Tech Know Learn (2017) 22:335–352
DOI 10.1007/s10758-017-9317-0

ORIGINAL RESEARCH

Balancing Student Success: Assessing Supplemental Instruction Through Coarsened Exact Matching

Maureen A. Guarcello[1] · Richard A. Levine[2] · Joshua Beemer[3] · James P. Frazee[1] · Mark A. Laumakis[4] · Stephen A. Schellenberg[5]

Published online: 11 July 2017

Abstract Supplemental Instruction (SI) is a voluntary, non-remedial, peer-facilitated, course-specific intervention that has been widely demonstrated to increase student success, yet concerns persist regarding the biasing effects of disproportionate participation by already higher-performing students. With a focus on maintaining access for all students, a large, public university in the Western United States used student demographic, performance, and SI participation data to evaluate the intervention's efficacy while reducing selection bias. This analysis was conducted in the first year of SI implementation within a traditionally high-challenge introductory psychology course. Findings indicate a statistically significant relationship between student participation in SI and increased odds of successful course completion. Furthermore, the application of Coarsened Exact Matching

✉ Maureen A. Guarcello
mguarcello@mail.sdsu.edu

Richard A. Levine
rlevine@mail.sdsu.edu

Joshua Beemer
joshbeemer@hotmail.com

James P. Frazee
jfrazee@mail.sdsu.edu

Mark A. Laumakis
mlaumakis@mail.sdsu.edu

Stephen A. Schellenberg
saschellenberg@mail.sdsu.edu

[1] Instructional Technology Services, San Diego State University, San Diego, CA, USA

[2] Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA

[3] Computational Science Research Center, San Diego State University, San Diego, CA, USA

[4] Department of Psychology, San Diego State University, San Diego, CA, USA

[5] Division of Undergraduate Studies, San Diego State University, San Diego, CA, USA

reduced concerns that increased course performance was attributed to an over-representation of higher performing students who elected to attend SI Sessions.

Keywords Learning analytics · High impact practice · Program assessment · Propensity score matching

1 Introduction

Supplemental Instruction (SI) is a well-developed model for voluntary peer-assisted group learning sessions intended to increase student success within high-challenge courses. The model was developed by the University of Missouri-Kansas City (UMKC) in 1973 and has been implemented by over 1500 universities worldwide (Martin and Arendale 1993; Dawson et al. 2014). The SI practitioner community uses various approaches to assess the effects of SI participation upon student performance. For example, the UMKC Supplemental Instruction Supervisor Manual (International Center for Supplemental Instruction 2014) outlines program evaluation instructions by way of tabulating the total number of SI participants, and comparing differences between the mean course grade point averages of SI and non-SI attending populations. Additional assessments include analyses of covariance and mixed and qualitative methods, with many of these measuring pass/fail rates at the course level (Arendale 1997; Dawson et al. 2014; Fayowski and MacMillan 2008).

In their comprehensive review of SI literature from 2001 to 2010, Dawson et al. (2014) concluded that greater retention, higher mean exam grades, and lower course failure were largely attributable to SI and that the approach produced no negative consequences. However, these authors also stressed that none of these outcomes were "supported by a gold standard study involving random assignment to groups and sufficient detail about methodology, participants, and the SI intervention in practice" (pp. 26–27). This lack of random assignment (e.g., randomized controlled trials) in the SI program assessment literature is not unexpected, because voluntary student participation is a key tenet of the SI model (International Center for Supplemental Instruction 2014). Such self-selection may bias commonly practiced analyses towards positive intervention effects (e.g., already higher performing students disproportionately attending SI Sessions which could have, in and of themselves, no significant positive effects).

SI was recently piloted at a large, public university in the Western United States as a potential means to improve student success within one historically high-challenge general-education course in introductory psychology. Given the relative high cost of the SI model, a robust quantitative assessment using a diverse suite of data was deemed essential to determine the pilot's efficacy and return on investment in terms of increased student success. This paper presents the multi-disciplinary strategy used to evaluate this pilot SI program. We began with some of the traditional statistical analyses commonly used within the SI literature and training materials. Some of these methods (specifically those that do not address covariate imbalance) have been deemed insufficient to assess biases from student self-selection into the voluntary SI treatment (McCarthy et al. 1997; Dawson et al. 2014). To address this concern, we identified and applied a relatively new analytical approach, termed Coarsened Exact Matching (CEM), to increase covariate balance and thereby decrease selection bias (Blackwell et al. 2009). To our knowledge, this is the first application of CEM to assess SI, and the goal of this paper is to present this approach and

our findings for consideration by the SI and broader educational effectiveness communities.

2 Overview of the Pilot Study

2.1 Course Design

This pilot study was based on two Spring 2016 semester sections of an introductory psychology course taught by the same instructor, with Tuesday/Thursday lectures from 8:00 to 9:15 a.m. (N = 221 students) and from 9:30 to 10:45 a.m. (N = 492 students), over a traditional 15-week semester. The sections were offered in a hybrid format, with Tuesdays in a face-to-face modality and Thursdays in a synchronous (i.e., live) online modality using a video-conferencing tool within the university's learning management system. Students who were unable or chose not to participate in the synchronous online sessions were provided with links to recordings of the sessions. The two sections had a common syllabus and identical structure in terms of textbook, homework assignments, number of exams, and exam items (each with identical questions drawn from two question repositories). The course is required for psychology majors and fulfills a lower-division general education requirement in social and behavioral sciences for other majors.

