HELPING CRIMINAL JUSTICE STUDENTS LEARN STATISTICS: A QUASI-EXPERIMENTAL EVALUATION OF LEARNING ASSISTANCE*

Shawn D. Bushway
Shawn M. Flower
University of Maryland

Undergraduate statistics courses in criminology and criminal justice programs tend to have high failure rates relative to other courses in the curriculum. One solution to this problem is to focus on helping the students "learn how to learn" difficult material. Drawing upon learning theory and incorporating web-based technology, the authors conducted an evaluation of several learning enhancement tools including Supplemental Instruction (SI), automated quizzes based on the reading, and required class attendance for at-risk students. The results indicate that Supplemental Instruction and quizzes substantially improve the failure rate in the class (by 47% and 52% respectively) but required attendance by itself is not effective.

In the early days of four-year degree programs in criminology and criminal justice, the emphasis was on police practice, and most curricula did not include statistics and/or research methods courses (Tenney 1971). Over time, there has been a realization that practitioners need some basic understanding of how research questions are asked and answered. In 1984, this insight was codified by the Joint Commission on Criminology and Criminal Justice Education and Standards when it recommended that courses in research methods and statistics be a required part of the curriculum (Ward and Webb 1984). A subsequent study of 91 undergraduate four-year programs during the 1987-1988 school year found partial compliance with that recommendation: 30% of the programs required statistics, and 56% required research methods (Southerland 1991). Our own review of the top 20 programs as rated by this journal (Cohn and Farrington 1998)

*This work represents the result of a true collaboration between the two authors. Almost by definition, the help of many people is needed to conduct an evaluation of this nature. For their assistance in the design of the evaluation, we thank Paulette Robinson, Ellen Borkowski, Sharon Rondlely, David Crammigio, Jim Greenberg, Marcia V. Fallon, Ian Kellems, David Iay, Don Berkoweit and Taifa Hillbert. For assisting with the implementation of the study, we thank Leana Allee, Melissa Kellstrom, Jill Farrell, Barbara Scotto, Haisha Thompson and Daniel Twombly. Finally, we thank Charles Wellford for important financial and institutional support. All errors remain our own.

JOURNAL OF CRIMINAL JUSTICE EDUCATION, Vol. 13 No. 1, Spring 2002
© 2002 Academy of Criminal Justice Sciences
found that 14 (70%) required a statistics course for the undergraduate degree, and an additional 15% required a course in research methods. While these programs do not represent a cross-section of criminal justice programs in the United States, it is safe to say that statistics and research methods have become a fixture in the undergraduate curriculum for a large proportion of criminal justice majors in the United States.

A review of the literature on statistics education and our own experience reveal that professors and students often have very different perceptions of these courses. Professors tend to be concerned that the students are simply learning how to “plug and chug” without any sense of when or why they would actually use these tools (Garfield and Ahlgren 1988). Students in these courses, on the other hand, tend to have poor math skills and are plagued by math anxiety. This anxiety is exacerbated by the fact that statistics classes tend to have a high failure rate relative to other social science courses and are sometimes viewed as an obstacle to graduation (Gardner and Hudson 1999).

The desire to improve statistical understanding and the problem of poor math skills creates a conundrum for the criminology or criminal justice professor teaching statistics. Requiring more from the students in terms of understanding is likely to result in even higher failure rates. Yet continuing the traditional course of action in which students can pass without understanding basic statistical concepts risks admitting that, for many students, statistics is a painful exercise that has no real meaning.

One way out of this problem may be for professors to focus not only on successfully communicating the content of the course, but also on teaching students the most effective ways to learn the material (Weinstein and Meyer 1991). This requires a basic understanding of the learning process. According to learning theorists, students develop expertise in a subject by integrating prior knowledge with new concepts. When this happens, new information is either organized within prior knowledge or prior knowledge is reorganized around new concepts. This reorganized knowledge is then placed into long-term memory. However, when this reorganization process does not occur, new information “is recalled in isolated bits (as answers to particular questions in a specific course) . . . and those bits are subject to our forgetting” (Weinstein and Meyer 1991:24). This description is a particularly apt for statistics classes because the material builds cumulatively throughout the semester. A successful student integrates each new piece of information, slowly building statistical understanding that can then be applied to solve problems.

---

1 Many new statistics texts are responding to this concern with an enhanced emphasis on intuition and application. See Smith (1998) for an example of this type of text.

2 Articles on teaching statistics often emphasize new teaching techniques designed to better illustrate key concepts (Smith 1998).
Our reading of learning theory leads us to focus remediation efforts on two stages: the information intake stage and the process of integration. It seems axiomatic that students will not be able to integrate what they do not understand. There are several clear tasks that will aid understanding of the key building blocks of knowledge, such as reading the book before class, attending lectures, and solving practice problems. Ideally, students would perform these tasks on their own. However, it is our experience that many criminal justice students fail to perform these basic learning tasks, leading to inevitable confusion at the integration stage. Repeatedly, we find that the students who struggle in statistics are also the students who fail to do the reading, attend class and/or do the homework. There are any number of reasons for this basic failure, including a fear of math and math texts, as well as a lack of motivation. It is our contention that relatively painless procedures can be implemented to encourage good student learning practices. In this paper, we evaluate two such mechanisms: automated quizzes on the reading material given before the lecture on the material and required class attendance for at-risk students.\footnote{For the purposes of this paper, at-risk students are defined as those students in danger of receiving a D, F, or withdrawing from the course.}

