The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Does Supplemental Instruction Improve Grades and Retention? A Propensity Score Analysis Approach

Nicholas A. Bowman, Stephanie Preschel & Danielle Martinez

To cite this article: Nicholas A. Bowman, Stephanie Preschel & Danielle Martinez (2021): Does Supplemental Instruction Improve Grades and Retention? A Propensity Score Analysis Approach, The Journal of Experimental Education, DOI: 10.1080/00220973.2021.1891010

To link to this article: https://doi.org/10.1080/00220973.2021.1891010

Published online: 06 Mar 2021.

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2021.1891010

Routledge
Taylor & Francis Group

LEARNING, INSTRUCTION, AND COGNITION

Does Supplemental Instruction Improve Grades and Retention? A Propensity Score Analysis Approach

Nicholas A. Bowman ,Stephanie Preschel, and Danielle Martinez

University of Iowa

ABSTRACT

Many colleges and universities seek to promote student success through targeted strategies for individuals or groups of students who are believed to have a higher risk of attrition. Taking a different focused approach, Supplemental Instruction (SI) provides voluntary collaborative learning sessions that are generally linked to specific undergraduate courses with a high percentage of students who either receive low grades or do not complete the course. Although a substantial body of literature has examined the outcomes associated with SI, many of these studies have notable methodological limitations, which include problems with student self-selection into SI participation. The present study examined the effects of SI using doubly robust propensity score analyses with a total of 12,641 observations from 21 different courses across 2 semesters. In both semester samples, SI participation led to higher course grades and retention. The strongest relationships were often observed for underrepresented racial minority students and for students who attended at least five SI sessions. The results did not differ systematically by students' sex, first-generation status, high school grades, and precollege standardized test scores. The findings have important implications for the use of SI to help students overcome challenges within early college coursework.

KEYWORDS

Academic achievement; peer assisted learning; peer assisted study sessions; retention; Supplemental Instruction

Introduction

COLLEGES AND UNIVERSITIES have a substantial interest in promoting college students' adjustment, academic achievement, and ultimately degree attainment. These efforts take the form of not only implementing programs and practices to help all students, but also identifying students who are perceived to be "at risk" of attrition and providing them with targeted assistance and resources (e.g., Hossler & Bontrager, 2014). At the individual level, these resources may include mental health counseling, disability services, financial aid, and academic advising, whereas group-level interventions may include summer bridge programs, first-year success seminars, and learning communities. Although such approaches have the potential to promote student persistence and graduation (e.g., Douglas & Attewell, 2014; What Works Clearinghouse, 2016a), they also run the risk of conveying a deficit message to students that they need additional help, which can potentially backfire and thereby undermine their intended effect (Yeager & Walton, 2011).

Supplemental Instruction (SI) constitutes a somewhat different approach that focuses on improving student success within high-risk courses, which are often defined as those that have sizable percentages of students who receive low grades or do not pass the course (see Martin & Arendale, 1992). SI is a widely used practice that originated at the University of Missouri-Kansas

CONTACT Nicholas A. Bowman nick-bowman@uiowa.edu University of Iowa, N491 Lindquist Center, Iowa City, IA 52242, USA.

City (UMKC) in 1973. Just 35 years after the creation of SI, UMKC's International Center for Supplemental Instruction had trained practitioners and administrators from more than 1,500 colleges and universities in 29 countries (Wilcox & Jacobs, 2008). Although a substantial body of literature has investigated the link between SI participation and college student outcomes (see Arendale, 2020; Dawson et al., 2014), this research generally does not account for student self-selection into SI, it is often limited to outcomes that occur during the semester of that course, and it only occasionally considers whether the potential impact of SI differs by student demographics. Therefore, the present study examined the efficacy of SI in promoting both proximal outcomes (grades within the course in which SI was implemented) and more distal outcomes (retention in the following year as well as 2 years later) in two different semesters. It also explored whether SI may be most strongly related to these outcomes among particular student subgroups as well as the extent to which different levels of participation in SI are each associated with student outcomes.

Supplemental Instruction and academic support in higher education

As higher education professionals shift institutional practices to support student retention, the ways in which students are supported academically have changed as well (Berger & Lyon, 2005). Tutoring was the original form of academic support for students, but a stigma developed over time to view tutoring as intended for students who were inadequately prepared for college. While tutoring remains a prevalent service at many institutions, additional academic support services began to be offered (Arendale, 1994). Due to enrollment growth and changes in student needs across higher education, learning centers were established to provide an array of activities and services for promoting students' academic success (McGee, 2005). With the incorporation of concepts and strategies from psychology, sociology, and student development theory, learning centers have worked to change the message to students, as many communicate that all students can use academic support to achieve academic success (Zimmerman, 2001). Academic support in general was initially considered a tool to utilize only when students were struggling with material; institutions have instead shifted to messaging that academic support should be used early as a tool to stay on track with coursework (Arendale, 1994, 2002).

Academic support offerings vary across institutions of higher education, yet they commonly target gateway courses. Students' success in gateway courses contributes to their academic progress and momentum and ultimately their likelihood of being retained and graduating (e.g., Adelman, 2006; Kalsbeek, 2013). Gateway courses are often high-enrollment, foundational in content, and "high-risk"; these courses are deemed high-risk when they have a sizable percentage of students who have D and F grades, withdrawals, and incompletes, also known as a DFWI rate (Koch, 2017). High-risk courses tend to have challenging content, readings, assignments, and/or exams.

Many learning centers provide SI as a form of academic support that differs from more traditional methods such as tutoring. SI is a structured, non-remedial program implemented in various countries; it is called Peer Assisted Learning in the United Kingdom and Peer Assisted Study Sessions in Australia (Arendale, 2002). When providing training for SI, UMKC's International Center for Supplemental Instruction offers some modest translation of the traditional SI model to fit well at each institution, but it emphasizes maintaining the core components of SI (Dawson et al., 2014; UMKC, 2013). SI seeks to help students gain a greater understanding of course material and to teach study skills and learning strategies (e.g., Blanc et al., 1983). These strategies support students in not only the course for which they attend SI, but also their other courses as well (McGuire, 2006). SI consists of peer-facilitated study sessions built on collaborative and active learning strategies. SI sessions occur outside of the classroom and are regularly scheduled

throughout the semester; these voluntary sessions are open to all students enrolled in the course (Dawson et al., 2014; McGuire, 2006; UMKC, 2013).

Four key characteristics of SI include peer facilitation, lecture attendance, collaborative learning activities, and question redirection (UMKC, 2013). SI sessions are led by an advanced student, referred to as an SI leader, who has already taken the course and received a strong grade. The SI leader attends the course lecture and actively engages in notetaking and reviewing readings and course materials. The SI sessions are structured around learning activities that allow the students to dive deeper into relevant content. These activities are collaborative and actively engage all students in the discussion or practice. SI leaders redirect questions—which involves reframing a student's question, asking the question back to the group of students, and/or breaking down the question—to allow students to work up to answering their question. "Facilitate" is arguably the best word to describe the role of SI leaders rather than "teach," "instruct," or "lecture," as this role involves engaging the students with the material and with one another; the leaders do not directly answer questions and are trained on how to break down questions and course content for student understanding (Stone & Jacobs, 2006; UMKC, 2013).

Another hallmark feature of SI is the incorporation of study skills and learning strategies. Study skills are best learned when rooted in meaningful content instead of abstract advice. During the activities within SI sessions, SI leaders will integrate study strategies to help students recognize varying ways of mastering course content and how to implement these approaches in other courses (Ning & Downing, 2010; UMKC, 2013; Van der Meer & Scott, 2009).

Research on the impact of Supplemental Instruction

A large number of studies have explored the extent to which SI predicts desired outcomes over the past several decades, and this work has frequently examined grades within the SI course as the outcome of interest (see Arendale, 2020). In an influential early study, Martin and Arendale (1992) analyzed data that were primarily collected from UMKC. They found that SI participation was associated with a greater likelihood of receiving an A or B grade; a reduced likelihood of receiving a D, F, or W; and a higher final course grade in every year out of 11 years of data. SI participants also had higher retention and graduation rates at UMKC than did non-participants. Moreover, when examining national data from 49 institutions, Martin and Arendale found that students who participated in SI had higher grades and lower DFW rates than those who did not; these patterns were consistent across 2-year public, 4-year public, and 4-year private institutions (also see Ogden et al., 2003).

Dawson et al. (2014) conducted a qualitative systematic review of 29 studies from 2001 to 2010 that examined the link between SI and various student outcomes. The effect sizes for SI participation predicting course grades were all positive across studies (Cohen's d ¼ .29 to .60). Although course grades constituted the primary outcome examined within this literature, they also found that SI participation was associated with a variety of other desired outcomes, including successful course completion (e.g., Cheng & Walters, 2009; Hensen & Shelley, 2003), study skills (Ning & Downing, 2010; van der Meer & Scott, 2009), academic motivation (Mack, 2007; Ning & Downing, 2010), reduced anxiety (Bronstein, 2008), development of social relationships with peers (Court & Molesworth, 2008; Dobbie & Joyce, 2008), college retention (Bowles & Jones, 2004), and graduation (Bowles et al., 2008). Among the few exceptions to these patterns, some scholars identified nonsignificant or mixed results for college retention and graduation (Ogden et al., 2003; Oja, 2012; Rath et al., 2007).

To a large extent, previous research has compared students who attended any SI sessions to those who attended no sessions, or they compared students who attended at least a certain number of SI sessions (e.g., three or five) to students who attended fewer sessions than that threshold (Arendale, 2020; Dawson et al., 2014). This approach makes it difficult to determine how much

SI is necessary to improve student outcomes. Among the notable exceptions, some research has found no significant differences in course grades between students attending no SI sessions versus those who attended one session (Romoser et al., 1997), 1–2 sessions (Malm et al., 2011; Pryor, 1990), or 1–4 sessions (Malm et al., 2011). In contrast, others have identified significant grade improvements for attending only 1–2 SI sessions (Arendale, 1997; Kochenour et al., 1997). Favorable outcomes are often reasonably large in magnitude among students who attended at least several SI sessions (e.g., Congos & Mack, 2005; Fayowski & MacMillan, 2008; Kochenour et al., 1997; Malm et al., 2011, 2018).

Despite the preponderance of positive findings from the SI literature, the quality of existing research is often less than ideal. Dawson et al. (2014) summarized some of the fundamental issues with this work: "a considerable number of studies did not provide all the details that would have allowed for a comprehensive assessment of their findings. For example, studies omitted definitions of what constituted SI attendance, number of students involved, p values, mean grades, and standard deviations" (p. 632). As a result of these problems with reporting methodology and results, Dawson et al. could not calculate effect sizes for the majority of eligible studies that predicted course grades.