A variety of high-stakes and low-stakes assessments accounted for a total of 700 possible points for the course. Four "high-stakes" exams, each worth 120 points, were administered in weeks 4, 8, 11, and 15 (480 points total). Students also completed weekly online quizzes, via a publisher-hosted web site, that were worth a total of 120 points. Finally, students used an audience response system (i.e., clickers) to answer instructor-posed questions during each face-to-face lecture that were worth up to 100 points, with up to 60 points earned for correct responses and up to 40 points earned for lecture attendance. Thus, 31% of the total points for the course could be earned through "low-stakes" online quizzes and face-to-face class participation.

During the six semesters prior to the implementation of the SI pilot, D and F grades (D+, D, D−, and F) ranged from 14 to 33%, with a mean of 22%, and the average course GPA was 2.38 (approximately a C+ letter grade). Notably, the course design and the instructor remained consistent during this time and through the SI pilot. Sometimes poor student performance is attributed to the instructor or course design; this is arguably not the case here: the instructor has been teaching this course for more than a decade, is the recipient of multiple student-nominated teaching awards, and consistently receives exceptional peer and student evaluations. Similarly, the course was carefully designed to meet and exceed both hybrid and traditional pedagogical standards (Laumakis et al. 2009). Based on this confluence of large enrollments, high repeatable grades (C− and below), and strong course design and delivery, the university decided to invest in a focused application and analysis of SI as a means to increase student success in the course.

2.2 Supplemental Instruction

The SI model is highly prescriptive, as is the recruitment and training of the student SI Leaders who conduct the SI Sessions. Potential SI Leaders are required to have earned a B+ or higher in the high-challenge course, have an overall grade point average of 3.0, and be formally recommended by their professor for the course. SI Leader training is required,

and includes effective practices for managing active learning environments (e.g., redirection of questions, increased wait time, frequent checks for understanding) and a diverse collection of active learning strategies applicable to a broad range of disciplines (e.g., think-pair-share, incomplete outlines, learning games, exam question prediction). Following this training, SI Leaders attend the high-challenge course again, serve as model students for the enrolled students, and lead regularly-scheduled, weekly 90-min SI Sessions available to all students enrolled in the course.

SI Sessions are promoted and administered as non-remedial opportunities for students to increase their ownership of, and expectations for, their own learning, with session participants commonly ranging from already high-achieving to those who are struggling in the course. While course instructors and teaching assistants are encouraged to identify the SI Leaders in lecture and to actively promote their SI Sessions as a resource for all students, student participation in SI Sessions is voluntary and not shared with the course instructors or teaching assistants (International Center for Supplemental Instruction 2014; Martin and Arendale 1993). In this pilot study, nine SI Leaders were recruited and trained for the introductory psychology course and collectively offered 18 regularly-scheduled weekly SI Sessions during the 15-week semester. For each SI Session, participation was recorded for subsequent analysis by swiping student identification cards through a magnetic stripe reader; this participation information was not shared with the instructor during the course.

2.3 Traditional Assessment of Supplemental Instruction Participation

For all students in the course, demographic data was obtained from the university's student information system and course performance data was obtained from the learning management system (see Tables 1, 2, respectively). These student demographic and course performance data were merged with student participation data for all SI Sessions into a single data matrix. Per the SI model, these data were updated throughout the semester to assess differences between SI-attending and non-SI-attending student performance. These aggregated data were routinely reported to the faculty member (see Table 2, Columns 3–5), who shared them with the entire class to demonstrate the potential for SI participation to help increase course mastery as measured in exam scores.

Of the 713 students enrolled in the psychology class, 305 (43%) students participated in at least one SI Session during the semester, with a maximum number of sessions attended by a student being 20 (Fig. 1). Six SI-attending students lacked sufficient demographic data and were removed from the subsequent analyses. Comparisons of the SI-attending and non-SI-attending populations revealed statistically significant differences in SI usage with respect to some categorical demographic variables (e.g., gender, on-campus residence, and student level) and some continuous variables (e.g., grade point average), with the SI-attending population outperforming the non-SI-attending population on exams throughout the course.