The task of integration is perhaps more difficult to understand and influence. The strategies for learning and integration are sometimes referred to collectively as "study skills." More than the concrete tasks of reading the book or attending class, study skills focus on how students build their understanding of material over the course of the semester. Congos and Schoeps (1998:52) point out that students need to be aware that effective study skills can be learned and that these "skills are not magically and mysteriously endowed on some and not others". Learning theorists suggest that these study skills can be developed best by working together in a group learning environment. In groups, students can work out problems and difficult concepts while "thinking out loud." Consequently, they are more apt to identify incorrect prior knowledge "before misunderstandings become a cohesive part of students' understanding" (Weinstein and Meyer 1991:24). Moreover, students can help each other figure out how to apply concepts to new problems, which is the key learning task that demonstrates that the concepts are understood and "owned" by the student.\footnote{Although these skills can be taught by themselves, it appears to be important to teach these skills in the context of a course where these skills are needed, since studies reveal that "when study skills are taught in isolation from course content, there is no positive impact on student academic performance" (Congos and Schoeps 1998:47).}

There is a well-developed model of group learning in a class context called Supplemental Instruction (SI) (Center for Supplemental Instruction (CSI) 1995). This tutoring program is designed specifically for high-failure courses such as statistics in the criminal justice major. In this paper, we introduce and evaluate the use of this program in the criminal justice environment.
This evaluation of SI is centered on a large lecture-based, required statistics class for criminal justice majors at a large public university. Previous evaluations of SI have been plagued by concerns about selection bias (Congos and Schoeps 1995). Specifically, the concern is that pre-existing differences between program participants and non-participants can explain the differences in student outcomes that are otherwise attributed to the program. This evaluation makes use of a quasi-experimental structure with statistical controls specifically designed to give the reader confidence that any observed difference in student outcomes was the result of the program interventions, and not the result of selection bias. We find that both quizzes and SI appear to substantially improve the failure rate in the class, but required attendance by itself is not effective. Although most professors would prefer to avoid too much hand-holding in the learning process, many students in our criminal justice statistics classes apparently do not have the necessary learning skills to succeed in class. As a result, it might be necessary to make the extra effort to require quizzes and SI attendance in order to teach the students how to learn.

In the next section, we provide more background and discuss each of the interventions in turn. We then discuss the research strategy and present the results. Finally, we offer some conclusions about how to improve student learning.

BACKGROUND

Statistics is a required course for criminology and criminal justice majors at the university. It is preceded by a required course in probability taught by the math department and followed by a course on research methods. A grade of C is required for graduation. The explicit goal of the class is to develop skills and intuition that then can be applied in computer-based projects in the research methods class. The statistics course starts with descriptive statistics, teaches the concepts of sampling and hypothesis testing, and applies those concepts with the binomial, normal, F and chi-square distributions. Historically, the class has a DFW rate (students who receive a D, F or withdraw) of 40%, a subject of some concern for the department.

Efforts to improve performance in this class have occurred over several semesters. In the spring of 2000, we focused on improving student study practices by attempting to eliminate barriers to the learning process. We chose a book written specifically for the class and tied lectures directly to the book. We also required weekly homework assignments designed specifically to prepare students for exams. To provide structure and clarity.

---

5 The class meets for 2 1/2 hours in lecture per week, and has a 50-minute discussion section. Graduate teaching assistants run the discussion sections and offer regular office hours. The class has between 90 and 150 students, depending on the semester. The class assessment is based on four timed exams consisting almost entirely of word problems to be solved with the help of a calculator. The exams are hand graded with partial credit awarded.

6 Bachman and Paternoster (1997).
we also placed the course on the Internet using WebCT, a popular web-based course interface supported by the university. The web site included detailed outlines of the notes that corresponded with overheads used in class, which were made available prior to each class. This allowed the students to avoid the constant struggle of writing down tables and equations during class, and allowed them to concentrate on learning. Answers to the homework and exams were also made available online to encourage the students to review these assessment materials.

An evaluation of this effort revealed that although the overall average improved, the failure rate remained virtually unchanged over previous semesters. After conducting focus groups, it became obvious that the at-risk students did not take advantage of the information provided on the Internet and did not benefit from the homework. In fact, many of the students most in need of help did not attend class, sometimes mistaking the availability of the notes to mean that class attendance was not needed to advance learning. Moreover, students as a whole were failing to read the book and were continuing to miss basic core concepts such as sampling, the Central Limit Theorem, and hypothesis testing. As a result, they could not generalize from one problem to the next, and struggled on any exam question that was not virtually identical to a homework or practice problem.

In an effort to improve student learning among the at-risk students and to decrease the failure rate, we identified three goals that corresponded with the key learning skills of information intake and information integration. The first goal was to encourage class attendance. The second was to encourage students to read the book through the use of online quizzes, and the third was to provide a mechanism for group learning that focused specifically on developing study skills, (Supplemental Instruction).

**Required Attendance**

Other than the basic sentiment that student attendance is positively correlated with student performance, we found no studies of class attendance (Devdoss and Foltz 1996; Lai and Chan 2000). Since mandatory class attendance for all students would be viewed as punitive, we took an alternative approach that could be viewed as a welcome offer of assistance in a difficult class. Specifically, we created an "I Need a C" (INAC) club whereby any student who was identified as at-risk was required to attend lecture and discussion section as part of their homework grade. Students could also volunteer to participate in the INAC club. Although there was no penalty if an individual who was not in the club did not attend class, all students were required to sign an attendance sheet at the beginning of each class. For those students in the INAC club, 25% of the weekly homework assignment was based on class attendance. In addition, these students had access to two points of extra-credit problems each week that were added to their homework grades. At-risk students were identified four times during
the semester, with the initial assessment based on their performance on a pre-test given on the second day of class and their performance in the prerequisite class. At-risk status was then reevaluated after the first, second and third exams. Students who volunteered in the club followed this time schedule, and, once they volunteered, they were committed to the rules of the club until the next reassessment. Over half of the class volunteered to be a part of the club. Qualitative evaluations revealed that the extra-credit points and the incentive to attend class motivated students to volunteer. In addition, WebCT allowed us to identify members of the INAC club discretely by allowing only those individuals identified as at-risk, along with those who volunteered, to read a bulletin board with information including the extra-credit assignments and announcements about this component of the course.

