– 0.145	0.429	0.415	0.075	0.402	– 0.672
	p	0.576	0.091	0.444	0.187	0.050	0.814	*0.020	**0.001
SE-HT	*d*	0.613	1.133	0.390	– 0.748	0.705	– 0.232	– 0.321	– 0.554
	p	**0.003	**0.000	*0.040	*0.025	**0.002	0.438	*0.045	**0.005
SE-DC	*d*	0.017	– 0.023	0.697	0.107	0.932	– 0.050	0.177	– 0.307
	p	0.927	0.914	**0.000	0.750	**0.000	0.861	0.296	0.112
SE-Gen	*d*	– 0.094	0.035	– 0.134	– 0.376	0.580	– 0.627	– 0.130	– 0.485
	p	0.600	0.888	0.479	0.252	*0.014	0.057	0.425	*0.013
PU	*d*	– 0.265	– 0.202	0.206	– 0.216	0.254	– 0.217	0.037	– 0.195
	p	0.160	0.377	0.277	0.516	0.256	0.487	0.825	0.305

Positive effect sizes favor treatment; negative effect sizes favor control.

3.4. INSTRUCTOR TIME USING TREATMENT

The remaining analyses only apply to the three instructors who were available to extend the study beyond its originally planned conclusion by conducting additional treatment semesters after their initial treatment semester was completed. Figure 1 shows mean scores from control semester to the second treatment semester; scores are represented as percentages of the max, allowing all outcomes to be displayed comparably on one plot. The only variables plotted are those for which trends are visibly evident or for which statistically significant differences were detected.

It is important to note that Figure 1 shows only the averages for instructors who remained involved in the study for more than one treatment semester; hence, the implications of the chart are not simply the result of drop-out bias. Treatment 3 is also not included in Figure 1 because only one instructor completed a third treatment semester. A similar plot is shown in Figure 2 solely for the instructor who completed all three treatment semesters.

For instructors who conducted additional treatment semesters after the first, analysis suggested that some outcome variables showed a decrease from the control semester to the first treatment semester, followed by an increase during subsequent treatment semesters that often matched or exceeded the initial decline. Although some other outcomes did not initially decline, they saw minimal or modest gains during the first treatment semester, but then continued to improve in a later treatment semester. This trend is particularly evident for some content knowledge outcomes (Figures 1 and 2).

A likely factor playing a role in this trend is the pilot instructors' own level of familiarity with the materials and with facilitating these kinds of projects, as previously discussed. As the instructors continued to facilitate student-directed projects, their familiarity with project implementation methods undoubtedly increased; during the same time, data reveal increases in many student outcome measures in the subsequent treatment semesters.

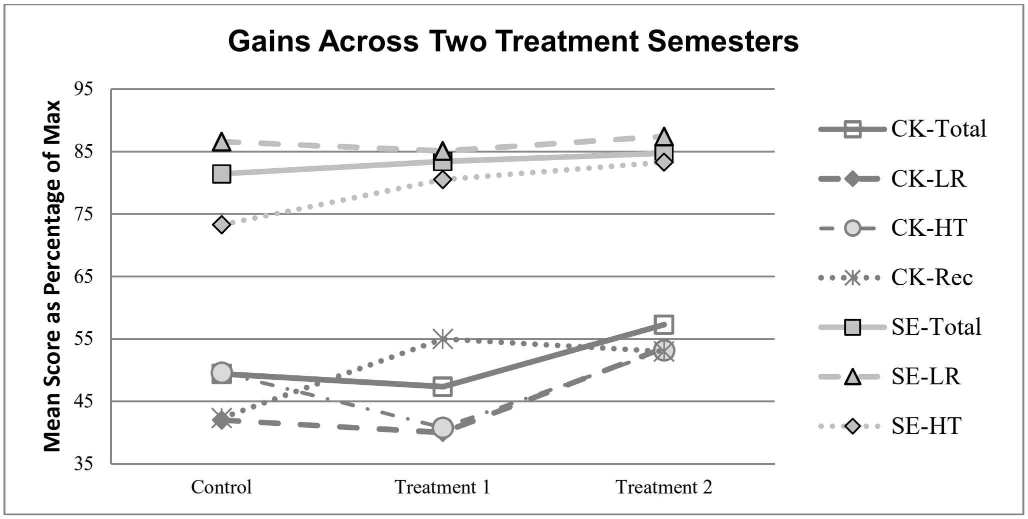

Figure 1. Average variable scores over three semesters for instructors participating in control and treatment semesters 1 and 2