Given that the SI model is based on voluntary participation, studies of SI effectiveness have been necessarily observational or quasi-experimental in nature (i.e., students are not randomly assigned to a treatment). The challenge is, without random assignment, such studies subsume an unknown amount of self-selection bias into the determined effect of the treatment. For example, in this study, students displaying certain demographic characteristics or academic preparation may be overrepresented in the SI-attending population (see above and Table 1). This self-selection problem has a long history, with researchers arguing that standard statistical adjustments and approaches are largely inadequate for

Table 1 Demographic comparison of all students, SI-attending students, and pre/post-CEM non-SI-attending students

Variable	All students ($N = 1145$)	Pre-CEM unmatched			Post-CEM matched			Pre/post-CEM comparison		
		SI-attending (N = 299)	Non-SI-attending (N = 846)	p value	SI-attending (N = 299)	Non-SI-attending (N = 299)	p value	Pre-CEM unmatched difference	Post-CEM matched difference	Increase in difference
Categorical variables (percentiles)										
Female[a]	63%	*71%*	*60%*	*0.002**	71%	70%	0.333	11%	1%	
Male	37%	*29%*	*40%*		29%	30%		11%	1%	
Enrolled in honors college	5%	5%	5%	0.899	5%	6%	0.720	0%	1%	Y
Registered with disabled student services	1%	2%	1%	0.259	2%	2%	1.000	1%	0%	
Pell-eligible (high financial need)	31%	34%	29%	0.129	34%	32%	0.664	5%	2%	
Lived in residence halls	69%	*79%*	*66%*	*0.000**	79%	78%	0.238	13%	1%	
First generation, no parent college	16%	15%	16%	0.745	15%	16%	0.655	1%	1%	
First generation, some parent college	36%	33%	37%	0.376	33%	34%	0.441	4%	1%	
Admitted via compact scholars program	14%	*10%*	*16%*	*0.016**	10%	10%	0.599	6%	0%	
Pre-major status	75%	76%	74%	0.539	76%	76%	0.923	2%	0%	

Table 1 continued

Variable	All students ($N = 1145$)	Pre-CEM unmatched			Post-CEM matched			Pre/post-CEM comparison		
		SI-attending (N = 299)	Non-SI-attending (N = 846)	p value	SI-attending (N = 299)	Non-SI-attending (N = 299)	p value	Pre-CEM unmatched difference	Post-CEM matched difference	Increase in difference
Student level										
Freshmen[a]	51%	*57%*	*49%*	*0.013**	57%	49%	0.355	8%	8%	
Sophomore	36%	*36%*	*36%*		36%	43%		0%	7%	Y
Junior	8%	*5%*	*10%*		5%	6%		5%	1%	
Senior	4%	*3%*	*5%*		3%	3%		2%	0%	
College										
Arts and letters	5%	*5%*	*5%*	*0.005**	5%	6%	0.989	0%	1%	Y
Business[a]	8%	*7%*	*8%*		7%	9%		1%	2%	Y
Education	3%	*3%*	*3%*		3%	3%		0%	0%	
Engineering	7%	*2%*	*8%*		2%	2%		6%	0%	
Health and human services	32%	*37%*	*30%*		37%	38%		7%	1%	
Professional studies and fine arts	12%	*14%*	*12%*		14%	15%		2%	1%	
Sciences	23%	*22%*	*23%*		22%	21%		1%	1%	
Undergraduate studies	12%	*9%*	*12%*		9%	7%		3%	2%	

Table 1 continued

Variable	All students ($N = 1145$)	Pre-CEM unmatched			Post-CEM matched			Pre/post-CEM comparison		
		SI-attending (N = 299)	Non-SI-attending (N = 846)	p value	SI-attending (N = 299)	Non-SI-attending (N = 299)	p value	Pre-CEM unmatched difference	Post-CEM matched difference	Increase in difference
Ethnicity										
African American[a]	4%	*5%*	*3%*	*0.043**	5%	5%	0.989	2%	0%	
Asian	5%	*5%*	*5%*		5%	5%		0%	0%	
Filipino	10%	*8%*	*11%*		8%	9%		3%	1%	
International	4%	*4%*	*4%*		4%	4%		0%	0%	
Mexican–American	23%	*22%*	*23%*		22%	22%		1%	0%	
Multiple ethnicities, non-Hispanic	8%	*6%*	*8%*		6%	6%		2%	0%	
Native American	1%	*1%*	*0%*		1%	0%		1%	1%	
Other Hispanic, Latino	5%	*8%*	*4%*		8%	6%		4%	2%	
Other, not stated	2%	*2%*	*2%*		2%	2%		0%	0%	
Pacific Islander, Native Hawaiian	1%	*0%*	*1%*		0%	0%		1%	0%	
SE Asian	3%	*3%*	*3%*		3%	5%		0%	2%	Y
White	35%	*35%*	*35%*		35%	35%		0%	0%	

Table 1 continued

Variable	All students ($N = 1145$)	Pre-CEM unmatched			Post-CEM matched			Pre/post-CEM comparison		
		SI-attending (N = 299)	Non-SI-attending (N = 846)	p value	SI-attending (N = 299)	Non-SI-attending (N = 299)	p value	Pre-CEM unmatched difference	Post-CEM matched difference	Increase in difference
Proficiency at university matriculation										
Proficient in English and Math[a]	91%	89%	92%	0.262	89%	90%	0.617	3%	1%	
Remedial in English; Proficient in Math	2%	2%	2%		2%	2%		0%	0%	
Proficient in English; Remedial in Math	5%	5%	4%		5%	6%		1%	1%	
Remedial in English and Math	2%	3%	1%		3%	2%		2%	1%	
Proficiency at high school graduation										
Proficient in English and Math[a]	88%	*82%*	*90%*	*0.006**	82%	86%	0.753	8%	4%	
Remedial in English; Proficient in Math	4%	*6%*	*3%*		6%	4%		3%	2%	
Proficient in English; Remedial in Math	5%	*7%*	*4%*		7%	6%		3%	1%	
Remedial in English and Math	3%	*5%*	*3%*		5%	4%		2%	1%	