As with most research on college student experiences and outcomes, self-selection into SI sessions constitutes a notable problem for drawing causal inferences about its potential effects. Many prior studies have conducted bivariate comparisons between students who did or did not participate in SI; other work has used multiple regression or analyses of covariance that controlled for demographics and/or precollege achievement. Potential issues with self-selection can lead to findings that seem unlikely when interpreted as the potential impact of SI. For instance, Malm et al.'s (2018) examination of engineering majors found that students with frequent SI attendance (more than 10 meetings) were over twice as likely to graduate within 6 years as students who did not attend any SI meetings.

Although SI participation is likely related to a variety of motivational factors, SI students tend to have standardized test scores and high school grades that are not higher than—and are sometimes actually lower than—those of non-SI students (e.g., Congos & Mack, 2005; Hensen & Shelley, 2003; Peterfreund et al., 2008; Terrion & Daoust, 2011). In addition, most of the stronger studies that sought to account directly for student motivation as well as prior achievement have generally found positive results for SI predicting course grades, college persistence, and graduation (Buchanan et al., 2019; Fayowski & Macmillan, 2008; Kenney, 1989; Terrion & Daoust, 2011). However, in arguably the most rigorous study to date, Paloyo et al. (2016) randomly assigned nearly 6,000 students to be eligible (or not) for a lottery that had a large financial incentive (gift cards of $1,000 or $5,000 Australian dollars); these randomly assigned students would be entered in the lottery in exchange for their SI session attendance. Assignment to the treatment condition substantially increased students' participation in SI, but it did not yield significant effects on course grades within the full sample and virtually all subgroups.

Theoretical framework

To understand the ways in which SI may contribute to student success overall as well as potential differential effects among student subgroups, the Multicontextual Model for Diverse Learning Environments (MMDLE) was used to inform the present study (Hurtado et al., 2012). The MMDLE offers insights into course-level dynamics; specifically, it posits that student learning and success within coursework are shaped by pedagogy and teaching methods, course content, and the identities of instructors and students. Many courses that have SI use large lectures in which students are listening and taking notes, so students are often expected to passively and individually receive information. In contrast, SI facilitates active group-based engagement in a collaborative format; this type of approach may be helpful for all students, with particular benefits for

students whose racial, socioeconomic, or gender identities are marginalized (Bowman & Culver, 2018). The large lecture structure also reinforces a hierarchical relationship between the lead instructor, who is likely to be White at many colleges and universities (National Center for Educational Statistics, 2019), and students in the U.S. who are much more diverse than their instructors in terms of race and other demographics (U.S. Census, 2018). Given that SI sessions are facilitated by advanced undergraduates, students may appreciate having the opportunity to engage with someone who is closer to a peer, but who still has expertise and training. SI leaders may have even taken the course with the same instructor, so they can provide insights and advice that are narrowly tailored toward succeeding within that class. Depending, in part, on the identities of the student and SI leader, SI may also provide a role model who may be missing from other instructional positions. These dynamics may be particularly important for marginalized identities that are often visible (e.g., underrepresented racial minorities [URM], women in STEM courses), which may lead to particularly strong effects of SI among students from these groups.

The MMDLE also describes three interrelated processes that may bolster inclusion, motivation, and success: socialization, creating community and a sense of belonging, and validation. SI is generally linked with early college courses, which serve to socialize students toward certain knowledge, skills, attitudes, and values. However, many aspects of this socialization are tacit, so students may struggle to understand these unstated norms, which can be as fundamental as how to study effectively in college. SI sessions can help address such issues through students teaching one another and through study skills instruction from the SI leader. SI sessions may also provide a space in which students find community and have the opportunity to engage meaningfully with one another. Therefore, students who may feel left out or isolated in large classrooms, such as URM students, may particularly benefit from SI. In fact, the modest representation of URM students in many courses at predominantly White institutions may result in an adverse classroom climate that contributes to racial equity gaps in academic achievement (Bowman & Denson, 2021; Oliver, 2020; but see Dills, 2018). The structure of SI may also provide opportunities for validation that are rare within lecture halls of hundreds of students and that may be particularly salient for students from marginalized identities. Rendon (1994) defines validation in terms of supportive interactions with in-class and out-of-class agents that acknowledge students' self-worth and highlight their potential for college success. Many interpersonal interactions with lead instructors or teaching assistants can be transactional in nature, especially within large courses in which instructors' attention is inherently divided among many students. In contrast, SI can become a community in which all students who have chosen to attend are dedicated to learning and success, and validation behaviors can come from the SI leader and from fellow students.

Present study

The present study seeks to improve and expand upon the previous SI literature in several ways. First, it provides a more rigorous examination of the potential causal impact of SI by using propensity score analyses to reduce or eliminate selection bias, which is a considerable concern for this voluntary form of engagement. This inquiry also directly examines the extent to which results from propensity score analyses diverge from simple bivariate comparisons between groups. Second, it explores the extent to which different levels of participation in SI may lead to desired outcomes. In other words, how much attendance of SI sessions is enough to bolster college success? Previous research has considered this issue infrequently and often relies on modest sample sizes. Third, it considers how the potential effect of SI may vary based on student demographics and precollege academic achievement. As discussed in the Theoretical Framework section, SI may be especially influential for groups of students whose identities are minoritized with higher education and society more generally. Fourth, this study uses large samples, which are especially important when conducting propensity score analyses (Shadish, 2013), whereas previous research

has frequently relied on small samples. In fact, the present subgroup analyses generally examined over 1,000 observations for first-generation students and for URM students, thereby providing sufficient statistical power to detect even modest effects of SI participation among these groups. Fifth, these analyses draw upon data from two different semesters to determine the consistency of findings across multiple samples. Finally, this study considers both short-term, course-based outcomes (overall grades and receiving a grade of D, F, W, or I) as well as longer-term outcomes (retention at the institution into the following year as well as 2 years later). Overall, this examination provides critical and rigorous evidence about whether and when SI accomplishes its primary goal of promoting student success in high-risk undergraduate courses and beyond.

Materials and methods

Data sources and participants

Participants were undergraduates who took at least one course that offered SI in Fall 2017 or Spring 2018 and who initially enrolled at a large, Midwestern public research university in Fall 2016 or later. Therefore, all participants were in their first or second year at the institution, but some had accrued higher levels of academic or class standing as a result of prior dual enrollment, Advanced Placement, International Baccalaureate, and/or transfer credits. Students were included in the study if they received any grade in the course (including a W or I); participants were also included regardless of their undergraduate major, transfer status, whether they had previously taken the course, or other such considerations. Many of the covariates were obtained from a survey that was given to all incoming students. This survey was administered for the first time in Fall 2016 through a required online college transition course; therefore, virtually every eligible student completed the survey (97% response rate).

The study employed two different samples to separately examine the impact of SI coursework in Fall 2017 and in Spring 2018. The set of SI courses was very similar, but not identical, across the two semesters. The retention outcomes were considered for two timepoints (Fall 2018 and Fall 2019), so combining those SI semesters would have led to different lengths of time between SI and retention. Moreover, this semester-based approach also reduces the number of times in which the same student took multiple courses within a particular analysis. The Fall 2017 sample consisted of 7,295 course enrollments for 4,869 students within 19 courses (six chemistry, six mathematics, three biology, three health and human physiology, and one psychology). Among these students, 59% were female, 79% were White/Caucasian, 8% were Latinx/Hispanic, 5% were Asian/Pacific Islander, 4% were multiracial, 3% were Black/African American, 1% were unknown race/ethnicity, and 26% were first-generation college students (i.e., neither parent had received a postsecondary degree).

The Spring 2018 sample consisted of 5,346 course enrollments for 3,631 students within 20 courses (these classes were the same as Fall 2017, except for two additional biology courses and one fewer health and human physiology course). The demographics of this sample were nearly identical to the prior semester: 60% were female, 78% were White/Caucasian, 9% were Latinx/Hispanic, 5% were Asian/Pacific Islander, 4% were multiracial, 3% were Black/African American, 1% were unknown race/ethnicity, and 26% were first-generation students. As shown in the Appendix, students were also similar across semesters on a variety of precollege academic achievement, psychological, very early college adjustment, and very early college engagement measures. Across the 21 total SI courses offered in one or both semesters, 17 of these were eligible to satisfy a general education requirement, and 19 of these counted toward satisfying a requirement for at least one major at the university.

Measures

Dependent variables

Two outcome variables indicated grades within the SI course. The first indicator consisted of grade points that correspond to the course letter grade (0 = F, to 4.33 = A+); students who withdrew from the course (with a W on the transcript) or had an incomplete at the time of data collection (with an I) were not included in analyses predicting this particular outcome. Moreover, because SI is intended to reduce the frequency of receiving low grades or withdrawals, a binary variable was used to indicate whether a student received a grade of D (including D + and D-), F, W, or I (0 = C- or higher, 1 = DFWI). Two additional binary variables indicated retention at the institution in Fall 2018 and Fall 2019. Some students had entered the university with additional credits, so a small percentage received a degree before the Fall 2019 semester (1.4% for the Fall 2017 semester sample and 1.2% for the Spring 2018 semester sample), and it would seem misleading to code their lack of Fall 2019 enrollment as indicating non-retention. As a result, these students were also coded as being "successful," along with the retained students (0 = neither enrolled nor graduated, 1 = retained or graduated). The same coding was also used for the exceedingly small number of students who graduated before Fall 2018 (five students total across both semester samples). Given the very small proportion of graduates and the desire to keep the language simple, these outcomes are described as "retention" throughout the paper.

Treatment variables

The primary treatment variable indicated whether students had attended any SI sessions associated with that course (0 = no, 1 = yes). Students attended at least one SI session within only 26% of course enrollments in Fall 2017 and 22% in Spring 2018, so the considerable majority of students did not participate at all in SI. In addition, a categorical measure indicating several levels of SI attendance was also used to examine these disparate amounts of treatment exposure. Among students who engaged in any SI sessions, a sizable portion attended only one SI session within the course (40% for Fall 2017 and 35% for Spring 2018), and many others attended just 2–4 SI sessions (32% for Fall 2017 and 26% for Spring 2018). Figure 1 provides a visual overview of participation levels among students who attended at least one SI session; the bar farthest to the left represents the number of students who participated in one or two sessions. To achieve sufficient sample size to identify any potential effects, four categories were examined in these analyses: attending one session, 2–4 sessions, or five or more sessions, with zero sessions as the referent or comparison group. Preliminary analyses found that using a larger number of categories led to problems with achieving appropriate covariate balance across conditions, since too few observations were present within one or more of the treatment categories.