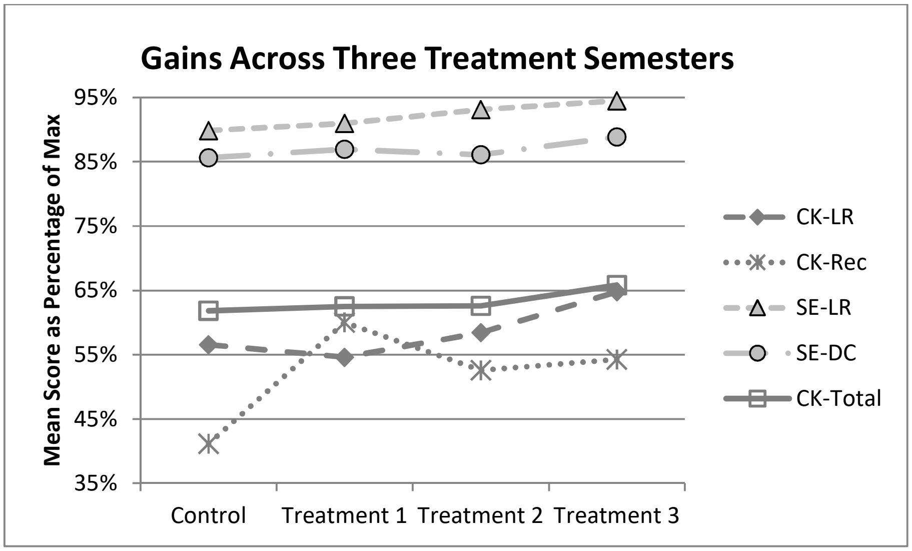

Figure 2. Average variable scores over four semesters for instructor who completed control and three treatment semesters

For all instructors who completed two treatment semesters, pairwise tests revealed significant differences between the control and each treatment semester, as well as between the first treatment and the second (Table 8).

Table 8. p-Values for Significant Pairwise Gains: Instructors with Two Treatments

Variable	Control versus Treatment 1	Control versus Treatment 2	Treatment 1 versus Treatment 2
CK-Total		$p = .001$	$p = .016$
CK-LR		$p = .000$	$p = .002$
CK-Rec		$p = .002$	
SE-Total	$p = .046$		
SE-HT		$p = .001$	
SE-DC	$p = .003$		

Control N=122 Treatment #1 N=122 Treatment #2 N=103

The only comparison possible with 'Treatment 3' is a comparison for a single instructor, as the other instructors completed their participation after the second treatment semester. For the instructor who completed a third treatment semester, there were significant pairwise gains from the control to each treatment semester, as well as gains from the first treatment semester to subsequent treatment semesters (Table 9).

Among instructors who conducted additional treatment semesters after the first, most content knowledge gains emerged in the second or third treatment semester. The only exception to this trend is CK-Rec, the recognition of the most appropriate type of analysis for a given context or data set; this gain was evident even in the first treatment class for the instructor who conducted a treatment class for three semesters. The most consistent content knowledge gains for all instructors who repeated treatment semesters were in the areas of linear regression and recognition of appropriate types of analysis.

Table 9. p-Values for Significant Pairwise Gains: Instructor with Three Treatments

Variable	Control versus Treatment 1	Control versus Treatment 2	Control versus Treatment 3	Treatment 1 versus Treatment 2	Treatment 1 versus Treatment 3
CK-LR			$p = .035$		$p = .029$
CK-Rec	$p = .003$	$p = .043$	$p = .032$		
SE-LR		$p = .022$	$p = .001$		$p = .009$
SE-DC			$p = .047$		

Control N=47 Treatment #1 N=45 Treatment #2 N=45 Treatment #3 N=43

4. DISCUSSION AND RECOMMENDATIONS

The study results indicate that the inclusion of student-directed projects tends to improve students' knowledge of statistics in selected domains. The students in classes with student-directed projects showed slightly higher overall content knowledge than did those in classes without student-directed projects, with strongest gains in the areas of linear regression and identifying the appropriate form of analysis for a given scenario. Results also indicate that students who complete student-directed projects may benefit from stronger self-beliefs in various domains of statistical practice, including the areas of hypothesis testing and data collection methods.

Least surprising is the observation that student-directed projects have the capacity to sharpen students' recognition of appropriate methods of analysis and to boost their beliefs in their own ability to collect data. These two outcomes are among the most direct ways that these projects engaged students with statistics. The students conducted two different kinds of analysis for themselves after formulating objectives and actively collecting data, so they could experience the differences between the two types of analysis. Perhaps more satisfying is the observation that students demonstrated content knowledge gains in linear regression by actively engaging in their own linear regression projects. These results are consistent with many previous findings and recommendations that students' statistics learning can be improved through student-centered authentic projects with real data (e.g., Bryce, 2005; Hogg, 1991; Roseth et al., 2008).