Quizzes

Perhaps because web-based learning is still a relatively new classroom enhancement, there is little in the existing literature that addresses automated quizzes or their possible impact on student learning.\(^7\) The one study we found used quizzes to prepare chemistry students for laboratories in lieu of a traditional pre-laboratory preparation (McKelvy 2001a, 2001b). A comparison of the grades revealed that overall, the quiz group fared worse than the traditional group. This study concluded that using WebCT in conjunction with (not instead of) traditional methods (e.g., combining the instructional video over WebCT with quizzes and the utilization of the laboratory manual) provides an environment that maximizes the student's learning experience. This fits very well with our intended goal of using quizzes to encourage traditional learning through the reading of the text. We chose to create one quiz for every book chapter. The questions were multiple-choice sentence completion exercises, with the sentences taken verbatim from the text. The quizzes themselves were available 48 hours prior to the lecture in which the material would be discussed for the first time. Each quiz included five multiple-choice questions and students were allowed to retake the quiz if they were not satisfied with their initial score. Each quiz had a bank of questions that rotated, so that within a topic, the quiz questions varied when the student retook the quiz. If the student got a question wrong, they were not told the correct answer, but were referenced to the section in the text where they could read the relevant material. It was expected that students would use their book while taking the quiz. The quizzes could be taken at any computer with an Internet connection, but students needed to use their student identification to gain access to the website. The quizzes were counted as 10% of the overall grade in the course.

\(^7\) However, we have noted a marked increase in the number of textbook manufacturers that now offer online quizzes on the book as part of their textbook.
Tutoring/Supplemental Instruction

After our first semester of implementing WebCT in the classroom, we found that the students who traditionally pass the course utilized the extra resources and thus improved their overall final grades. However, upon review of student performance and attendance in the classroom and during the professor and teaching assistant’s office hours, it became evident that at-risk students were not keeping up with the course, regardless of the extra resources of a web-enhanced classroom environment. This result is corroborated by a review of the literature, that revealed that at-risk students are the least likely group to take advantage of supplemental educational resources (Blanc, DeBuhr and Martin 1983; Hodges 2001; Ramirez 1997; Visor, Johnson and Cole 1992). This finding fits well with our understanding of study skills. Many students who fail these high-risk courses do so because they demonstrate poor learning practices, which include not taking advantage of resources provided by the instructor. Supplemental Instruction (SI) bills itself as a program that focuses specifically on improving learning skills in high risk classes, and therefore seemed particularly appropriate for this classroom situation.

Deanna Martin developed SI as a peer-assisted study session program in 1973. The SI program differs from traditional tutoring in several important ways. First, as SI is open to all students enrolled in the classroom, it is not considered remedial education, nor is it focused on at-risk students (Burmeister 1996: University of Missouri 1996). Conversely, traditional tutoring is usually directed toward at-risk students. Secondly, SI is course, not topic, specific. The SI leaders who facilitate the peer learning process are students who have successfully completed the course in a previous semester. In their role as leaders, they attend class, complete the homework assignments and take the exams side by side with the students they tutor. As a result, there is little confusion, or misinformation, about the course materials (Blanc et al. 1983). In a traditional tutoring environment, the tutors may be familiar with the overall topic (e.g., math, science) but may not necessarily be familiar with the specific course content.

Third, SI trains leaders to develop group learning techniques. During the study sessions, the leader serves as a facilitator in the learning process rather than an instructor, and they emphasize “what to learn and how to learn” (UM 1996). SI is considered to be a “marriage of content and skills” (Burmeister 1996:24). SI leaders are explicitly cautioned not to answer questions but are instructed to model the learning process through and within the group environment using collaborative learning techniques, including group discussion. The basis of this collaborative process is formed by the ideas of Piaget, who developed the “comprehensive model of cognitive development,” which was then adapted by his colleagues to an educational theory called “constructivism” (Martin and Arendale 1993:43).
Constructivism takes its name from the idea that "students must "construct" their own knowledge in order to be able to understand and use it" and that knowledge is gained through "expertise" (Martin and Arendale 1993: 43). As many college students have not yet developed the ability to reason abstractly, they struggle to learn new ideas through traditional teaching methods of lectures and reading. SI peer leaders can help students to gain a higher level of expertise by encouraging them to reason through problems rather than "telling" them the answer. In addition, as less experienced learners often fail to recognize what they do not know. SI leaders, through the collaborative learning process, attempt to help students to discover the concepts and ideas they do not understand (Congos and Schoeps 1998). Once a student knows what they do not know, SI leaders try to help them to "frame questions that will lead to a more complete understanding" (Martin and Arendale 1993:48). In addition, as cooperative learning focuses on learning and not on evaluation, SI seeks to foster improved "interpersonal relations, which further leads to increased self-esteem" (Lundeberg 1990:153) which is linked to improved student performance. Students feel more comfortable asking questions among peers and find that the realization that others struggle with the material "is relaxing - which aids learning" (Lundeberg 1990:150). Finally, SI provides instruction on learning techniques to assist with problem solving, exam preparation, effective note taking and other strategies to enhance student performance (Blanc et al. 1983).8

National evaluations of the SI program conclude that participants who attend SI average from one half to one full letter grade above the average of those who do not attend SI (CSI 1998: UM 1996). The evidence also suggests that courses that implement SI have a lower drop out rate, as well as fewer Ds and Fs (Burmeister 1996; CSI 1998; Kochenour et al. 1997; Lundeberg 1990). Overall, evaluations have shown that "across institutional types, disciplines, precollege student preparation levels and ethnic groups" those who participate in SI perform at a higher level than those who do not participate in SI (Lundeberg 1990: Ramírez 1997:3).