Several factors may account for the modest levels of SI engagement at this institution within this analytic sample. First, some students only attended SI immediately before a midterm or final exam, and they appear to represent a substantial portion of students who attended only one or two sessions. Second, SI was implemented in some extremely large courses, and the SI sessions had an attendance maximum (to facilitate an ideal environment and as a result of physical space constraints), so some students were not able to attend SI sessions if the maximum enrollment had already been reached. It seems likely that such students would be far less inclined to try to go to subsequent SI sessions after being prevented from attending. Third, as a result of these space constraints and the continued development of SI in this university, the messaging about SI from faculty and administrators was sometimes not prevalent and not always targeted toward encouraging frequent and early attendance. Finally, despite attempts to vary the days and times in which SI sessions were offered for each course, some students reported having schedule conflicts that prevented them from attending most or all of the sessions. (Since the data collection

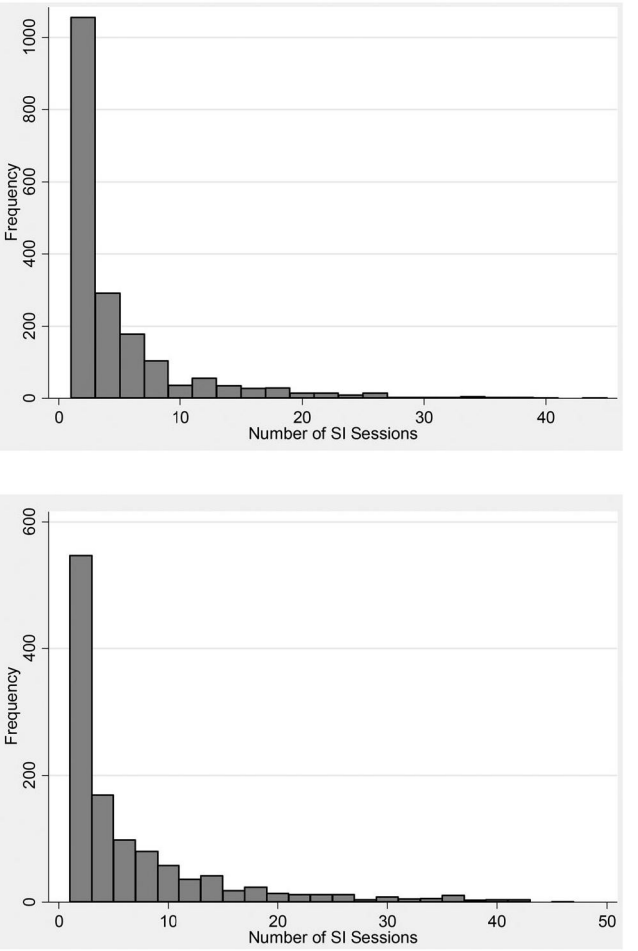

Figure 1. Number of SI sessions attended among students who engaged in any SI in Fall 2017 and Spring 2018, respectively.

period of this study, improvements have been made to the availability of SI in large courses and the messaging about SI from various constituents.)

Covariates

Prior research has demonstrated the benefits of including propensity score covariates that contribute to the treatment and/or the outcome (e.g., Brookhart et al., 2006; Patrick et al., 2011). The choice of covariates in this study was based on theory and research pertaining to college student success (e.g., Bean & Eaton, 2000; Braxton et al., 2004; Museus, 2014). Demographics, precollege academic preparation and achievement, and course subject variables were obtained from registrar data. Very early academic behaviors, nonacademic behaviors, college adjustment, and other psychological factors were obtained from a survey that served, in part, as an early alert system to identify students who may be struggling within their new college environment. This survey was administered during the third week of classes in students' first semester, so it technically could occur after some SI participation among some students who were in their first year at the university for the Fall 2017 analyses. However, this issue would likely affect a very small number of participants, since (a) the large majority of observations occurred among students who were in the Fall 2016 entering cohort and/or attended an SI course in the Spring 2018 semester, and (b) the

vast majority of students who participated in SI engaged in only one or a handful of SI sessions, and it is very unlikely that those students attended in the first couple of weeks in the semester.

Demographic variables included sex (0 = male or another response, 1 = female), race/ethnicity (dummy variables for Asian/Pacific Islander, Black/African American, Latinx/Hispanic, and other race/ethnicity, with White/Caucasian as the reference group), first-generation status (0 = continuing generation, 1 = first generation), and academic or class standing (based on the number of credits earned; 1 = first year, to 4 = senior). In addition to first-generation status, students provided their subjective social class or socioeconomic status (SES) for their family when growing up (1 = lower class, to 5 = upper class); such subjective measures add important information to "objective" SES indicators that are limited in the aspects of SES that they measure (see Rubin et al., 2014). Precollege academic achievement was assessed via high school GPA (weighted to provide an extra grade point for any honors/AP/IB courses) and ACT composite score (or SAT verbal + math equivalent if ACT score was not available). Given that virtually all SI courses were in math and science subjects, we initially considered using ACT subscores in those areas as covariates. However, all four ACT subject scores were very highly correlated with the composite score ($r_s > .80$), so the single composite score was used. Instead, math and science preparation were assessed via the number of years of high school coursework in biology, chemistry, physics, and mathematics beyond Algebra 2 (separate variables were created for each subject).

Several measures assessed students' very early academic behaviors in college. Productive academic behaviors included course attendance, notetaking, in-class participation, and completing readings (nine-item index, $\alpha = .82$ for Fall 2017 semester and $\alpha = .82$ for Spring 2018 semester). Students also reported their engagement to that point in visiting a professor during office hours (0 = none, to 4 = four or more times) or missing scheduled classes (1 = zero times, to 5 = 10 or more times). Students also reported whether they had declared a major, which may be associated with class participation and retention (Mayhew et al., 2016).

Additional variables indicated nonacademic behaviors that may reflect greater college engagement and/or experiences that may detract from time and effort spent on coursework. These single-item measures included time spent socializing (1 = less than one hour/day, to 6 = 21 or more hours/day), time spent with student or community organizations (1 = less than one hour/week to 6 = 21 or more hours/week), and time spent working on-campus and working off-campus (for each measure, 1 = less than one hour/week to 8 = more than 40 hours/week). Previous research has demonstrated consistent curvilinear relationships for co-curricular engagement and paid employment when predicting college student outcomes (Bowman & Trolian, 2017; Mayhew et al., 2016; Pascarella & Terenzini, 2005; Perna, 2010), so squared terms for these constructs were also included in the models.

Indicators of very early college adjustment were also used. Multi-item indices measured feelings of social and interpersonal belonging at the institution (four items, $\alpha = .83$ and .84, respectively), feelings of homesickness (four items, $\alpha = .89$ for both semesters), and satisfaction and intent to persist at that institution (five items, $\alpha = .91$ for both semesters). Students also reported whether a life event, such as a family emergency or financial changes, had altered their commitment to attending the university (0 = no, 1 = yes). Several additional psychological constructs were assessed; these did not directly ask about college adjustment, but some of them may have been shaped by students' very early college experiences. Students reported challenges with mental health, anxiety, and depression (three items, $\alpha = .83$ for both semesters); concerns about paying for their tuition, housing, and meals (two items, $\alpha = .95$ and .96); and expectations for their first-semester college GPA (1 = 0.0–0.5, to 9 = 4.0 or higher). Grit was assessed via two subscales: perseverance of effort (3 items, $\alpha = .87$ for both semesters) and consistency of interest toward specific long-term goals (4 items, $\alpha = .81$ and .82). The original short-form of the perseverance of effort scale had four items (Duckworth & Quinn, 2009), but including the item about discouragement from setbacks would have substantially reduced the internal reliability of the measure (α

¼ .77 or .76 instead of .87), so it was excluded from the present study. Finally, dummy variables indicated the subject of the SI course (biology, health and human physiology, mathematics, and psychology, with chemistry as the referent group). Descriptive statistics for all variables in each semester are provided in the appendix.

Analyses

This study used propensity score analyses with an augmented inverse probability weighting (AIPW) estimator. Generally speaking, propensity score analyses seek to remove selection bias by determining the likelihood that each participant will engage in the treatment and then conducting analyses that create treatment and control conditions consisting of participants who are equally likely to have engaged in the treatment. If participants in the treatment and control conditions do not differ on numerous covariates (after propensity score adjustment) but they exhibit different outcomes, then researchers can draw stronger conclusions about the potential causal effects of the treatment (for more information about these analyses, see Bai & Clark, 2019; Guo & Fraser, 2015; Holmes, 2014; Imbens & Rubin, 2015).

Several different approaches are available for using propensity scores to achieve comparable treatment and control conditions. Perhaps the most well-known technique is to directly match each participant in the treatment condition with at least one participant in the control condition who had a very similar or identical propensity score; the analyses predicting the outcome(s) then examine a sample of matched participants. Another technique compares groups of participants who are stratified or subclassified into narrow bands of propensity scores rather than matching individual participants. A third approach is to weight the sample based on the inverse probability of participating in the treatment so that the treatment and control groups as a whole have very similar or identical means on all observed covariates. For example, if female students are more likely to engage in SI (as we had expected), then participants will be weighted in a manner that corrects this imbalance (e.g., by providing higher weights to female students in the control condition). The present study employed weighting with an AIPW estimator; this involves a two-step process in which both the propensity score model and the outcome model include covariates. AIPW can use different covariates within the propensity score and outcome models, but we initially had no reason for creating divergent models, so the covariates in the original models were identical (as described in the Results section, a handful of covariates were eventually removed from the propensity score models to achieve balance across treatment and control conditions). The AIPW approach has a notable benefit: It provides unbiased estimates of the average treatment effect if either the propensity score model is correctly specified or the outcome regression is correctly specified, whereas both models must be correct for most other propensity score approaches. This methodological property is known as “double robustness” (for more information about AIPW, see Bang & Robins, 2005; Glynn & Quinn, 2010; Scharfstein et al., 1999).

Propensity score analyses often compare a single treatment and a single control condition; the present study did so when comparing participation in any SI session(s) with no participation at all. In addition to this binary treatment variable, our interest in the impact of different levels of SI engagement required us to examine multiple levels of treatment. Some prior higher education research has used a dose-response function to explore the curvilinear relationship between the amount of treatment and the outcome (e.g., number of credit hours taken per semester and upward transfer; Doyle, 2011). This dose-response modeling requires substantial sample size throughout the treatment distribution in order to achieve sufficient statistical power; therefore, it cannot be used here, since the overwhelming majority of participants have no participation or very limited participation in SI. Therefore, we used a multinomial or categorical approach in which each of the treatment levels was compared to the control condition (i.e., no SI participation) to determine what level(s) of SI might affect student success. Instead of creating the

propensity score using logistic regression, a multinomial logistic regression was used for this categorical treatment, and separate propensity scores were created to account for selection into each level of the treatment relative to the control condition (see Feng et al., 2012).