Another result that warrants attention is the evidence regarding the impact of these projects on students' ability to perform hypothesis testing. The largest overall effect size is for self-efficacy in hypothesis testing, suggesting that students who engage in these projects are more confident in their ability to conduct a hypothesis test. Yet no corresponding increase in content knowledge about hypothesis testing is evident in these data. A plausible explanation for this discrepancy is that the statistical reasoning underlying hypothesis testing was not a focus of these projects. Rather, as the materials used to implement the projects suggest, the focus was on carrying out a number of project-related tasks; although these tasks included conducting a hypothesis test, they did not necessarily foster the reasoning associated with such a test. It is conceivable that students increased their comfort level with the steps of a hypothesis test (stating hypotheses, finding the test statistic, reporting the statistic and p-value, and stating a conclusion) by carrying these steps out in their own projects, while doing little to increase their true understanding of how or why such a test works. The distinction between procedural knowledge and conceptual understanding is well-known to educators and researchers (e.g., Rittle-Johnson & Schneider, 2015; Whitaker, Foti, & Jacobbe, 2015). Further, statistics educators have paid a fair amount of attention to developing assessments that target conceptual understanding (e.g., delMas, Garfield, Ooms, & Chance, 2007). The discrepancy between students' reported self-beliefs and their demonstrated knowledge points to the possibility that students responded to the self-efficacy questions with reference to their procedural knowledge, but were unable to demonstrate conceptual understanding on the assessment.

The above observation highlights the need for an instructional approach that will not only engage students in statistical tasks, but also illuminate the reasoning behind statistical inference in a meaningful way. The content knowledge gains reported here fall short of showing any improvement in students' conceptual understanding of statistical inference. As such, the methods and materials used to facilitate these projects might be improved by the introduction of learning tasks with a greater focus on statistical reasoning. For instance, many researchers and educators now propose an approach based on randomization and simulation (e.g., Rossman & Chance, 2014). Early investigations of this method suggest that it can increase student knowledge in several areas, with tests of significance among the domains most favorably improved (Tintle et al., 2014). Obviously, the simulation-based approach and the use of student-directed projects are not mutually exclusive. To continue this line of inquiry, researchers would do well to explore the impact of instruction using both of these methods, and possibly gauge the benefit of adding each approach to the other.

Results across repeated treatment semesters also suggest that certain benefits may only be attained or become more pronounced after instructors become more familiar with these projects and thus more proficient at facilitating them. Many more significant gains were evident during second and third treatment semesters than were observed in the first treatment. Hence, statistics instructors intending to incorporate similar student-directed projects into their classes should be aware that improvement in their students' knowledge or attitudes about statistics may not be immediate; rather, instructors may need to refine the art of implementing such projects before looking for improved student outcomes.

Finally, it is important to acknowledge that although many results reported here are significant, the effect sizes are small, indicating very modest gains in student outcomes of interest. Nonetheless, the consistency with which observed differences suggest a positive effect is promising. If instructors seize each opportunity to modify their teaching approach to tweak student outcomes favorably, it stands to reason that the cumulative effect of such a practice will be highly beneficial in the long run; a series of small improvements will likely lead to greater impact than might result from any single instructional change. This observation echoes the above call for combining promising approaches in statistics instruction whenever possible.

ACKNOWLEDGEMENTS

This work is supported by NSF grant award DUE-1021584. Any findings or recommendations presented in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors also extend their thanks to the eight pilot instructors who participated in this study, as well as to the members of the advisory panel who provided input on many aspects of the curriculum materials, the instruments, and the study design.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J. Velleman, P., & Witmer, J. (2005) *Guidelines for assessment and instruction in statistics education: College report.* American Statistical Association. [Online: http://www.amstat.org/education/gaise/GaiseCollege_full.pdf]

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics* (Doctoral dissertation). Retrieved from SSRN. [Online: http://dx.doi.org/10.2139/ssrn.2130143]

Bailey, B., Spence, D.J., & Sinn, R. (2013) Implementation of discovery projects in statistics. *Journal of Statistics Education, 21*(3). [Online: www.amstat.org/publications/jse/v21n3/bailey.pdf]

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Scottsdale, AZ: Information Age Publishing.

Bryce, G. R. (2005). Developing tomorrow's statistician. *Journal of Statistics Education, 13*(1). [Online: http://www.amstat.org/publications/jse/v13n1/bryce.html]

Castro Sotos, A.E., Vanhoof, S., Van den Noortgate, W., Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of*

Statistics Education, 17(2). [Online:http://www.amstat.org/publications/jse/v17n2/castrosotos.html]

Cleary, T. J. (2006). The development and validation of the Self-Regulation Strategy Inventory–Self-Report. *Journal of School Psychology, 44*, 307-322.