Table 1 continued

Variable	All students ($N = 1145$)	Pre-CEM unmatched			Post-CEM matched			Pre/post-CEM comparison		
		SI-attending (N = 299)	Non-SI-attending (N = 846)	p value	SI-attending (N = 299)	Non-SI-attending (N = 299)	p value	Pre-CEM unmatched difference	Post-CEM matched difference	Increase in difference
Continuous variables (mean ± standard deviation)										
Units enrolled for term	15.0 ± 1.9	15.2 ± 1.9	15.0 ± 1.9	0.119	15.2	15.0	0.472	0.2	0.1	
Total GPA (includes transfer units)	2.5 ± 1.2	*3.0 ± 0.6*	*2.3 ± 1.3*	*0.000**	3.0	3.1	0.550	0.7	0.0	
Campus GPA	2.4 ± 1.3	*3.0 ± 0.6*	*2.2 ± 1.4*	*0.000*	3.0	3.1	0.670	0.8	−0.1	
Age	19.4 ± 1.9	19.4 ± 2.2	19.4 ± 1.8	0.949	19.4	19.4	0.695	0.0	0.0	
SAT composite score	1121 ± 192	1117 ± 140	1123 ± 207	0.574	1117	1129	0.634	−6.0	−12.0	Y
High school graduation year	2014 ± 2.0	2014 ± 2.1	2014 ± 1.7	0.087	2014	2014	0.634	0.0	0.1	
High school GPA	3.7 ± 0.5	*3.7 ± 0.3*	*3.6 ± 0.6*	*0.000**	3.7	3.7	0.866	0.1	0.0	
Transfer GPA	3.3 ± 0.6 (n = 184)	3.4 ± 0.6 (n = 49)	3.3 ± 0.6 (n = 135)	0.824	3.4	3.4	0.327	0.1	0.0	
Transfer units accepted	16.6 ± 15.6 (n = 746)	15.0 ± 12.9 (n = 196)	17.1 ± 16.5 (n = 550)	0.160	19.4	18.9	0.781	−2.1	0.5	
Average propensity score for population	–	0.43	0.20	–	0.43	0.39	—	0.23	0.04	

* Significant at 0.05

[a] MatchIt analysis defined this condition as the comparison and did not report it; for completeness, the value has been calculated from other conditions and is reported here

Table 2 Course component comparison of all students, SI-attending students, and pre/post-CEM non-SI-attending students

Course component	All students ($N = 1145$)	Pre-CEM unmatched			Pre-CEM matched		
		SI-attending (N = 299) (%)	Non-SI-attending (N = 846) (%)	p value	SI-attending (N = 299) (%)	Non-SI-attending (N = 299) (%)	p value
Exam 1	71 ± 15%	71 ± 14	71 ± 15	0.563	71 ± 14	73 ± 16	0.093
Exam 2	74 ± 17%	76 ± 16	74 ± 17	0.012*	76 ± 16	74 ± 96	0.192
Exam 3	76 ± 16%	78 ± 13	75 ± 17	0.006*	78 ± 13	78 ± 16	0.883
Exam 4	77 ± 18%	81 ± 11	76 ± 20	0.000*	81 ± 11	76 ± 21	0.003*
Test Total	75 ± 14%	76 ± 11	74 ± 15	0.002*	76 ± 11	75 ± 16	0.338
Online quizzes	84 ± 2%	87 ± 20	87 ± 20	0.000*	87 ± 20	86 ± 20	0.228
Clicker points	95 ± 22%	96 ± 14	95 ± 24	0.532	96 ± 14	94 ± 21	0.291
Final course percent	0.79 ± 0.14	81 ± 9	78 ± 15	0.000*	81 ± 9	79 ± 16	0.029*

* Significant at α-level of 0.05

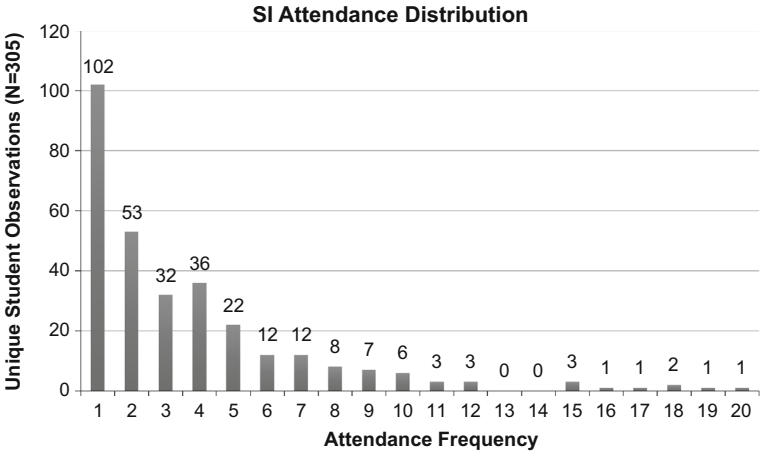

Fig. 1 Frequency of SI Session attendance by 305 students in the Spring 2016 introductory psychology course (e.g., 7 of the 305 students attended 9 SI Sessions during the semester)

identifying and addressing such biases (Rosenbaum and Rubin 1984; Fayowski and MacMillan 2008).