To determine whether the propensity score weighting analyses successfully reduced bias in the observed covariates, the standardized mean difference between treatment and control was computed for each covariate before and after the propensity score adjustment. This measure is well-suited as an indicator of covariate balance (Ali et al., 2014; Belitser et al., 2011), and scholars suggest that the difference between conditions should ideally be no larger than .10 standard deviations (e.g., Normand et al., 2001; Stuart et al., 2013; Zhang et al., 2019).

Different versions of the propensity score analyses were conducted to determine the robustness of the findings to alternative specifications. For instance, some models excluded the subjective SES measure, and others included students' self-reports of the types of financial aid that they were receiving (we ultimately omitted these from the analyses, in part because were skeptical about whether students could accurately report detailed financial aid information). Other analyses used inverse probability weighted regression adjustment (IPWRA), which is an alternative estimator that also conducts doubly robust propensity score analyses. As described in more detail below, some variables were removed from specific analyses if the covariates were not sufficiently balanced treatment and control conditions. Although the effect sizes sometimes varied across model specifications, the general findings and conclusions for the impact of SI remained the same.

Additional propensity score analyses examined several subgroups of interest: (1) URM students (i.e., American Indian/Alaska Native, Black/African American, Hispanic/Latinx, Native Hawaiian/Pacific Islander) and non-URM students (i.e., White/Caucasian and Asian); (2) female and male students; (3) first-generation and continuing-generation students; (4) students with relatively higher or lower weighted high school GPAs (median split of 3.8 or below versus above 3.8); and (5) students with higher or lower ACT composite scores or SAT equivalent (median split of 25 or lower versus 26 or higher). Given the substantially reduced sample sizes for many of these groups, the analyses only used the binary treatment variable comparing any SI participation to none at all.

We also conducted regression analyses to determine the simple association between SI and student success without accounting for self-selection; these results were contrasted with those from the propensity score analyses to explore how much selection bias may have inflated estimates that simply compare treatment and control conditions. Separate analyses were performed using either the binary SI variable or the four-category SI variable (three levels of treatment and one control) as the lone predictor. Ordinary least squares regression analyses were used for predicting grades, and both OLS and logistic regression examined the binary outcomes. The patterns of statistical significance and effect size estimates were virtually identical regardless of the linear versus binary modeling of the outcome variable, so OLS regression results are reported here to simplify interpretation of the results.

Robust standard errors were used in the regression analyses as well as the propensity score analyses. For all tables provided below, the regression or propensity score coefficients should be interpreted as the difference between the treatment and control conditions in terms of GPA points (for grades) or percentage points (for DFWI and retention).

Limitations

Some limitations should be noted. First, although propensity score analyses are designed to facilitate stronger causal inferences, we cannot be certain that the results presented here represent causal effects. We sought to improve our likelihood of examining causal relationships by conducting propensity score analyses that use a doubly robust AIPW estimator, employing a large number of covariates that are believed to influence selection into the treatment and/or subsequent outcomes,

establishing appropriate balance between treatment and control conditions, and exploring alternative model specifications to ensure that the results are not attributable to idiosyncratic decisions. Second, it is also unclear to what extent the present results may generalize to other institutions. Fortunately, given the substantial role of UMKC in training SI supervisors and disseminating information about SI, the implementation of this practice seems to be far more standardized and consistent than many other academic and/or programmatic interventions (e.g., first-year seminars). Third, although the patterns of results described below provide evidence about the potential mechanisms through which SI may operate, we do not have direct measures of such processes from the institutional data sources examined here. For instance, we were able to use students' academic motivation very early in college as a covariate in the propensity score analyses, but we did not subsequently assess this construct later in college, so we could not explore the potential role of SI in improving motivation and therefore contributing to students' grades and retention.

Results

Balance across treatment and control conditions

Before considering the relationship between SI and student outcomes, it is important to ensure that the propensity score analyses have successfully achieved balance between the treatment and control conditions. Table 1 contains the standardized mean differences between the SI and non-SI conditions for analyses that used a single SI treatment group. Every covariate in both semesters was below the recommended threshold of an absolute value of a .10 standard deviation (SD) difference—and nearly all were within a much more stringent threshold of .05 SD—which means the propensity score weighting analyses created treatment and control conditions that were similar on all observed covariates.

In the analyses that examined several levels of SI participation, the propensity score adjustment sufficiently balanced two of the treatment conditions (1 SI session and 2–4 SI sessions) with the control condition, but it did not yield sufficient balance for the treatment group with the highest participation (5+ SI sessions). This problem with balance was likely driven by the modest sample size within this most engaged group as well as the notable unweighted differences between students in this group and those who did not participate in any SI. To improve the balance, the variables indicating course subject were removed for predicting the treatment, but these were retained for predicting the outcome. As shown in the left-hand side of Table 2, 100 of the 102 group differences for the Fall 2017 semester were less than .10 standard deviations, and the two exceptions were not much larger than this threshold (*d*s ~ .14 for the linear and squared terms for working on-campus comparing 5+ sessions with no sessions), so this served as the final model for Fall 2017 SI participation. However, the Spring 2018 analyses were still too imbalanced when comparing the greatest level of treatment with the control condition; the largest differences occurred for students' very early college engagement with on-campus employment, off-campus employment, and co-curricular activities. Therefore, these variables were also removed as covariates for creating the propensity score, but not for predicting the outcome. After these changes, the standardized mean differences were virtually all below .10, with adverse life events for 5+ SI sessions versus none being the only exception (*d* = .11), so these served as the final analyses for Spring 2018 SI participation. The standardized mean differences between each treatment and the control condition in Spring 2018 appear on the right-hand side of Table 2.

Supplemental Instruction and college student outcomes

The results for participation in any SI session are reported in Table 3; the values for unadjusted regression analyses appear on the left, and those for the propensity score analyses appear on the

Table 1. Covariate balance between treatment and control conditions before and after propensity score adjustment (single treatment variable).

	Fall 2017 Semester Course		Spring 2018 Semester Course	
Covariate	Unweighted	Weighted	Unweighted	Weighted
Female	.332	−.006	.244	−.011
Black	.064	−.002	.081	−.005
Latinx	−.033	−.009	.032	.007
Asian	.057	.011	.053	.044
Other race	−.029	.014	.003	−.027
First-generation	−.026	−.030	.009	−.016
Subjective SES	.008	.003	−.029	−.029
Class standing	.048	−.004	.159	−.006
HS advanced math	.079	.002	.089	.003
HS biology courses	.094	.028	.015	.019
HS chemistry courses	.037	.020	.093	.001
HS physics courses	−.001	−.001	−.010	.019
HS GPA	.311	.024	.270	−.017
ACT/SAT score	.009	.024	−.008	.024
Perseverance of effort	.180	−.009	.188	.010
Consistency of interest	.127	.009	.085	−.003
Expected GPA	.218	.008	.177	−.057
Academic behaviors	.226	.007	.147	−.038
Missed classes	−.324	.013	−.264	.039
Attended office hours	.190	−.000	.162	.016
Declared major	.086	−.006	.106	−.015
Work on-campus	−.106	.013	−.023	.039
Work on (squared)	−.110	.018	−.033	.049
Work off-campus	−.134	−.003	−.096	−.040
Work off (squared)	−.138	−.001	−.109	−.044
Co-curriculars	−.034	.023	−.002	.093
Co-curriculars (squared)	−.065	.021	−.038	.097
Time socializing	−.128	.009	−.092	.037
Intent to persist	.091	−.003	.025	−.011
College belonging	−.029	−.004	−.037	.012
Homesickness	.038	.001	.130	−.032
Financial stress	.018	.002	−.007	−.020
Mental health	−.051	.015	−.048	−.009
Adverse life event	−.101	.007	−.092	.047
Enrolled in biology course	−.011	−.008	.150	−.016
Enrolled in health course	−.064	.016	.097	.002
Enrolled in math course	−.405	−.015	−.225	.001
Enrolled in psychology course	−.075	.002	−.329	.024

Note. "Unweighted" indicates the standardized mean difference between groups within the original sample; "weighted" indicates this same statistic after the propensity score weighting has been implemented. The last several covariates in this table refer to the course subject in which the student is currently enrolled.

right. Across all outcomes and in both semesters, the regression results indicate that SI participation was positively associated with grades in the SI course, retention to the next fall semester (i.e., Fall 2018), and retention into the following academic year (i.e., Fall 2019), whereas it was inversely related to receiving a DFWI in the course ($ps < .001$). All of these same relationships were also significant in the same direction when conducting the doubly robust propensity score analyses ($ps < .05$). The effect sizes were sometimes notably smaller when employing the propensity scores (these decrease by about half for course grade and DFWI in fall semester SI courses), but they were sometimes quite similar with and without propensity score analyses (for course grade and retention to the subsequent fall for spring semester SI courses).

Table 3 provides two different approaches for conceptualizing the effect sizes. One approach involves considering the simple difference between the treatment and control groups using the original metric of the outcome; these are represented by the B values on the left-hand side (presented above the standard errors, which are in parentheses). The propensity score analyses

Table 2. Covariate balance between treatment and control conditions before and after propensity score adjustment (with several levels of treatment).