The problem with this type of study is what Congos and Schoeps termed "pesky self-selection bias" (1999:73). Since the comparison is between people who chose to attend SI and those who chose not to attend, it is possible that the observed difference in outcomes is the result of pre-existing differences in ability and motivation between those who participate and those who do not. There have been several attempts to control for

---

8These features of SI denote SI as fundamentally different than traditional individual tutoring. In fact, traditional individual tutoring is viewed by some learning scholars as the least effective means of improving long-term student learning for at least three reasons (Keimig 1985): First, as students "drop in" rather than attend tutoring regularly, tutoring fails to be a "systematic activity," second, by the time students seek tutoring assistance, it is often too late to make a difference. Third, the students most in need of tutoring are the least likely to participate. A 1990 study by Maxwell found that "there is no evidence that tutoring helps the weakest students" (1990:4).
these differences, including one study that compared participants in SI with a motivated control group of students who said they wanted to attend SI but could not due to schedule conflicts, along with a third group of nonparticipating without schedule conflicts (Blanc et al. 1983). In keeping with the hypothesis of selection bias, the motivated non-attenders performed better than the unmotivated non-attenders. But, in a result that supports the efficacy of SI, the attenders still did better than the motivated non-attenders.

Peled and Kim (1996:12) attempted to circumvent the bias related to “comparing self-selected groups” by evaluating classes of students that were offered SI to classes that did not, within the assumption that the students were equivalent groups. Attendance at SI was voluntary, so not all students attended SI, and attendance varied for those that did attend. The authors also assumed “that the amount of participation in sessions increased the success of the class as a whole” (1996:6). This study found that classes with SI improved their final grades by 6.5 percentage points, reduced the number of grades below 60% and increased the number of students who received a grade above 80% (Peled and Kim 1996).

In both cases, of course, it is possible that there are still uncontrolled differences between SI participants and non-SI participants. This difference in the students, and not SI, might have explained the differences in performance/participation. In the same way, the classes that got SI, or the students that attended these classes, might well have been different than the classes that did not get SI. In light of this selection bias problem, it is important that the researcher controls for differences in GPA, math ability and other observed differences using statistical methods (Congos and Schoeps 1997, 1999). Congos and Schoeps (1997) appear to show that there was still a modest improvement in performance when these differences are controlled for statistically.

From our perspective, selection bias is not just a threat to internal validity, but also an indicator of the problems in this type of remediation. As mentioned earlier, it is a well accepted fact in the learning literature that the students most in need of help often do not attend the extra sessions, precisely because they are not good learners. As a result, SI is under-utilized. Nationwide, SI is focused on high-risk classes with D, F, & W rates between 30% to 40%. Yet nationwide, attendance at SI sessions is around 8%-10%. (Fallon 2001). Given the fact that included in this 10% are students who are going to pass anyway, it seems fairly clear that the majority of students who would benefit from SI do not attend. In effect, while SI is most helpful for the students with real learning needs, this student is the very student least likely to voluntarily take advantage of extra resources.

This conundrum was first identified in a SI program conducted at California State University, Long Beach. They found that the highly motivated individuals did fairly well with or without SI, while the low-motivation
group was "unable to sustain a mean grade of 'C'... without the support of SI" (Ramirez 1997:6).

One response to this problem is to think about making SI mandatory for at-risk students, despite the fact that the program's creator originally intended for the program to be voluntary to avoid stigmatizing at-risk students. At least one researcher has begun to evaluate the effect of making SI mandatory. In one study of a community college population in the early 1990s, Hodges found that students "required to attend SI earned significantly higher final course grades" (2001:3). Another program used extra credit as an incentive to participate, but students did not respond affirmatively to the inducement, as less than 10% attended SI (Hodges 2001). A more recent study by Hodges et al., found that there were no differences in final course grades between those students who were mandated to participate in SI and those who participated voluntarily, apparently showing that forcing students to attend does not appear to yield negative outcomes (2001). Hodges concludes by stating that "students may need to be required, not simply encouraged, to participate" in programs such as SI (2001:8). In response to the problem of non-attendance by the students most in need, and the early positive results of evaluations of forced attendance at SI sessions, we decided to require attendance at SI sessions for members of the INAC club. In response to concerns about stigmatization, we included components in the evaluation that would allow us to detect feelings of isolation/exclusion.

In implementing SI we asked two students who had previously received As in the class to be Supplemental Instruction Leaders. They offered four sessions over the course of the week, at times chosen by the students in the class. The leaders were given credit for a 3-credit teaching experience class. They were required to attend class and to take all exams. We chose students who we thought would relate well to at-risk students. One student was an African American female and the second was a white male. After an initial training day on the SI approach provided by the Learning Assistance Service, which administers the university-wide Supplemental Instruction program, the students met with the professor weekly to cover any issues that arose. Both students were evaluated by the on-campus SI staff twice during the semester. The SI coordinator felt that both students related well to their peers and did an excellent job of staying true to the SI model. They also received positive feedback from the students who were in the SI program. Overall, our leaders performed very well, with no missed sessions and a responsible attitude towards implementing SI.