2.4 Coarsened Exact Matching

A recent method, Coarsened Exact Matching (CEM), designed and pioneered by Gary King and colleagues (Iacus et al. 2012; Ho et al. 2011), essentially mimics a randomized treatment assignment, producing "treatment" and "control" groups after the treatment has

been administered. While some observations may be dropped in the process (detailed below), CEM reduces covariate imbalance for the subsequent determination of a treatment effect, which was our main objective in conducting this research. CEM has not been applied in SI studies, but has been used in epidemiology and other disciplines for studies that make voluntary treatment available to all participants (Iacus et al. 2012; Stevens et al. 2010). In short, CEM afforded us the ability to produce *ex post* matching of students who self-selected into the SI-attending population (N = 299) with a covariate-based "equivalent" student population drawn from a much larger non-SI-attending population (N = 846). We found that the Spring 2016 non-SI-attending pool did not provide sufficient matches with SI-attending students, so the N = 846 includes non-SI-attending students who were enrolled in the Fall 2015 introductory psychology course.

The complete theoretical and operational aspects of CEM are beyond the scope of this paper; we refer interested readers to the detailed methodological treatments of CEM (i.e., Blackwell et al. 2009; Keller and Tipton 2016; King and Nielsen 2016). Below we apply CEM to the SI pilot data set as a means to maximize covariate balance and minimize self-selection biases, followed by a logistic regression analysis of the resulting matched populations to test the hypothesis that student participation in SI increases their odds of receiving a passing grade in the course.

3 Methods

3.1 MatchIt

The MatchIt package in the R statistical programming environment (Ho et al. 2011; R Core Team 2016) was used to conduct the CEM. MatchIt provides an automated implementation of the CEM as described by Iacus et al. (2009), and produces a standard output including statistics, diagnostic plots, and summaries. Below we outline, and provide commentary on, the application of the MatchIt CEM analysis to the introductory psychology course dataset.

The MatchIt CEM analysis begins with a basic diagnosis of covariate balance between the SI-attending and non-SI-attending populations and reports means for each covariate in each population along with their difference (see Table 1). The next step in the CEM is to "coarsen" the covariates, and then create groups of similar individuals with respect to those covariates. For example, the coarsened variables could categorize age groups as <18, 18, 19, 20, 21, 22–24, and >25 years, or combine the twelve ethnicity categories into a smaller number of groups containing multiple ethnicities.

3.2 Research Design

The MatchIt package is flexible, with options to search for exact matches between treated and control subjects, where a failure to match will remove the treated subject from the treatment population. In order to preserve our initial SI-attending population through the CEM process, we opted for a different approach. Data were coarsened using CEM, then we employed propensity-score-based matching; in this case, the propensity score is the probability that a student would participate in the SI Session "treatment." This method produced a non-SI-attendee match for each of the SI attendees.

The CEM approach identifies control subjects that are closest to the treated subjects over the covariate set (Blackwell et al. 2009, Section 4.7; Iacus et al. 2012, Section 3.1).

Iacus et al. (2011) show that such methods maintain, in their words, "a surprisingly large number of attractive statistical properties" (p. 345) for causal inference. Furthermore, by balancing covariates across treatment and control groups, we are in essence mimicking a randomized treatment assignment and can thereby minimize biases in estimating the treatment effect.

Following the production of matched pairs of SI-attending and non-SI-attending students through the above CEM process, we performed a logistic regression analysis on the matched pairs to assess the impact of SI attendance on students' odds of passing the introductory psychology course (using final course grade point averages).

4 Results

Mean values for each covariate for the pre-CEM unmatched SI-attending and non-SI-attending populations as well as the post-CEM matched SI-attending and non-SI-attending populations are reported in Table 1. Comparison of these pre-CEM unmatched and post-CEM matched populations shows that the mean difference for 45 of the 51 covariates marginally to markedly decreased in the matched populations. The mean difference increased slightly in six of the covariates.