	Fall 2017 Semester Course						Spring 2018 Semester Course					
	1 SI session		2–4 SI sessions		5+ SI sessions		1 SI session		2–4 SI sessions		5+ SI sessions	
Variable	Unwt	Wght	Unwt	Wght	Unwt	Wght	Unwt	Wght	Unwt	Wght	Unwt	Wght
HS advanced math	.021	−.011	.058	−.006	.187	.067	.070	−.012	.033	−.025	.147	.032
HS biology courses	.084	.011	.031	.032	.131	.068	−.014	−.003	.035	.027	.027	.049
HS chemistry courses	−.012	−.020	.038	.083	.103	.093	.088	.010	.107	.047	.088	.006
HS physics courses	−.080	−.032	.010	.010	.099	.070	.011	.010	−.111	.017	.042	.034
HS GPA	.172	.019	.334	.009	.497	.055	.148	−.011	.363	.014	.327	−.023
ACT/SAT score	−.019	−.006	−.031	.009	.097	.079	−.044	−.019	−.070	.024	.071	.036
Female	.333	.000	.394	−.041	.266	−.062	.230	.026	.376	−.038	.170	.019
First-generation	−.012	.003	−.028	.017	−.043	−.048	.008	.005	−.002	.019	.017	−.030
Black	.033	−.009	.093	−.003	.072	−.000	.034	−.016	.148	.059	.070	−.026
Latinx	−.020	−.038	−.019	.033	−.068	−.005	.074	−.000	.059	−.008	−.028	−.001
Asian	−.013	.017	.081	.046	.119	−.016	−.014	.008	.014	.053	.132	−.007
Other race	−.066	−.013	.004	.008	−.020	−.017	.053	−.010	−.092	−.042	.003	−.008
Class standing	−.072	.001	.075	.008	.176	.013	.105	.012	.161	.025	.207	.017
Subjective SES	.009	−.005	.062	.002	−.055	−.019	.042	.006	−.063	.016	.007	−.078
Intent to persist	.067	.010	.113	−.096	.101	−.089	.048	−.018	.020	.057	.008	−.017
College belonging	−.036	−.001	−.005	−.021	−.045	−.063	.022	.025	.009	−.009	−.122	−.071
Homesickness	−.005	−.012	.077	.001	.053	.042	.111	.000	.136	.016	.143	−.051
Time socializing	−.126	−.015	−.085	.041	−.179	.066	.040	.041	−.120	−.013	−.204	.022
Financial stress	.014	−.003	−.001	−.010	.047	−.031	−.120	.015	.007	.004	.090	−.072
Mental health	.004	−.006	−.046	−.013	−.140	.052	−.022	−.001	−.059	−.053	−.063	.057
Adverse life event	−.032	.009	−.138	.032	−.165	.055	−.048	.003	−.075	−.082	−.149	.113
Academic behaviors	.163	.006	.185	−.015	.371	.035	.060	−.030	.201	−.057	.192	−.057
Expected GPA	.149	.020	.168	−.010	.375	−.086	.133	−.034	.126	.043	.258	−.030
Missed classes	−.213	−.023	−.283	−.043	−.547	.066	−.118	.034	−.322	−.046	−.370	.003
Attended office hours	.109	−.028	.221	.028	.264	.011	.174	.009	.151	.013	.158	.009
Declared major	.064	−.004	.070	−.034	.136	−.069	.101	−.011	.067	−.041	.139	.077
Perseverance of effort	.099	−.003	.176	−.087	.309	.046	.075	−.020	.302	.008	.219	.051
Consistency of interest	.094	.015	.117	−.046	.185	.038	.022	.022	.060	.044	.161	−.006
Work on-campus	−.107	.028	−.141	.050	−.066	.137						
Work off-campus	−.138	−.001	−.167	−.014	−.095	.078						
Work on (squared)	−.111	.022	−.138	.043	−.079	.146						
Work off (squared)	−.151	−.012	−.176	−.033	−.084	.098						
Co-curriculars	−.058	.009	.005	.022	−.045	.085						
Co-curriculars (squared)	−.091	.015	−.034	.027	−.066	.080						

Note. "Unwt" indicates the standardized mean difference between groups within the original sample; "wght" indicates this same statistic after the propensity score weighting has been implemented. The final propensity score models for the Spring 2018 SI semester did not include paid employment or co-curricular variables when predicting the treatment.

indicate that SI contributed to a .10–.19 improvement on course grade points (using the traditional four-point scale), a 4–6 percentage-point decrease in DFWI rate, and a 3–4 percentage-point increase in retention. Another way of conceptualizing effect sizes for binary outcomes is the overall increase or decrease in the rate of an event occurring, which appears for binary outcomes on the right-hand side of each pair of columns in Table 3. For instance, in Spring 2018, 17% of students who attended no SI sessions received a DFWI grade, whereas only 11% of those who attended at least one SI session received a DFWI within the propensity score weighted sample. This result would be considered a 6 percentage-point decrease by the first effect size metric (since $17 - 11 = 6$) or a 35% decrease in the frequency of DFWI grade by this second metric (since $[17 - 11]/17 = .35$ or 35%). In terms of this latter metric across analyses, attending SI led to a 25%–35% decrease in the likelihood of receiving a DFWI grade, and it resulted in an approximately 4% increase in the number of retained students.

The results of regression and propensity score analyses examining different levels of SI participation are provided in Table 4. The regression analyses were uniformly in the expected direction and largely significant (with a few exceptions for attending one SI session versus none). For the

Table 3. Results for regression and propensity score analyses for participation in any Supplemental Instruction (SI) predicting college grades and retention.

	Unadjusted Regression Analyses		Propensity Score Analyses	
Outcomes from Fall 2017 Coursework	β (SE)	Cohen's d or 96 difference	β (SE)	Cohen's d or 96 difference
Course grade	.209*** (.023)	.23 SD	.104*** (.023)	.11 SD
DFWI in course	−.075*** (.008)	−4996	−.036*** (.010)	−2596
Retention to Fall 2018	.046*** (.008)	5.396	.037*** (.008)	4.396
Retention to Fall 2019	.051*** (.010)	6.596	.035** (.012)	4.496
Outcomes from Spring 2018 Coursework				
Course grade	.202*** (.032)	.21 SD	.188*** (.035)	.20 SD
DFWI in course	−.079*** (.010)	−4696	−.060*** (.012)	−3596
Retention to Fall 2018	.036*** (.008)	3.996	.038*** (.010)	4.196
Retention to Fall 2019	.051*** (.012)	6.296	.029* (.015)	3.696

Note. The coefficients for regression or propensity score analyses predicting binary outcomes should be interpreted as the percentage-point difference between students who did versus did not participate in SI. Cohen's d (i.e., standardized mean difference) is provided for the continuous outcome of course grade, whereas the percentage change (for the treatment relative to the control group) is presented for the binary outcomes of DFWI and retention. Participation in any SI sessions was the lone predictor for the regression analyses, whereas the propensity score analyses with augmented inverse probability weighting included covariates in both the treatment and outcome models. Robust standard errors were used in all analyses.

*$p < .05$,
**$p < .01$,
***$p < .001$.

most part, these relationships were notably stronger for attending 5+ SI sessions than for attending one or 2–4 SI sessions. For the propensity score analyses, attending just one SI session versus none was associated with higher course grades (for fall semester coursework only), lower DFWI rate (fall courses), greater retention to the following year (fall and spring courses), and greater retention 2 years after SI (spring courses). Attending 2–4 SI sessions also had benefits for DFWI and both retention indicators for fall semester courses as well as retention to Fall 2019 for spring courses. Moreover, the outcomes for five or more SI sessions were notable in magnitude for both semesters. Although these results were nonsignificant for predicting retention to Fall 2019, significant findings included notable associations for retention to Fall 2018 (4–7 percentage points), higher course grades (.24–.37 grade points) and much lower chances of receiving a DFWI in that course (8–12 percentage points). Using a different metric of effect size, these DFWI findings for 5+ SI sessions were especially impressive when viewed as 55%–69% percentage declines in DFWI rates relative to students who did not participate in any SI (those figures are not displayed in Table 4 due to space constraints).

Subgroup analyses examined the impact of SI by race/ethnicity, sex, first-generation status, high school GPA, and ACT scores. Race/ethnicity was the only precollege characteristic for which the results differed consistently across semesters, so these subgroup findings are displayed in Table 5. Although URM students only comprise about 15% of each sample and therefore the corresponding subgroup analyses had much less statistical power, virtually all effects of SI were significant and in the expected direction for URM students (except for a nonsignificant 4.2 percentage-point relationship for retention to Fall 2019 for spring coursework). The results for non-URM students were generally positive and significant, but the findings for retention to Fall 2019 were nonsignificant in both semesters, and the effect sizes were generally smaller. For instance, SI was associated with 5%–11% gains in retention for URM students versus ~3% for non-URM students; SI also led to a 34%–50% decline in the DFWI rate for URM students versus a 20%–30% decline for non-URM students. Supplemental analyses compared the coefficients across these subgroup results to examine significant differences in the effect of SI for URM versus non-URM students (see Cohen et al., 2003). Overall, within fall SI coursework, the relationships for URM students were significantly stronger when predicting course grade and Fall 2018 retention than those for non-URM students; the decline in DFWI associated with SI attendance was also significantly stronger for URM students than for non-URM students within spring SI coursework.

Table 4. Results for regression and propensity score analyses for participation in different levels of Supplemental Instruction (SI) predicting college grades and retention.

Outcomes from Fall 2017 Coursework	Unadjusted Regression Analyses			Propensity Score Analyses		
	1 session	2–4 sessions	5+ sessions	1 session	2–4 sessions	5+ sessions
Course grade	.152*** (.034)	.129*** (.036)	.371*** (.035)	.077** (.032)	.072* (.042)	.240*** (.041)
DFWI in course	−.052* (.012)	−.067*** (.013)	−.119*** (.009)	−.038** (.012)	−.037* (.018)	−.080*** (.018)
Retention to Fall 2018	.030 (.012)	.045*** (.012)	.071*** (.011)	.034 (.012)	.031* (.015)	.069*** (.014)
Retention to Fall 2019	.018 (.015)	.055*** (.016)	.093*** (.015)	.024 (.015)	.043* (.019)	.038 (.026)
Outcomes from Spring 2018 Coursework						
Course grade	.036 (.048)	.171*** (.051)	.363*** (.040)	.009 (.047)	.095 (.055)	.365*** (.042)
DFWI in course	−.033** (.018)	−.071*** (.018)	−.126*** (.011)	−.021* (.019)	−.031 (.025)	−.116*** (.014)
Retention to Fall 2018	.034** (.012)	.031* (.014)	.041*** (.011)	.030* (.013)	.029 (.017)	.038** (.013)
Retention to Fall 2019	.049** (.018)	.063** (.020)	.043* (.017)	.043* (.020)	.062** (.023)	.011 (.022)

Note. The coefficients for regression or propensity score analyses predicting binary outcomes should be interpreted as the percentage-point difference between students who did versus did not participate in SI. Cohen's d (i.e., standardized mean difference) is provided for the continuous outcome of course grade, whereas the percentage change (for the treatment relative to the control group) is presented for the binary outcomes of DFWI and retention. Three dummy-coded variables indicating participation in different numbers of SI sessions (with zero sessions as the referent group) were the lone predictors for the regression analyses, whereas the propensity score analyses with augmented inverse probability weighting included covariates in both the treatment and outcome models. Robust standard errors were used in all analyses.

*$p < .05$,
**$p < .01$,
***$p < .001$.

Discussion

This paper provides some of the strongest evidence to date about the efficacy of Supplemental Instruction for promoting student success outcomes within and beyond SI coursework by conducting propensity score analyses using two large samples of students and SI courses. The positive relationships between SI and student success outcomes are consistent with previous studies (see Arendale, 2020; Dawson et al., 2014), but the present research design supports stronger conclusions about the extent to which SI actually caused these improvements. Many of the results reported here fall near Mayhew et al.'s (2016) guidelines of "small" effect sizes within college impact research, which they describe as .15 SDs for continuous outcomes and five percentage points for binary outcomes. That said, these authors also recommended that their guidelines be contextualized based on a variety of factors, several of which suggest that the magnitude of the present findings should be viewed more favorably. The ability to draw causal inferences in the present study is enhanced by the use of doubly robust propensity score analyses that employed an extensive set of covariates that may shape both participation in SI and college student success. The same effect size is more impressive when selection bias has been largely or entirely eliminated and therefore provides a better estimate of the true causal effect. In addition, SI is a reasonably modest intervention in terms of scope and cost when compared with some other initiatives. The fact that voluntary collaborative learning sessions associated with a single course may promote retention at the university is certainly noteworthy. In a related consideration, most students who engaged in SI attended only one or a handful of 50-minute sessions, so the "dosage" of this intervention was quite small within the present study. Finally, SI led to greater retention not only in the following year, but also in the year after that; the lasting nature of these results is also impressive.