**EVALUATION STRUCTURE**

When creating the evaluation, we were particularly interested in establishing the internal validity of our two main substantive contributions to the
literature: testing the usefulness of automated quizzes on reading material and evaluating a mandatory program of SI. In a happy coincidence, making the SI sessions mandatory for a large minority of the students also made it possible to deal with selection bias. Using our class of 92 students in the fall of 2000, we created a quasi-experiment, in which we gave quizzes to one section of 31 students and gave another section of 30 students access to SI. All students were subject to the conditions of the INAC club (required attendance). By focusing on the students who had access to SI rather than only on students who actually attended SI, we should be able to claim that any observed differences between the sections are the result of SI and not selection bias. The key advantage of this study versus other “intent to treat” studies such as Peled and Kim (1996) is that all of the students in the study had the same class with the same professor and teaching assistant. However, this is called a quasi-experiment because we were not able to randomly assign students to sections. As a result, we also controlled for observable pre-existing differences between students. We collected information from student records (GPA, race, gender, age, and performance on the prerequisite) and gave all of the students a pretest of basic math skills.

In this evaluation structure, the baseline treatment consisted of access to WebCT (online lecture notes and other course information), and membership in the “I Need a C” (INAC) club, which required lecture and discussion attendance. Recall that membership was mandatory for students deemed at-risk and voluntary for all other students. Use of computerized membership lists, allowing all students to volunteer, and requiring all students to sign in for class, appeared to minimize any stigma that at-risk students may have felt about being required to attend class. We conducted four qualitative evaluation surveys and specifically asked students how they felt about being a member of the INAC club, including whether they felt stigmatized. Only one student responded that he felt stigmatized, but he had volunteered for the club. Most students responded positively to INAC because it motivated them to attend class and they liked to earn the extra-credit points. From our experience, stigmatization was not an issue with our students. INAC club and WebCT was offered to all students and was

---

9. We received approval from our Internal Review Board. Although technically experiments involving the evaluation of course improvements are exempt from IRB review, we felt that going through the IRB process enhanced the research evaluation and gave us some credibility with the students. All students signed an informed consent form advising them of their rights with respect to the use of the information collected for the evaluation. Students did not have the right to avoid participation, although they were able to drop the class — one student did drop the class because of the experiment. Most students came to accept the experiment after some initial doubt about its merits/fairness.

10. *A priori*, it is not clear that the process is not random (especially since students did not know about the experiment ahead of time). However, it is possible that some sections attract better students because of the timing of the discussion sections. In the present case, none of the sections were at a particularly bad time (e.g., 8 am Monday morning or 3 pm on a Friday afternoon) and we will show that the sections appear to be very similar.
the only treatment offered to the first section (we call this Group B in the following discussion).

A second section (Group C) included SI. Four SI sessions were offered per week throughout the course at times that maximized attendance. Students in INAC were required to attend these sessions, while others could attend if desired. The actual penalty for failing to attend was very minor (a 10% reduction on the homework). But because of the additional time requirement, only 50% of Group C belonged to INAC, as opposed to 70% in the other sections. Nine (approximately 30%) of the students attended SI consistently, seven of whom were at-risk. Although we were pleased to get 30% of the students regularly attending SI, we were concerned that requiring SI may have led some at-risk students to drop or avoid class. But the drop rate for the group with SI was lower than the drop rate for Group B without SI. The students who failed to attend SI also failed to attend class, suggesting that the problem was not SI but the class in general.

All members of the third section (Group D) were required to take quizzes on WebCT. The only additional change was a minor modification in how the grades were distributed. Members of this section earned 10% from quizzes, 10% from homework, and the remainder from exams. All other students earned 15% on homework and the remainder from exams. In this evaluation, we will focus on the average of the exams only.

Since all members of the class had access to WebCT, and those who were members of INAC were required to attend lectures and discussion section, it was not possible within this class to evaluate the impact of the required attendance. As a partial solution to this problem, we identified a second section of the same class taught by another professor. This professor had similar experience teaching this class (three previous semesters) and agreed to teach the class using standard techniques, with no required attendance, and without the use of WebCT. Moreover, the professor agreed to use the same tests, homework and discussion questions. To allow for the fact that the students may differ across classes, we also collected information about the student’s prior records in this class. We refer to this class as Group A.

With four groups, we can now identify the basic structure of the evaluation, described in Figure 1. The basic impact of forced attendance and WebCT can be seen as the difference between Groups A and B. The impact of quizzes over and above WebCT and mandatory attendance is the difference between D and B. Finally, the impact of SI net of WebCT and coerced attendance is the difference between Groups C and B. In the next section, we discuss the results of the analysis.
**Figure 1. Basic Structure of the Evaluation**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Traditional Class</td>
<td>N=86</td>
</tr>
<tr>
<td>B</td>
<td>A + WebCT + INAC</td>
<td>N=31</td>
</tr>
<tr>
<td>C</td>
<td>B + Tutoring</td>
<td>N=30</td>
</tr>
<tr>
<td>D</td>
<td>B + Quizzes</td>
<td>N=31</td>
</tr>
</tbody>
</table>

- Professors X and Y had both taught the class on three previous semesters.
- Students in all four groups received the same exams, homework assignments and discussion questions.
- Students in Group A were taught by Professor X using standard techniques with no required attendance and without WebCT and served as the evaluation control group.
- Students in Groups B, C, and D, the treatment groups, were taught by Professor Y, and they received WebCT and were either mandated or could volunteer to participate in the INAC club.
- Students in Group C had access to Supplemental Instruction.
- Students in Group D were required to complete WebCT online quizzes throughout the semester.