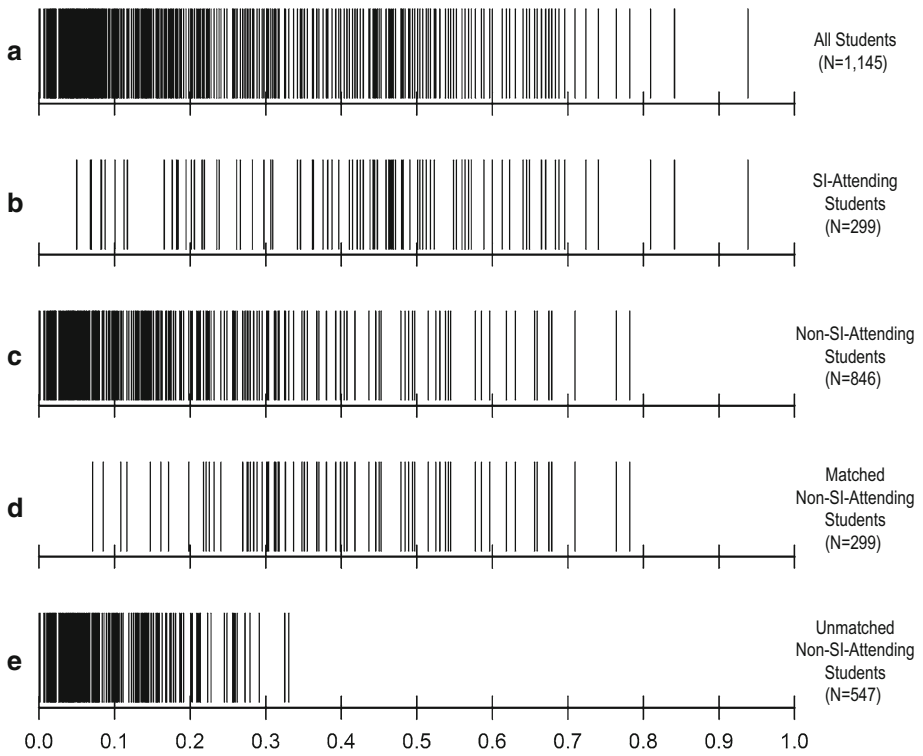

Fig. 2 Individual propensity score values, represented as vertical lines, for (**a**) all students (N = 1145), (**b**) the SI-attending population (N = 299), (**c**) the non-SI-attending population (N = 846), (**d**) the matched non-SI-attending population (N = 299), and (**e**) unmatched non-SI-attending students (N = 547)

Propensity score distributions for all students in the dataset are presented in Fig. 2, with individual propensity score values represented as vertical lines, for (a) all students (N = 1145), (b) the SI-attending population (N = 299), and (c) the non-SI-attending population (N = 846). The CEM paired each SI-attending student with a non-SI-attending student, and this matching is qualitatively evident in the similar distributions of propensity scores for the (b) SI-attending population and (d) matched non-SI-attending populations (N = 299). In contrast, the propensity scores for (e) unmatched non-SI-attending students (N = 547) have a right-skewed distribution.

Figure 3 presents a cross-plot of total course scores (as a percentage) between the matched SI-attending and non-SI-attending populations. Data above the diagonal no-difference line represent matched pairs where SI-attending students performed better than their non-SI-attending partners. Data to the left of 70% mark, represent where non-SI-attending students received a D+ or lower (repeatable grades) while their SI-attending partners performed better. Data in the upper-right corner represent high performing matched pairs. In these cases, the non-SI-attending student was successful in the course without the addition of SI support, but the SI-attending student may have benefitted from SI.

Table 2 presents the differences in students' course performance outcomes pre- and post-CEM. The focus of this research was to examine and reduce self-selection bias when measuring the efficacy of SI. In the Pre-CEM analysis, the SI-attending student group outperformed non-SI-attending students on three of the four exams, and ultimately

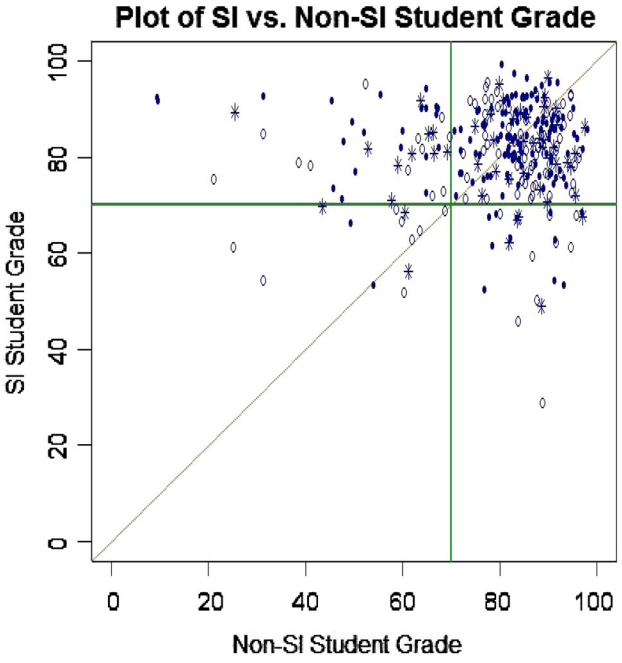

Fig. 3 Cross-plot of final course percentage for matched pairs of SI-attending and non-SI-attending students. Horizontal and vertical lines represent passing and non-passing boundaries, and the diagonal line represents no difference. Open circles represent those students who attended one SI Session; asterisks represent those students who attended two SI Sessions; solid circles represent those students who attended three or more SI Sessions

received a higher final course score (Table 2). It should be noted, while SI-attending student scores remain higher before and after CEM, post-CEM performance averages are significant only in the case of Exam 4 and the final course percentage.