Table 5. Results of propensity score analyses for participation in any Supplemental Instruction (SI) predicting college grades and retention among underrepresented racial minority (URM) and non-URM students.

	Underrepresented Racial Minority Students		Non-URM Students	
Outcomes from Fall 2017 Coursework	β (SE)	Cohen's d or % difference	β (SE)	Cohen's d or % difference
Course grade*	.219*** (.066)	.23 SD	.077** (.025)	.09 SD
DFWI in course*	−.079*** (.028)	−3%	−.026** (.011)	−2%
Retention to Fall 2018	.090* (.021)	10.7%	.026 (.010)	3.0%
Retention to Fall 2019	.067 (.030)	8.7%	.023+ (.013)	2.9%
Outcomes from Spring 2018 Coursework				
Course grade*	.169*** (.073)	.18 SD	.186*** (.042)	.20 SD
DFWI in course*	−.121* (.026)	−5%	−.047** (.013)	−3%
Retention to Fall 2018	.052 (.021)	5.8%	.031 (.011)	3.4%
Retention to Fall 2019	.042 (.035)	5.3%	.026 (.016)	3.2%

Note. The asterisk(s) next to outcome variable names indicate that these coefficients differ significantly between URM and non-URM students. The coefficients for regression or propensity score analyses predicting binary outcomes should be interpreted as the percentage-point difference between students who did versus did not participate in SI. Cohen's d (i.e., standardized mean difference) is provided for the continuous outcome of course grade, whereas the percentage change (for the treatment relative to the control group) is presented for the binary outcomes of DFWI and retention. Participation in any SI sessions was the lone predictor for the regression analyses, whereas the propensity score analyses with augmented inverse probability weighting included covariates in both the treatment and outcome models. Robust standard errors were used in all analyses.

* $p < .05$,
** $p < .01$,
*** $p < .001$.

The findings for the amount of SI participation are intriguing. Consistent with prior literature that used a different research design (e.g., Malm et al., 2011, 2018; Romoser et al., 1997), the course-related outcomes in this quasi-experimental study were particularly impressive for attending five or more SI sessions. Relative to not attending any SI sessions, students with this higher level of SI engagement had considerably greater academic achievement (equivalent to adding a plus or removing a minus from a letter grade), and their chances of receiving a DFWI decreased by over half in both semesters. SI was originally designed to reduce these poor course outcomes (e.g., Blanc et al., 1983), so it appears to be working quite well for students who participate in more than several sessions. The highest engagement group in these analyses of 5+ sessions included a fair number of students who participated 5–7 times over the semester, which is less than once every other week (out of a possible 45 sessions total). Unfortunately, there was not sufficient sample size in either semester to provide accurate estimates for students who engaged in very high levels of SI attendance, as these students may have exhibited even more favorable outcomes.

When considered together, several sets of results suggest that SI likely has effects that extend beyond the mastery of course content, which may be driven by students who attend SI coming to feel that the institution cares about their success and well-being. Multiple theories and frameworks state that this perception is critical for retention, regardless of whether that dynamic is framed as institutional commitment to the welfare of students (Braxton et al., 2004, 2014), receiving validation (Rendon, 1994, 2002), or fostering culturally engaging campus environments (Museus, 2014). As the first piece of evidence for this potential mechanism, the absolute value of the effect size for predicting DFWI rates is basically identical to that for retention within fall semester SI courses. The spring SI coursework results are more disparate, but the relationships for retention are not much smaller than those for DFWI, especially for retention to Fall 2018. If the impact of SI were driven entirely by preventing low SI course grades, then this confluence of results would be extremely unlikely; it would suggest that basically every student who received a C- or better grade as a result of SI attendance would have dropped out of the university if their grade in this single course had instead been a DFWI.

Second, attending only one SI session was significantly associated with greater course grades and retention. It seems highly doubtful that the course-related content and/or relevant study skills obtained from 50 minutes of engagement were entirely responsible for a several percentage-point effect on retention; therefore, some other process must be driving at least part of these results, which could include greater motivation that occurs after engaging in the SI learning environment (Mack, 2007; Ning & Downing, 2010). We encourage treating the results of individual analyses for participating in a single SI session cautiously, since most of the findings for each outcome were not replicated across semesters. That said, the presence of five different significant results using propensity score analyses with a wide range of relevant covariates is noteworthy, thereby suggesting that even minimal exposure to SI may be playing some role in improving student outcomes. Although the prior research is mixed on attending very few SI sessions, this finding is consistent with a couple of prior studies that have found significant results for attending 1–2 SI sessions versus no attendance (Arendale, 1997; Kochenour et al., 1997).

Third, the apparent impact of SI was frequently higher among URM students than non-URM students; several of the results differed significantly between those two groups. Previous research has also sometimes found that the relationships between SI participation and student success were larger among URM students than non-URM students (e.g., Peterfreund et al., 2008; Rath et al., 2007; Wilson & Rossig, 2014). URM students often encounter a more hostile campus climate at predominantly White institutions than do White students (see Harper & Hurtado, 2007; Hurtado et al., 2012), and URM students are therefore less likely to believe that their college or university cares about them (Hurtado & Ruiz Alvarado, 2015; Zhou & Castellanos, 2013). It makes sense, then, that SI would have a stronger impact among groups of students who tend to doubt their institution's support and concern if this constitutes a salient mechanism for SI promoting student success.

Fourth, although the effects of SI were more pronounced for URM students, the subgroup analyses also showed favorable results for non-URM students, and the results did not differ systematically by students' sex, first-generation status, high school GPA, or standardized test scores. This consistency across several precollege characteristics indicates that SI may be a potential fruitful approach for promoting the success of all students, which runs counter to the notion that academic support is only helpful or should only be tailored toward students who enter college with lower academic preparation. When considered alongside the differential effects for URM versus non-URM students, the lack of divergent results by precollege academic achievement further suggests that SI may improve student outcomes, at least in part, through its role in fostering a sense of belonging and community. The similar findings for first-generation and continuing-generation students also suggests that the visibility of minority status may play a role in this process, as first-generation students may be less able to assess whether they share this identity with their SI leader or instructor. More generally, the present findings of overall main effects of SI among all students, along with some stronger results for students from certain minoritized identities, fits well with patterns from the broader literature on collaborative learning and college student outcomes (Mayhew et al., 2016; Pascarella & Terenzini, 2005).

As a methodological consideration, the propensity score analyses generally reduced the effect size estimates relative to a simple comparison between students who attended versus did not attend any SI sessions. The overall unadjusted relationships for course grades observed in the present study (Cohen's $d = .23$ for fall semester and .21 for spring semester coursework) actually fall just below the range of values that Dawson et al. (2014) identified in their systematic review ($d = .29$ to .60), which often reflected studies that conducted bivariate comparisons. The changes in these estimates as a result of employing propensity score analyses varied somewhat by semester and by outcome. These reductions in effect size were relatively modest—and sometimes virtually non-existent—for spring SI coursework and for retention to Fall 2018. In contrast, using propensity score analyses reduced the effect size estimate by approximately 1/3 for retention to Fall 2019

in both semesters and by about 1/2 for course grades and DFWI in the fall semester. The balance statistics in the present study provide further insight into these dynamics. Students in the treatment and control conditions had similar average ACT scores, which is consistent with previous research (e.g., Congos & Mack, 2005; Peterfreund et al., 2008). However, SI participants had much better high school grades than non-participants, so this precollege achievement measure appears to be useful for removing selection bias (at least within the present sample). Students in the treatment and control conditions differed in a variety of other ways before the propensity score weighting, which further illustrates potential role of selection bias in obscuring the effectiveness of SI.

Future research

Further research is needed to foster an understanding of whether, when, and how SI contributes to student success. Additional quasi-experimental or experimental studies at other institutions could help bolster—or perhaps challenge—the generalizability of the present findings. A handful of prior studies have randomly assigned some required discussion sections to use an approach that mirrors SI sessions, but the sample sizes were quite small, and the findings have been mixed (Fest, 2000; Kenney, 1989; Khan, 2018). Obtaining an even larger sample of courses and institutions would also allow researchers to explore other conditional effects, such as possible differential results for general education versus major-specific classes, introductory versus more advanced classes, attending SI sessions earlier versus later in the semester, and attending larger numbers of SI sessions (e.g., 6–10 versus 11 or more). Although SI is fairly uniform in its implementation, such work could also consider the effects of the number of sessions offered per week, SI leaders' relative emphasis on course content versus relevant studying strategies, and other logistical choices.

Specific efforts to bolster SI attendance could be tested via a cluster-randomized trial that randomly assigns lab or discussion sections to receive (or not) a targeted persuasion technique. SI proponents tout the voluntary nature of this intervention (Hurley et al., 2006), but providing a modest amount of extra credit for SI attendance may be helpful and consistent with this philosophy. For decades, research has demonstrated that people who engage in behaviors in exchange for a very small incentive tend to adopt attitudes and values that are consistent with internal motivation (e.g., Festinger & Carlsmith, 1959), so this strategy should not undermine student engagement in the same way that a large financial incentive might do so. Generally speaking, these types of psychologically informed approaches toward practical implementation decisions may be fruitful for maximizing the potential benefits of SI and other college student success efforts.

Conclusion

The present study provides intriguing evidence for the benefits of Supplemental Instruction on academic achievement and retention. Relative to previous research on SI, this study offered stronger causal inferences through the use of doubly robust propensity score analyses within a sizable number of courses and students across multiple semesters, along with identifying larger effects among underrepresented racial minority students and students who participated in at least 5 SI sessions. Some rigorous research has found no significant impact of widely used approaches for fostering college student success, such as summer bridge programs, first-year seminars, and linked learning communities (e.g., Culver & Bowman, 2020; Lesik, Santoro, & DePeau, 2015; What Works Clearinghouse, 2014, 2015, 2016b). Therefore, the present inquiry adds important support and nuance to research on a common postsecondary intervention, as prior inquiry has often been limited in its attempts to rule out alternative explanations for its findings (Dawson et al., 2014).