**RESULTS**

We calculated the final grade for this evaluation based on an equal weighting of the three exams and the final. We chose to use a cutoff for passing of 70 or better. This decision rule results in a higher failure rate for this evaluation than what the students actually experienced, since the homework assignments were worth 15% of the final grade, and most students did better on the homework than on the exams.\(^1\) However, the homework was only partially graded, and precautions to avoid grader bias

\(^1\) In addition, because the failure rate for Group A was so high, the professor curved the final grade so that she had a 40% DFW rate matching previous semesters.
Figure 2. Failure (DFW*) Rates by Group, With and Without Statistical Controls for Observable Differences Between the Groups

* The DFW rate stands for the proportion of the class that earns a D, F or withdraws from the class.

** The controls included variables measuring gender, race, and year in school, and traditional student status (e.g., age 23 or under). Additional control variables included transfer status, GPA, performance on the class prerequisite and past class performance.

*** Groups C and D are significantly different than Group A in the baseline model, and significantly different than Groups A and B in the model with controls. Statistical significance means that we can be reasonably certain, with some chance of error (in this case 10%), that the observed difference is an accurate reflection of reality.

were not as strong as with the exams. In the exams, each grader (two teaching assistants and two professors) evaluated all students on a given question, using criteria agreed on by the two professors in order to insure consistency across groups. Moreover, the SI group students did better on the homework because they worked together on the homework. Including the homework grade as part of the evaluation risks biasing the results in favor of finding a result supporting SI merely because SI students had assistance with their homework. Finally, all of the learning techniques in this class, including attendance, quizzes and homework, were intended to improve learning, as measured by exam performance. Therefore, the exam average provides the best gauge for the success of the class.

We chose to focus on the DFW rate, (the percent of the class that would receive a D, F or W– withdraw) rather than the exam average for two reasons. First, we are primarily concerned with improving the DFW rate. Second, focusing on the average rather than the DFW rate forces us to ignore data from students who dropped the course, an important subset of students in this evaluation.
Figure 2 shows the baseline DFW rates for the four groups. The DFW rates are 34.7%, 41.9%, 33.3% and 25.8% for Groups A-D respectively. Only the use of SI and automated quizzes showed a significant impact over the basic course offering reflected by Group A. The impact is substantively quite large. The SI group (Group C) had a DFW rate 39% lower than the DFW rate of the control group (Group A), and the quiz group (Group D) had a DFW rate 53% lower than the DFW rate of Group A. In contrast, it appeared that WebCT and required attendance by itself (Group B) did not significantly improve the DFW rate over that of the control group.

Of course, as people were not randomly assigned to the groups, other differences might exist between these groups, other than the treatment, which could explain these results. Table 1 provides data on how each of the groups varies on important characteristics. The variables that we included in our model to control for student differences included both demographic and academic indicators. We controlled for gender, race, year in school, and based on their date of birth, we categorized students as traditional (e.g., age 23 or under) or nontraditional students (older than age 23). In addition, we looked to see if the student had transferred into our school from another college or university. Academic indicators included GPA, whether or not the student had completed and/or passed the prerequisite course, and if the student was a "repeater." A "repeater" is someone who had attempted the course in a prior semester and had either withdrawn or failed the course.

The most obvious finding is that Group A's students were. a priori, much more at risk for failing the class than the students in the other three groups. In contrast, with the exception of the percent of seniors, the three groups with WebCT, plus enhancements (Groups B-D) were remarkably similar. This similarity is consistent with the idea that the students randomly sorted themselves into sections within a class during the registration process, although the students clearly did not randomly sort themselves into classes by professor.

Because of the differences observed between the groups, it is important to try and control for these differences statistically. Table 2 provides the results of a probit regression predicting the probability of passing the class. Column 1 provides the results for the baseline regression with no controls, and Column 2 provides the results for the regression with controls for the variables in Table 1. The predicted probabilities for each group, all other factors held constant at the mean, are included in Figure 2. Substantively, the difference between A and B decreases, while D (quizzes) looks even better than before. Moreover, Group D is now significantly better than Group B at the 10% level, allowing us to conclude that quizzes improve the DFW rate by 52% over and above the WebCT and required
Table 1. Baseline Characteristics by Group

<table>
<thead>
<tr>
<th>Variable</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>49</td>
<td>45</td>
<td>37</td>
<td>45</td>
</tr>
<tr>
<td>White*</td>
<td>50</td>
<td>77</td>
<td>67</td>
<td>65</td>
</tr>
<tr>
<td>Senior Year*</td>
<td>21</td>
<td>13</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>Nontraditional Students (24 Years Old &amp; Up)</td>
<td>13</td>
<td>10</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Students Who Did Not Pass the Math Prerequisite*</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Students Who Have Failed or Dropped This Class Before*</td>
<td>17</td>
<td>7</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Transfer Students</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Score on Math Pretest**</td>
<td></td>
<td>85</td>
<td>83</td>
<td>84</td>
</tr>
<tr>
<td>GPA</td>
<td>2.6</td>
<td>2.7</td>
<td>2.6</td>
<td>2.6</td>
</tr>
</tbody>
</table>

* Significantly different at the 10% level. Statistical significance means that we can be reasonably certain, with some chance of error (in this case 10%), that the observed difference is an accurate reflection of reality.