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

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences,* (2nd ed.). Hilldale, NJ: Lawrence Erlbaum Associates.

DaSilva, M.P.M., & Pinto, S. S. (2014). Teaching statistics through learning projects. *Statistics Education Research Journal, 13*(2), 177-186. [Online: http://iase-web.org/documents/SERJ/SERJ13(2)_daSilva.pdf]

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. [Online: https://www.stat.auckland.ac.nz/~iase/serj/serj6(2)_delMas.pdf]

Finney, S. J. and Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161-186.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education, 2*(2). [Online: http://www.amstat.org/publications/jse/v2n2/gal.html]

Hall, S., & Vance, E. A. (2010). Improving self-efficacy in statistics: Role of self-explanation and feedback. *Journal of Statistics Education*, 18(3). [Online: http:// www.amstat.org/publications/jse/v18n3/hall.pdf]

Hogg, R. V. (1991). Statistical education: Improvements are badly needed. *The American Statistician, 45*, 342-343.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Landrum, R. E., & Smith, R. A. (2007). Creating syllabi for statistics and research methods courses. In D. S. Dunn, R. A. Smith, & B. C. Beins (Eds.), *Best practices for teaching statistics and research methods in the behavioral sciences* (pp. 45-57). Mahwah, NJ: Lawrence Erlbaum.

Mvududu, N. (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education, 11*(3). [Online:http://www.amstat.org/publications/jse/v11n3/mvududu.html]

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology 25*, 241-286.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*(2), 193-203.

Ramirez, C., Schau, C., & Emmioglu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal, 11*(2), 57-71. [Online: http://iase-web.org/documents/SERJ/SERJ11(2)_Ramirez.pdf]

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. Cohen Kadosh & A. Dowker (Eds.), *Oxford handbook of numerical cognition* (pp. 1102-1118). Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/978019942342.013.014

Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in learning and teaching statistics. *Journal of Statistics Education, 16*(1). [Online: http://www.amstat.org/publications/jse/v16n1/roseth.html]

Rossman, A. J. & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*, 211–221. doi: 10.1002/wics.1302

Schau, C. (2000). Survey of attitudes toward statistics. In J. Maltby, C. A. Lewis, & A. Hill (Eds.), *Commissioned reviews on 250 psychological tests* (pp. 898-901). Lampeter, Wales: Edwin Mellen Press.

Schunk, D., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 15-31). San Diego, CA: Academic Press. doi:10.1016/B978-012750053-9/50003-6

Singer, J. D., & Willett, J. B. (1990).) Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician, 44*(3), 223-230. [Online: http://dx.doi.org/10.1080/00031305.1990.10475726]

Spence, D. J., Sharp, J. L., & Sinn, R. (2011). Investigation of factors mediating the effectiveness of authentic projects in the teaching of elementary statistics. *Journal of Mathematical Behavior, 30*, 319-332.

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. [Online: http://iase-web.org/documents/SERJ/SERJ6(2)_Tempelaar.pdf]

Thompson, W. B. (1994). Making data analysis realistic: Incorporating research into statistics courses. *Teaching of Psychology, 21*, 41–43.

Tintle N.L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. [Online: http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf]

Tucker, L.R. (1949). A note on the estimation of test reliability by the Kuder-Richarson formula (20), *Pyschometrika*, *14*(2).

Whitaker, D., Foti, S., Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy, 8*(2), Article 3. [Online: http://scholarcommons.usf.edu/numeracy/vol8/iss2/art3]

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement, 45*, 401 – 405.

Yesilcay, Y. (2000). Research project in statistics: Implications of a case study for the undergraduate statistics curriculum. *Journal of Statistics Education, 8*(2). [Online: http://www.amstat.org/publications/jse/secure/v8n2/yesilcay.cfm]

DIANNA J. SPENCE
University of North Georgia
82 College Circle
Dahlonega, GA 30597

APPENDIX: SAMPLE ITEMS FROM SELF-EFFICACY INSTRUMENT

Sample items from the statistics self-efficacy scale are shown below. One or two items are shown for each of the 4 subscales. Items were scored on a Likert-style scale from 1 to 6, with 6 being the highest rating.

Subscale	Sample Items
Linear Regression	• I am confident that I can use a set of data collected for two variables to determine the equation of a regression line correctly. • I am confident that I can correctly use a linear regression line to make predictions.
Hypothesis Testing (t-Tests)	• I am confident that I can conduct a t-test to compare the means of two populations using data collected from two independent samples. • I am confident that I can correctly interpret the p-value of a statistical test (such as a t-test).
Data Collection	• I am confident that I can devise a sampling strategy that would help to ensure a representative sample.
Learning and Using Statistics	• I am confident that I can learn statistical concepts.