These findings not only suggest that institutions should invest in SI offerings, but they also provide some insight into the ways in which practitioners should promote SI attendance at the available sessions. A large-scale experimental study on SI found no significant effect on course grades of being randomly assigned to entry in a (substantial) lottery in exchange for attending SI sessions (Paloyo et al., 2016). However, this extrinsic incentive may not have motivated students to engage meaningfully in SI activities, since those students may have been largely concerned with simply becoming eligible for the lottery drawing. Therefore, higher education staff and instructors should instead promote engagement in SI sessions by emphasizing the learning and achievement gains that may result from attendance.

Note

1. As supplemental analyses, we explored the overall impact of SI separately within each of two individual courses that met the sample size conditions for propensity score analyses offered by Shadish (2013); both of these introductory chemistry courses contained over 200 students who participated in SI and more than 800 students total. This examination of individual courses is consistent with a substantial amount of prior literature on the outcomes associated with SI. The same AIPW propensity score weighting approach was utilized, and sufficient balance between treatment and control conditions was achieved. SI participation had a positive effect on grades in both courses (.12–.13 grade points, ps $\leq$ .01); SI was also significantly associated with lower DFWI rates in one course (6.0 percentage points and 42% decrease, $p <$ .01), but this result was not significant for the other course (3.2 percentage points and 23% decrease, $p = .16$). All results for SI and retention were non-significant, which stems from a combination of the smaller sample sizes and the relatively modest positive effect sizes in these course-specific analyses (1–3 percentage points).

ORCID

Nicholas A. Bowman http://orcid.org/0000-0001-8899-7383

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Office of Vocational and Adult Education, U.S. Department of Education.

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., de Boer, A., & Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: A simulation study. *Pharmacoepidemiology and Drug Safety*, *23*(8), 802–811. https://doi.org/10.1002/pds.3574

Arendale, D. R. (1994). Understanding the Supplemental Instruction model. In D. C. Martin & D. R. Arendale (Eds.), *Supplemental Instruction: Increasing student achievement and retention* (New Directions for Teaching and Learning, no. 60, pp. 11–21). Jossey-Bass. https://doi.org/10.1002/tl.37219946004

Arendale, D. R. (1997). 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. In S. Miodusk & G. Enright (Eds.), *Proceedings of the 17th and 18th annual institutes for learning assistance professionals: 1996 and 1997*. University Learning Center, University of Arizona.

Arendale, D. R. (2002). History of Supplemental Instruction (SI): Mainstreaming of developmental education. In D. B. Lundell, & J. L. Higbee (Eds.), *Histories of developmental education* (pp. 15–28). Center for Research on Developmental Education and Urban Literacy, University of Minnesota.

Arendale, D. R. (2020). *Annotated bibliography—Supplemental Instruction*. https://www.arendale.org/peer-learning-bib

Bowman, N. A., & Denson, N. (2021). *Institutional racial representation and equity gaps in college graduation*. Unpublished manuscript.

Bai, H., & Clark, M. H. (2019). *Propensity score methods and applications*. SAGE.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bean, J., & Eaton, S. (2000). A psychological model of college student retention. In J. Braxton (Ed.), *Rethinking the departure puzzle: New theory and research on college student retention* (pp. 48–62). Vanderbilt University Press.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., de Boer, A., & Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, *20*(11), 1115–1129. https://doi.org/10.1002/pds.2188

Berger, J. B., & Lyon, S. C. (2005). Past to present: A historical look at retention. In A. Seidman (Ed.), *College student retention: Formula for student success* (pp. 1–29). American Council on Education/Praeger.

Blanc, R. A., Debuhr, L. E., & Martin, D. C. (1983). Breaking the attrition cycle: The effects of supplemental instruction on undergraduate performance and attrition. *Journal of Higher Education*, *54*(1), 80–90. https://doi.org/10.1080/00221546.1983.11778153

Bowles, T. J., & Jones, J. (2004). An analysis of the effectiveness of Supplemental Instruction: The problem of selection bias and limited dependent variables. *Journal of College Student Retention*, *5*(2), 235–243.

Bowles, T. J., McCoy, A. C., & Bates, S. C. (2008). The effect of supplemental instruction on timely graduation. *College Student Journal*, *42*, 853–859.

Bowman, N. A., & Culver, K. (2018). Promoting equity and student learning: Rigor in undergraduate academic experiences. In C. M. Campbell (Ed.), *Reframing notions of rigor: Building scaffolding for equity and student success* (New Directions for Higher Education, no. 181, pp. 47–57). Jossey-Bass. https://doi.org/10.1002/he.20270

Bowman, N. A., & Trolian, T. L. (2017). Is more always better? The curvilinear relationships between college student experiences and outcomes. *Innovative Higher Education*, *42*(5–6), 477–489. https://doi.org/10.1007/s10755-017-9403-1

Braxton, J. M., Doyle, W. R., Hartley, H. V., Hirschy, A. S., Jones, W. A., & McClendon, M. K. (2014). *Rethinking college student retention*. Jossey-Bass.

Braxton, J. M., Hirschy, A. S., & McClendon, S. A. (2004). *Understanding and reducing college student departure (ASHE-ERIC Higher Education Report, vol. 30, no. 3)*. Wiley.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156. https://doi.org/10.1093/aje/kwj149

Bronstein, S. B. (2008). Supplemental Instruction: Supporting persistence in barrier courses. *The Learning Assistance Review*, *13*(1), 31–45.

Buchanan, E. M., Valentine, K. D., & Frizell, M. L. (2019). Supplemental Instruction: Understanding academic assistance in underrepresented groups. *Journal of Experimental Education*, *87*(2), 288–298. https://doi.org/10.1080/00220973.2017.1421517

Cheng, D., & Walters, M. (2009). Peer-assisted learning in mathematics: An observational study of student success. *Journal of Peer Learning*, *2*, 23–39.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum.

Court, S., & Molesworth, M. (2008). Course-specific learning in peer assisted learning schemes: A case study of creative media production courses. *Research in Post-Compulsory Education*, *13*(1), 123–134. https://doi.org/10.1080/13596740801903729

Culver, K., & Bowman, N. A. (2020). Is what glitters really gold? A quasi-experimental study of first-year seminars and college student success. *Research in Higher Education*, *61*(2), 167–196. https://doi.org/10.1007/s11162-019-09558-8

Congos, D., & Mack, A. (2005). Supplemental instruction's impact in two freshman chemistry classes: Research, modes of operation, and anecdotes. *Research & Teaching in Developmental Education*, *21*(2), 43–64.

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: A systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research*, *84*(4), 609–639. https://doi.org/10.3102/0034654314540007

Dills, A. K. (2018). Classroom diversity and academic outcomes. *Economic Inquiry*, *56*(1), 304–316. https://doi.org/10.1111/ecin.12481

Dobbie, M., & Joyce, S. (2008). Peer-assisted learning in accounting: A qualitative assessment. *Asian Social Science*, *4*(3), 18–25.

Douglas, D., & Attewell, P. (2014). The bridge and the troll underneath: Summer bridge programs and degree completion. *American Journal of Education*, *121*(1), 87–109. https://doi.org/10.1086/677959

Doyle, W. R. (2011). Effect of increased academic momentum on transfer rates: An application of the generalized propensity score. *Economics of Education Review*, *30*(1), 191–200. https://doi.org/10.1016/j.econedurev.2010.08.004

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, *91*(2), 166–174.

Fayowski, V., & MacMillan, P. D. (2008). An evaluation of the Supplemental Instruction programme in a first year calculus course. *International Journal of Mathematical Education in Science and Technology*, *39*(7), 843–855. https://doi.org/10.1080/00207390802054433

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., & Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, *31*(7), 681–697. https://doi.org/10.1002/sim.4168

Fest, B. J. R. (2000). The effects of Supplemental Instruction (SI) on student performance in a college-level biology course [Doctoral dissertation, The University of Texas at Austin. *Dissertation Abstracts International*, *80*(09), 3311.].

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*(1), 36–56. https://doi.org/10.1093/pan/mpp036

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). SAGE.

Harper, S. R., & Hurtado, S. (2007). Nine themes in campus racial climates and implications for institutional transformation. In S. R. Harper & L. D. Patton (Eds.), *Responding to the realities of race on campus* (New Directions for Student Services, no. 120, pp. 7–24). Jossey-Bass. https://doi.org/10.1002/ss.254

Hensen, K. A., & Shelley, M. C. (2003). The impact of supplemental instruction: Results from a large, public, Midwestern university. *Journal of College Student Development*, *44*(2), 250–259. https://doi.org/10.1353/csd.2003.0015

Holmes, W. M. (2014). *Using propensity scores in quasi-experimental designs*. SAGE.

Hossler, D., & Bontrager, B. (Eds.). (2014). *Handbook of strategic enrollment management*. Jossey-Bass.

Hurley, M., Jacobs, G., & Gilbert, M. (2006). The basic SI model. In M. E. Stone & G. Jacobs (Eds.), *Supplemental instruction: New visions for empowering student learning* (New Directions for Teaching and Learning, no. 106, pp. 11–22). Jossey-Bass. https://doi.org/10.1002/tl.229

Hurtado, S., Alvarez, C. L., Guillermo-Wann, C., Cuellar, M., & Arellano, L. (2012). A model for diverse learning environments. In J. C. Smart & M. B. Paulsen (Eds.), *Higher education: Handbook of theory and research* (Vol. 27, pp. 41–122). Springer.

Hurtado, S., & Ruiz Alvarado, A. (2015). *Discrimination and bias, underrepresentation, and sense of belonging on campus (HERI Research Brief)*. Higher Education Research Institute, University of California, Los Angeles.

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

Kalsbeek, D. H. (2013). Framing retention for institutional improvement: A 4 Ps framework. In D. H. Kalsbeek (Ed.), *Reframing retention strategy for institutional improvement* (New Directions for Higher Education, no. 161, pp. 5–14). Jossey-Bass. https://doi.org/10.1002/he.20041

Kenney, P. A. (1989). Effects of Supplemental Instruction (SI) on student performance in a college-level mathematics course [Doctoral dissertation, The University of Texas at Austin]. *Dissertation Abstracts International*, *50*(2), 378A.

Khan, B. R. (2018). *The effectiveness of Supplemental Instruction and online homework in first-semester calculus* [Doctoral dissertation, Teachers College, Columbia University]. https://academiccommons.columbia.edu/doi/10.7916/D8CN8MW4

Koch, A. K. (2017). It's about the gateway courses: Defining and contextualizing the issue. In A. K. Koch (Ed.), *Improving teaching, learning, equity, and success in gateway courses* (New Directions for Higher Education, no. 180, pp. 11–17). Jossey-Bass. https://doi.org/10.1002/he.20257

Kochenour, E. O., Jolley, D. S., Kaup, J. G., Patrick, D. L., Roach, K. D., & Wenzler, L. A. (1997). Supplemental Instruction: An effective component of student affairs programming. *Journal of College Student Development*, *38*(6), 577–586.