** The students in Group A were not told about the pretest, while students in Groups B-D were told and were encouraged to study. Perhaps as a result, the performance on the pretest was significantly worse in Group A. In what follows, the pretest grade for Group A was curved to equal the average of the other three groups, meaning the pretest variable can only explain variation with the groups, not between them. Also, 8 students in Group A either did not take the pretest or got a zero. All 8 dropped the class eventually. This helps to explain, but does not completely explain, the correlation between the pretest and the failure rate.

Attendance. This comparison between B and D is stronger than the comparison between A and D because Groups B and D had similar students, and each group experienced the same lecture setting. In contrast, Group A had a different professor and a more at-risk population.

The impact of SI (Group C) on the DFW rate is large (47%) but not statistically different than that of Group B using a two-tailed test. It is significant at the 10% level using a one-tailed test. Although this result is not as strong as for the quiz group, the use of a one-tailed test in this case is justified given the strong theoretical predictions of SI. Based on Table 2, we conclude that quizzes, and to a lesser extent, SI, significantly improve the DFW rate over both a traditional course and a course with WebCT and required attendance. We also feel fairly certain that our results are not the outcome of selection bias, at least not selection on variables that we included in the model.

Among the control variables, race and prior academic performance (as measured by performance on the prerequisite, the pretest and GPA) are all significant and substantial predictors of passing. For example, having a 3.0
Table 2. Probit Analysis of Treatment Effect on Failure (DFW*) Rate

<table>
<thead>
<tr>
<th>Variable</th>
<th>Without Controls</th>
<th>With Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A (Comparison Group)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Group B</td>
<td>-0.32</td>
<td>-0.048</td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>Group C</td>
<td>-0.55**</td>
<td>-0.68*</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.39)</td>
</tr>
<tr>
<td>Group D</td>
<td>-0.77***</td>
<td>-0.76**</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>Women</td>
<td>-</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.27)</td>
</tr>
<tr>
<td>White</td>
<td>-</td>
<td>-0.46*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.27)</td>
</tr>
<tr>
<td>Asian</td>
<td>-</td>
<td>-0.089</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.50)</td>
</tr>
<tr>
<td>Senior Year</td>
<td>-</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.39)</td>
</tr>
<tr>
<td>Junior Year</td>
<td>-</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.284)</td>
</tr>
<tr>
<td>Non-Traditional Student</td>
<td>-</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.43)</td>
</tr>
<tr>
<td>GPA</td>
<td>-</td>
<td>-1.06***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.30)</td>
</tr>
<tr>
<td>D on Math Prerequisite</td>
<td>-</td>
<td>0.55*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.29)</td>
</tr>
<tr>
<td>Did Not Take Math Prerequisite</td>
<td>-</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.45)</td>
</tr>
<tr>
<td>Failed or Dropped Class Before</td>
<td>-</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.39)</td>
</tr>
<tr>
<td>Transfer Student</td>
<td>-</td>
<td>-0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.501)</td>
</tr>
<tr>
<td>Performance on Pretest</td>
<td>-</td>
<td>-0.19***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.046)</td>
</tr>
<tr>
<td>Pseudo R²**</td>
<td>0.04</td>
<td>0.40</td>
</tr>
</tbody>
</table>

The Probit coefficients represent movement in Z scores, which can than be mapped to changes in the probability of failure. For convenience, assume that the original Z score is 0, representing a DFW rate of 50% for Group A.

**** The Pseudo R² obtained from the Probit maximum likelihood technique is a reasonable approximation of the more conventional R² derived from Ordinary Least Squares (OLS) analysis.

* Statistically significant at the 10% level
** Statistically significant at the 5% level
*** Statistically significant at the 1% level

GPA instead of the average 2.6 GPA would decrease the probability of failing by 35%, assuming that the average chance of failure is 40%. A close look at the gain in R² from Column 1 to Column 2 demonstrates that prior
ability explains far more of the overall variation in class performance than does learning assistance — yet learning assistance can and does improve student learning.

One concern that we had in reviewing these results was that the students who dropped the course drove these results. This is clearly a possibility, since 22% of Group A, 19% of Group B, 16.6% of Group C, and 13% of Group D dropped by the end of the semester.\footnote{These include both formal drops and informal drops. Informal drops are those people who stopped coming to class and taking exams. Although they are officially counted as F's rather than withdrawals by the registrar’s office, their final grades are skewed by a score of zero on the final.} So we also examined the distribution of numerical final grades by group. We found that, of those who took the final, 42% of Group A, 28% of Group B, 20% of Group C, and 15% of Group D averaged below a 70 on all four exams. Based on these results, we conclude that quizzes and SI improve the pass rates, both by decreasing the number of students who drop the class and by increasing the performance of the students who did not drop the class.

The average final grades for the non-droppers also varied in the expected direction: \((A=70.9, B=77.2, C=78.9, D=79.5)\) although only the differences between Group A and all other groups is statistically significant. The fact that the average for Group B is similar to the average for Group C and D, despite the higher DFW rate, confirms the earlier finding that supplemental information available on WebCT does not help the at-risk student but does help the competent learner perform better in class.

We also surveyed the students multiple times during the course of the semester to get feedback on their experience. WebCT itself was generally well received, with the exception of technical problems involved with downloading the lecture notes. Most students liked that information was readily available when they wanted it. The feedback on required attendance, implemented through the INAC club, was almost uniformly positive. 83% of the respondents said that it encouraged them to come to class, and many students seemed to appreciate the encouragement. For example, one student said “It makes me come to class and try when otherwise, I probably would have become frustrated and given up a long time ago.” Another student highlighted the fact that the creation of the INAC club was actually excellent public relations in a class long known for its somewhat tense atmosphere. The student said, “I feel that it gives me every opportunity to do well in this class. All I have to do is put forth the effort, and I have the chance to do really well.”