In an effort to determine the impact of SI participation on course performance, specifically the passing of the course, we applied a logistic regression model on final course percentages amongst the matched students, controlling for student demographics (i.e., differences in ethnicity, grade point average, etc.; Table 1). We discovered the odds of passing the course (i.e., final course grade of C or better) for students who attended at least one SI Session were 2.2 times higher than those who did not attend any SI Sessions (n = 299; p value = 0.006; 95% CI of 1.3–3.8). Furthermore, students who attended two or more SI Sessions were 2.8 times more likely to pass the course than those who did not (n = 196; p value = 0.03; 95% CI of 1.2–6.9).

5 Discussion

Over the past forty years, SI research has worked to measure and demonstrate the program's efficacy across disciplines and institutions. For those institutions who fully adhere to the University of Missouri-Kansas City's model, SI Session attendance is voluntary and sessions are available to all students within targeted high-challenge courses. While conventional assessment of SI outcomes have been arguably sufficient (Arendale 1997; McCarthy et al. 1997), the advent of large quantities of performance and demographic data, together with new tools to analyze those data, have made it possible to address some longstanding questions about the efficacy of SI upon student performance.

This study demonstrates the potential for SI programs to increase covariate balance and decrease selection bias when evaluating the overall impact of SI participation upon students' performance, thus providing a more accurate characterization of the treatment. By analyzing the introductory psychology class in this way, outcomes indicated that covariate imbalance was indeed producing an inaccurate overall measure of the SI treatment. However, we discovered that by reducing selection bias, SI attendees still outperformed their non-SI-attending counterparts (Table 2).

Investigating this SI pilot using traditional SI methods, the CEM process, and statistical modeling, enabled us to approach the "gold standard study" that Dawson et al. (2014) found lacking in their comprehensive review of the SI literature. This research opens the door for the strategic use of data with multi-disciplinary cross-campus partnerships such as the ones formed through this study, in an effort to support student success through programming and acute assessment of those offerings. However, these discoveries come with some limitations and questions, discussed below, surrounding the potential for data use in student success initiatives.

5.1 Evolution of Supplemental Instruction Assessment

Twenty years have passed since McCarthy et al. (1997) raised concerns about the sophistication of SI program assessments and the model's direct positive impact on student performance. However, limited time and resources can curtail the extent and complexity of program assessment as efforts focus on more immediate logistical operations such as student training, session scheduling, and faculty collaboration. Yet, the paucity of time and resources are why it is critical to accurately evaluate the effect of SI within and among

courses. Perhaps most importantly, such information provides the means to identify opportunities for program improvement and research to inform further optimization of student success strategies. These efforts can raise awareness, increase the university's return on investment, and inform broader decisions regarding allocation of limited institutional resources. Moving forward, incorporating qualitative research, specifically triangulation and embedded design (Creswell et al. 2003), would inform the more nuanced outcomes that the SI treatment generates.

5.2 Coarsened Exact Matching

Although logistic regression, ANCOVA, and propensity score matching have been used as a means to reduce selection bias within SI and other voluntary student success programs (Stock et al. 2013; Dawson et al. 2014), CEM is not currently among the *ex ante* methods employed in this SI literature. It was only after the application of CEM, which controlled for the multiple variables present in the data, that we were able to confidently infer SI was a contributing factor in the higher performance of those students who participated, versus attributing higher scores to the overrepresentation of already high achieving students (Table 2).

5.3 Learning Analytics and Student Success Alliances

Conducting this SI analysis required partnerships and resources from the administration, academic deans, departments, faculty, instructional technology services, and institutional research. Each of these units understands the value of data to support and improve student success. Although SI does not require tremendous effort on the part of the faculty partners, they must regularly advocate on behalf of, and endorse the program in order to make students aware of its availability and utility. If faculty, chairs, deans, and upper-level administration are not presented with evidence of program effectiveness, then it is unreasonable to request their partnership and advocacy. Providing campus stakeholders with regular reports to share aggregate student exam scores, comparisons between SI-attending and non-SI-attending students, and SI Session attendance frequencies enables them to see the impact of the program even though they do not know which students are participating. Oftentimes, the faculty shared these data with students in their courses as a means to promote the value of participating in SI.

6 Limitations and Considerations for Future Research

Employing CEM in conjunction with conventional statistical analyses has provided this institution with an empirical response to questions about the SI program's effectiveness, specifically with respect to student demographics and course performance. These outcomes will help inform the university administration as they evaluate the effectiveness of multiple programs and allocate often limited resources for student success programming. Although the literature indicates that SI is a proven active-learning strategy that increases student exam scores and final grades, we are now able to demonstrate that it is effective in this local context, and to continue with more granular measurements of its effectiveness in other courses.

There are some limitations to acknowledge as we move forward with similar analyses in other high-challenge courses where SI is offered. Although CEM is a powerful instrument, and the MatchIt package is freely available through the R programming environment, the method is not a replacement for the outcomes which may result from a randomized controlled trial. The automated nature of using the MatchIt package leads to additional questions about the algorithm and how it operates. By electing to automatically, rather than manually, coarsen the data, we recognize that we left those decisions up to the MatchIt algorithm. However, this is not so much a limitation, as it is an acknowledgement that large quantities of data now afford us the opportunity to conduct deeper analysis with more accuracy using powerful data processing software, and to carefully investigate how those programs operate.