Lesik, S. A., Santoro, K. G., & DePeau, E. A. (2015). Evaluating the effectiveness of a mathematics bridge program using propensity scores. *Journal of Applied Research in Higher Education*, *7*(2), 331–345. https://doi.org/10.1108/JARHE-01-2014-0010

Mack, A. C. (2007). Differences in academic performance and self-regulated learning based on level of student participation in Supplemental Instruction. (Ph.D. dissertation), University of Central Florida.

Malm, J., Bryngfors, L., & Fredriksson, J. (2018). Impact of Supplemental Instruction on dropout and graduation rates: An example from 5-year engineering programs. *Journal of Peer Learning*, *11*(1), 76–88.

Malm, J., Bryngfors, L., & Morner, L.-L. (2011). Improving student success in difficult engineering education courses though Supplemental Instruction (SI): What is the impact of the degree of SI attendance? *Journal of Peer Learning*, *4*(1), 16–23.

Martin, D. C., & Arendale, D. R. (1992). *Supplemental instruction: Improving first-year student success in high-risk courses*. The National Resource Center for The Freshman Year Experience, University of South Carolina.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Seifert, T. A., Wolniak, G. C., With Pascarella, E. T., & Terenzini, P. T. (2016). *How college affects students (Vol. 3): 21st century evidence that higher education works*. Jossey-Bass.

McGee, J. (2005). *Cognitive, demographic, and motivational factors as indicators of help-seeking in supplemental instruction* [Doctoral dissertation]. Texas A&M University. https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/2325/etd-tamu-2005A-EDAD-McGee.pdf?sequence=1&isAllowed=y

McGuire, S. Y. (2006). The impact of Supplemental Instruction on teaching students *how* to learn. In M. E. Stone & G. Jacobs (Eds.), *Supplemental Instruction: New visions for empowering student learning* (New Directions for Teaching and Learning, no. 106, pp. 3–10). Jossey-Bass. https://doi.org/10.1002/tl.228

Museus, S. D. (2014). The Culturally Engaging Campus Environments (CECE) Model: A new theory of college success among racially diverse student populations. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (Vol. 29, pp. 189–227). Springer.

National Center for Educational Statistics. (2019). Race/ethnicity of college faculty. https://nces.ed.gov/fastfacts/display.asp?id=61

Ning, K., & Downing, K. (2010). The impact of Supplemental Instruction on learning competence and academic performance. *Studies in Higher Education*, *35*(8), 921–928.

Normand S. L. T., Landrum M. B., Guadagnoli E., Ayanian J. Z., Ryan T. J., Cleary P. D., & McNeil B. J. (2001). Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*(4), 387–398.

Ogden, P., Thompson, D., Russell, A., & Simons, C. (2003). Supplemental Instruction: Short and long-term impact. *Journal of Developmental Education*, *26*, 2–8.

Oja, M. (2012). Supplemental instruction improves grades but not persistence. *College Student Journal*, *46*(2), 344–349.

Oliver, D. M. (2020, March). *Classmates like me: Race and ethnicity in college [Paper presentation]*. Paper Presented at the Annual Meeting of the Association for Education Finance and Policy.

Paloyo, A., Rogan, S., & Siminski, P. (2016). The effect of supplemental instruction on academic performance: An encouragement design experiment. *Economics of Education Review*, *55*, 57–69. https://doi.org/10.1016/j.econedurev.2016.08.005

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students. (Vol. 2): A third decade of research*. Jossey-Bass.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology – An empirical illustration. *Pharmacoepidemiology and Drug Safety*, *20*(6), 551–559. https://doi.org/10.1002/pds.2098

Peterfreund, A. R., Rath, K. A., Xenos, S. P., & Bayliss, F. (2008). The impact of supplemental instruction on students in STEM courses: Results from San Francisco State University. *Journal of College Student Retention: Research, Theory & Practice*, *9*(4), 487–503. https://doi.org/10.2190/CS.9.4.e

Pryor, S. A. (1990). The relationship of Supplemental Instruction and final grades of students enrolled in high-risk courses [Doctoral dissertation, Western Michigan University]. *Dissertation Abstracts International*, *50*(7), 1963A.

Rath, K. A., Peterfreund, A. R., Xenos, S. P., Bayliss, F., & Carnal, N. (2007). Supplemental instruction in Introductory Biology I: Enhancing the performance and retention of underrepresented minority students. *CBE Life Sciences Education*, *6*(3), 203–216. https://doi.org/10.1187/cbe.06-10-0198

Rendon, L. I. (1994). Validating culturally diverse students: Toward a new model of learning and student development. *Innovative Higher Education*, *19*, 33–50.

Rendon, L. I. (2002). Community college Puente: A validating model of education. *Educational Policy*, *16*(4), 642–667. https://doi.org/10.1177/0895904802016004010

Romoser, M. A., Rich, C. E., Williford, A. M., & Kousaleous, S. L. (1997). Supplemental Instruction at Ohio University: Improving student performance. In P. L. Dwinell & J. L. Higbee (Eds.), *Developmental education: Enhancing student retention* (pp. 37–44). National Association for Developmental Education.

Rubin, M., Denson, N., Kilpatrick, S., Matthews, K. E., Stehlik, T., & Zyngier, D. (2014). I am working-class": Subjective self-definition as a missing measure of social class and socioeconomic status in higher education research. *Educational Researcher*, *43*(4), 196–200. https://doi.org/10.3102/0013189X14528373

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametrics nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120. https://doi.org/10.1080/01621459.1999.10473862

Shadish, W. R. (2013). Propensity score analysis: Promise, reality, and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129–144. https://doi.org/10.1007/s11292-012-9166-8

Stone, M., & Jacobs, G. (Eds.). (2006). *Supplemental Instruction: New visions for empowering student learning* (New Directions for Teaching and Learning, no. 106). Jossey-Bass.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8 Suppl), S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013

Terrion, J. L., & Daoust, J. (2011). Assessing the impact of supplemental instruction on the retention of undergraduate students after controlling for motivation. *Journal of College Student Retention: Research, Theory & Practice*, *13*(3), 311–327. https://doi.org/10.2190/CS.13.3.c

University of Missouri-Kansas City (UMKC). (2013). *Supplemental Instruction supervisor manual*. Curators of the University of Missouri.

U.S. Census. (2018). *More than 76 million students enrolled in U.S. schools*. Census Bureau reports. https://www.census.gov/newsroom/press-releases/2018/school-enrollment.html

van der Meer, J., & Scott, C. (2009). Students' experiences and perceptions of peer assisted study sessions: Towards ongoing improvement. *Australasian Journal of Peer Learning*, *2*(1), 3–22.

What Works Clearinghouse. (2014). *WWC Intervention Report: Linked learning communities*. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2015). *WWC Intervention Report: Developmental summer bridge programs.* Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2016a). *WWC Intervention Report: First year experience courses*. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2016b). *WWC Intervention Report: First year experience courses for students in developmental education*. Institute of Education Sciences, U.S. Department of Education.

Wilcox, F. K., & Jacobs, G. (2008). Thirty-five years of Supplemental Instruction: Reflections on study groups and student learning. In M. E. Stone & G. Jacobs (Eds.), *Supplemental Instruction: Improving first-year student success in high-risk courses* (Monograph no. 7, 3rd ed., pp. vii–vix). National Resource Center for the First-Year Experience and Students in Transition, University of South Carolina.

Wilson, B., & Rossig, S. (2014). Does Supplemental Instruction for Principles of Economics improve outcomes for traditionally underrepresented minorities? *International Review of Economics Education*, *17*, 98–108. https://doi.org/10.1016/j.iree.2014.08.005

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, *7*(1), 16. https://doi.org/10.21037/atm.2018.12.10

Zhou, J., & Castellanos, M. (2013). *Examining the influence of campus climate on students' time to degree: A multilevel discrete-time survival analysis*. Paper presented at the annual meeting of the Association for the Study of Higher Education.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D.H. Schunk, (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 1–37). Lawrence Erlbaum Associates.

Appendix.

Descriptive statistics for all variables by semester.

	Fall 2017 SI Coursework		Spring 2018 SI Coursework	
Variable	Mean	SD	Mean	SD
Course grade	2.75	0.92	2.76	0.94
DFWI	0.13	0.34	0.16	0.36
Retention to Fall 2018	0.88	0.33	0.92	0.27
Retention to Fall 2019	0.80	0.40	0.83	0.38
SI participation (any vs. none)	0.26	0.44	0.22	0.41
SI participation (4 categories)	1.48	0.92	1.44	0.93
Female	0.59	0.49	0.60	0.49
Black	0.03	0.17	0.03	0.17
Latinx	0.08	0.28	0.08	0.28
Asian	0.05	0.22	0.05	0.23
Other race	0.01	0.11	0.01	0.11
First-generation	0.26	0.44	0.26	0.44
Subjective SES	3.37	0.87	3.36	0.87
Class standing	1.42	0.61	1.68	0.70
HS advanced math	1.59	0.81	1.59	0.81
HS biology courses	1.26	0.48	1.28	0.50
HS chemistry courses	1.16	0.41	1.18	0.43
HS physics courses	0.74	0.52	0.75	0.52
HSGPA	3.73	0.41	3.74	0.41
ACT/SAT score	25.76	3.89	25.81	3.97
Perseverance of effort	4.10	0.72	4.12	.71
Consistency of interest	3.17	0.78	3.20	.79
Expected GPA	7.35	0.84	7.39	.85
Academic behaviors	5.40	0.87	5.58	.86
Missed classes	1.62	0.76	1.60	.76
Attended office hours	0.71	1.05	0.71	1.05
Declared major	5.54	1.94	5.58	1.93
Work on-campus	1.38	1.05	1.38	1.06
Work on (squared)	3.01	6.22	3.03	6.40
Work off-campus	1.37	1.12	1.38	1.15
Work off (squared)	3.14	7.25	3.24	7.58
Co-curriculars	2.18	1.11	2.16	1.10
Co-curriculars (squared)	5.99	6.63	5.85	6.64
Time socializing	2.45	0.92	2.44	0.93
Intent to persist	6.32	1.04	6.35	1.00
College belonging	5.66	1.08	5.65	1.09
Homesickness	2.31	1.15	2.26	1.12
Financial stress	5.78	1.50	5.79	1.50
Mental health	1.68	0.74	1.67	0.73
Adverse life event	0.17	0.49	0.15	0.47
Enrolled in biology course	0.12	0.32	0.15	0.36
Enrolled in health course	0.06	0.23	0.12	0.33
Enrolled in math course	0.34	0.47	0.24	0.43
Enrolled in psychology course	0.10	0.30	0.09	0.29