The reaction to the SI sessions was also extremely positive. Only one of the students reported feelings of stigma, while all other students who attended the SI sessions reported that they valued the experience. Most of the comments supported the claims made by the SI literature that SI helps aid the learning process. One student said “I found it extremely helpful,
not because we were given answers, but because we found our own answers and then got immediate feedback. If they were wrong, we could fix them. It really taught me how to do things right. I could not have done well without tutoring.” Other students commented on the peer context, saying, “I liked talking with other students. Sometimes it’s easier than speaking with a busy professor.”

The quizzes, which showed the largest and most unambiguous improvement in outcomes, were less well received by the students. On the one hand, 80% of the students felt that they were easy to take, and 75% felt that they were clear. And, consistent with the improved outcomes, these students do report more time with the book than other students (about 30 minutes more per week). On the other hand, less than half of the students liked the quizzes and only about half thought it helped them understand the lectures. There were complaints that the book and the lecture did not correspond and some students did not like the fact that the quizzes were very different than the exams. Many students admitted to just scanning the text while taking the quiz, looking for the right definition. However, even just becoming familiar with key terms might increase the chance that they understand the class material. Recall that the quizzes were not intended to be difficult, nor were they intended to provide direct practice for the exam. Rather, they were intended to familiarize the students with the material before it was taught in class and to keep the students up with the reading in order to increase understanding of the core concepts.

Despite the students’ somewhat negative attitude towards the quizzes, there was some additional evidence that quizzes did, in fact, increase learning, suggesting that the observed benefits of quizzes are valid. For example, the correlation coefficient between quiz performance and the final grade was .55 for those who did not drop, and the average quiz grade for the four students who scored below a 70 was a full 10% lower than the average quiz grade for those who scored above a 70. Moreover, there were a number of students who had positive things to say about the quizzes. For example, one student said that it was “helpful to take the quiz before class time, because once I got to class and listened to lecture I really understood what the quiz and readings were about.”

**CONCLUSION**

In this paper, we reported the results of an evaluation of three different interventions to increase learning: required attendance, Supplemental Instruction, and automated quizzes on the reading. Each of these improvements was evaluated in the context of a web-based course environment supported by WebCT. The quasi-experimental nature of the experiment, with statistical controls for pre-existing differences between the groups, provides strong controls for selection bias. This evaluation found support for the claim that Supplemental Instruction does improve at-risk student
performance, even if participation is mandatory rather than voluntary. This
represents an important contribution to the larger literature on Supplemental
Instruction, in addition to the present context of statistics education
in the field of criminal justice and criminology. The only caveat for professors
is the amount of effort required to administer this type of program.
This effort is mitigated somewhat if SI is supported by a university learning
assistance or counseling center.

The evaluation also found strong support for automated quizzes.
Quizzes decreased the DFW rate by over 50% over WebCT, a rather
dramatic result. They were also met with some resistance by the students.
Unlike the SI sessions, the quizzes were a relatively impersonal and persist-
ent infringement on student time. The fact that the quizzes were adminis-
tered in an experimental setting where other students did not take the
quizzes may have added to the resentment. In addition, since the professor
never met with the quiz group by themselves, he could not provide feed-
back or encouragement about the quiz process. As a result, many students
might not have understood the rationale behind the quizzes. If quizzes
were integrated into the course for all students, they could be discussed in
class, and the students could more directly see the benefits of reading
before class.

Despite the potential public relations problem, quizzes were unique in
their ability to help every student. SI only helped those students who at-
tended (who tended to be at-risk), and WebCT and INAC apparently only
helped those students not at risk for failing the class. In contrast, the quiz-
zes appeared to improve the grades for all members of the group, both
those who were at risk, and those who were not at risk. Furthermore, from
the professor's perspective, automatic quizzes are easier to administer than
SI, and many textbooks have begun to create online quiz packages that can
be accessed by the professor.

Of course, these results are not perfect and should be interpreted with
due caution. Many of the strongest results come from one class, with three
sections and 92 students, in one university. Although we feel that the con-
trols for selection bias were excellent, it is ultimately possible that the re-
sults were driven by selection bias since we did not conduct a true
experiment. Moreover, the results may not generalize to other settings,
and it is not clear if forced attendance is required to make SI and quizzes
effective. Clearly, other experiments of this type are needed to build on
the results in this paper.

In that vein, we would like to suggest that other faculty consider per-
forming similar evaluations in their own large classes. Although we met
with some student resistance, treating different sections in the same class
differently is relatively easy to accomplish. It also creates a fairly strong
research design, provided that some controls for pre-existing differences
are collected. In our experience, the cost of evaluating the course improvements that most professors may undertake as a matter of standard operating procedure is small, relative to the insight gained about what works to improve student outcomes.

REFERENCES


Center for Supplemental Instruction. 1998. Supplemental Instruction (SI) Review of Research Concerning the Effectiveness of SI from the University of Missouri-Kansas City and Other Institutions from Across the United States. University of Missouri-Kansas City.


---

13 By far the most time-consuming part of the evaluation was the coordination with the other professor (Group A); despite that professor's enthusiastic cooperation. In contrast, the evaluation within the one class (Groups B-D) was straightforward and easy to implement, simply because many of the potential confounding factors (e.g., differential treatment by professors and TAs) were held constant by definition.