Although covariate balance was increased through our approach, non-SI-attending students from the Fall 2015 semester were included in the CEM analysis in an effort to provide a larger pool of potential matches for those 299 SI-attending students from the Spring 2016 semester. As such, we encounter the following potential qualifier: the majority of students who attended this course were first-time freshmen, and those who attended the class in the Fall could perhaps be less prepared (study skills, campus navigation, etc.) when compared to the freshmen who attended in the Spring. There were 77 non-SI-attending students from Fall 2015 who were matched with SI-attending students in the Spring 2016 semester; 25 of those students were freshman. Though we match on the covariate set, unmeasured first-semester characteristics may in some way affect the overall outcomes. As we continue to offer SI, we will be able to increase our training data and pair similar semesters' data to control for differences between the Fall and Spring semesters.

There are a number of potential barriers to the voluntary, non-remedial nature of the SI program. University faculty and administrators ask, if Supplemental Instruction works, why not require everyone to participate? One consideration perhaps, is that the voluntary nature of SI is a contributing factor to the program's effectiveness. This condition comes with a number of circumstances that make program evaluation challenging. For example, we cannot know if, when, or how often a student will attend sessions. Furthermore, we do not know if those students who participate will be over- or under-represented in comparison to the broader class population's race, gender, historic academic performance, etc. However, this research and the implementation of CEM as a convention in the evaluation of SI serves as a confident, and formerly unavailable, new approach to exploring these questions.

7 Conclusions

Our assessment of this first-year SI pilot study has revealed the following key points:

- After increasing covariate balance, we determined that higher exam performance by SI-attending students was attributable in part to the SI intervention, which is designed to help students prepare for exams through active learning strategies and peer-facilitated study.
- Selection bias exists among those students who choose to attend SI Sessions, based upon significant differences among covariates (e.g., overall GPA). Although covariate balance was markedly increased (Table 1) using CEM, we still contend with a host of additional unmeasured variables (i.e., social, motivational, behavioral, etc.).

- The frequency and timing of students' SI attendance is an important factor in measuring overall performance and selection bias. In addition to increased exam scores and overall course performance, outcomes from the logistic regression analyses indicate that SI-attending students were two to three times more likely to pass the introductory psychology class than non-SI-attending students. Further investigation of course performance in relation to how often a student attends SI Sessions, and at what point they initiate SI attendance during the semester is forthcoming.
- While SI is an effective treatment in many cases, it obviously does not address all student success challenges (e.g., course design and modality, student preparedness, instructor behaviors). However, SI can provide a structure for conversations about course effectiveness from multiple perspectives and stakeholders (e.g., course faculty, SI Leaders, instructional designers, and administrators).
- The voluntary nature of the SI program presents institutions with benefits and challenges. Students who attend SI Sessions may range from already very successful in the course, to being in danger of failing. This is where deeper analysis of those students whose behavior indicate they may no longer be attending the course (an unauthorized withdrawal), or those whose only SI Session attendance is the night before the last exam will be beneficial to both the institution and to the program.

These conclusions highlight the need for ongoing analysis and optimization of SI operations. Institutional investments in SI programs and their evaluative operations can potentially yield more focused interventions, earlier indicators of student academic distress, and truer evaluation of effectiveness across this and other resource intensive student success programs.

References

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

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: coarsened exact matching in Stata. *The Stata Journal, 9,* 524–546.

Creswell, J. W., Plano Clark, V. L., Gutmann, M., & Hanson, W. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

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

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,* 843–855.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42,* 8.

Iacus, S. M., King, G., & Porro, G. (2009). CEM: Software for coarsened exact matching. *Journal of Statistical Software, 30,* 9.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106,* 345–361.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20,* 1–24.

International Center for Supplemental Instruction. (2014). *Supplemental Instruction supervisor manual*. Kansas City, MO.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics, 41*(3), 326–348.

King, G., & Nielsen, R. (2016). *Why propensity scores should not be used for matching.* Working paper.

Laumakis, M., Graham, C., & Dziuban, C. (2009). The Sloan-C pillars and boundary objects as a framework for evaluating blended learning. *Journal of Asynchronous Learning Networks, 13*(1), 75–87.

Martin, D., & Arendale, D. (1993). *Supplemental instruction: Improving first-year student success in high-risk courses* (2nd ed.). Columbia: National Resource Center for the First Year Experience and Students in Transition, University of South Carolina.

McCarthy, A., Smuts, B., & Cosser, M. (1997). Assessing the effectiveness of supplemental instruction: A critique and a case study. *Studies in Higher Education, 22,* 221–231.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Society, 79,* 516–524.

Stevens, G., King, G., & Shibuya, K. (2010). Deaths from heart failure: using coarsened exact matching to correct cause-of-death statistics. *Population Health Metrics, 8,* 6.

Stock, W. A., Ward, K., Folsom, J., Borrenpohl, T., Mumford, S., Pershin, Z., et al. (2013). Cheap and effective: The impact of student-led recitation classes on learning outcomes in introductory economics. *The Journal of Economic Education, 44*(1), 1–16.