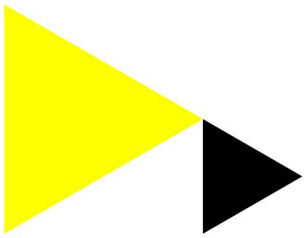
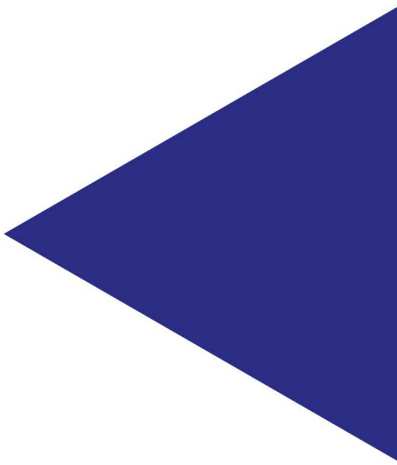
Effects of supplemental instruction on grades, mental well-being, and belonging: A field experiment



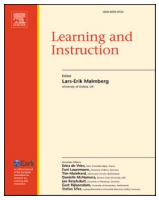


Learning and Instruction 87 (2023) 101805

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Effects of supplemental instruction on grades, mental well-being, and belonging: A field experiment

Izaak Dekker [a,*], Merel Luberti [a,b], Jantien Stam [a]

[a] Center for Applied Research of Education (CARE), Amsterdam University of Applied Sciences, Amsterdam, the Netherlands
[b] Department of Linguistics, University of Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Supplemental instruction
SI-PASS
Field experiment
Academic performance
Mental well-being
Sense of belonging

ABSTRACT

Supplemental Instruction (SI) is a form of structured peer guidance attached to a specific course, provided by an experienced and trained student to a group of students. Previous studies show a positive effect of SI on learning outcomes, some found effects on well-being, and sense of belonging. However, literature on SI lacks randomized controlled trials and does not fully address the risk of self-selection bias. The current study tested whether SI has an effect on grades, mental well-being, and sense of belonging with a pre-registered randomized field experiment and a sample of 493 Dutch first-year students. Students who were offered SI obtained significantly higher grades ($d = 0.26$) but did not score significantly different on mental well-being or belonging.

1. Introduction

Supplemental Instruction (SI)-, also referred to as peer assisted study sessions (PASS)-, consists of peer-led sessions that offer support for historically difficult courses with relatively high drop rates. For SI, experienced peers (the SI-leaders) are trained to design and execute 50–100 min voluntary sessions for a classroom of students on a weekly basis during the school term (Dawson et al., 2014). During these sessions, the SI-leader does not introduce new concepts but rather stimulates discussion and application of concepts that were introduced during the regular classes in order to consolidate the subject matter by means of collaborative learning techniques, practice assessments, and other exercises.

Since its introduction in the 1980s, higher education institutions implemented SI in more than 30 countries and 1500 universities under the guidance of international centers (Martin, 2008). The international center in Missouri Kansas (operates out of UMKC), and national and regional centers in North America, Europe, Africa, and Australasia train supervisors and support and accredit programs to ensure that SI is provided as intended and prevent program drift. There are many different types and forms of near-peer teaching and the different names that can be used to define and distinguish them are not always used consistently (Dawson, 2014). Peer Assisted Learning, for example, is another well-known form of peer-learning in medical and STEM education that is often used for practicums and functions embedded in courses without interaction with the international SI centers (Hermann-Werner et al., 2017). To be called SI, courses should organize substantial training and observations of SI-leaders by a certified SI-supervisor, and use structured lesson plans. The international SI centers also place emphasis on the importance of monitoring the performance of the students who voluntarily show-up compared to those who do not attend. In the 1990s, data thus accumulated from 49 institutions and 1447 courses was used to assess the effectiveness of SI by the U.S. Department of Education (Martin & Arendale, 1993; U.S. Department of Education, 1995). SI was chosen as an exemplary education program because students who attended SI earned higher course grades, succeeded more, and dropped out less than students who did not attend. These evaluations, and various subsequent effectiveness evaluations, controlled for previous academic achievement as a proxy for motivation. Studies such as McCarthy et al. (1997), Kochenour et al. (1997), and Ashwin (2003) criticized the SI literature for risking self-selection bias in relation to motivation, achievement, and ability variables. In order to reliably compare students who did and did not attend, it is necessary to control for all the relevant potential systematic differences between these groups.

During the 2000s, SI spread to other continents, which also induced new effectiveness studies from across the globe (Dawson et al., 2014). In their systematic literature review of SI's effectiveness, Dawson et al.

* Corresponding author. Center for Applied Research of Education (CARE), Amsterdam University of Applied Sciences, Wibautstraat 2-4, 1091 GM, Amsterdam, the Netherlands.
E-mail address: i.dekker@hva.nl (I. Dekker).

https://doi.org/10.1016/j.learninstruc.2023.101805
Received 29 September 2022; Received in revised form 22 March 2023; Accepted 2 July 2023
Available online 7 July 2023

summarized 29 studies published between 2001 and 2010. These supported an overall positive effect of SI on final course grades (seven included studies reported effect sizes, ranging from $d = 0.29$ to $d = 0.60$), passing, and retention. They also found some evidence of a positive effect of SI on academic skills (e.g., Ning & Downing, 2010), general satisfaction or well-being (e.g., Bronstein, 2008), and social relationships (e.g., Dobbie & Joyce, 2008). However, they conclude that the literature fails to convincingly control for self-selection bias because "none [of the findings] is supported by a gold standard study involving random assignment to groups and sufficient detail about methodology, participants, and the SI intervention in practice" (Dawson et al., 2014, p. 635). Recent studies compensated for the risk of self-selection bias with a multivariable analytic approach (Allen et al., 2021), randomized encouragement (Paloyo et al., 2016), or propensity score analyses (Bowman et al., 2021), but acknowledge the persisting risk of self-selection bias in the absence of evidence from randomized controlled trials. Two studies that did use random allocation (Parkinson, 2009; Stanich et al., 2018) to test the effects of SI did not meet the requirements to provide reliable evidence. Parkinson's (2009) controlled experiment was insufficiently powered (only 20 students in the treatment group) and basic statistical information was not reported. The study by Stanich et al. (2018) contained a more sufficient sample size (291 students in the treatment and 110 in the control condition), but combined their SI program with several other interventions. This makes it hard to deduce whether the significant positive effects they found should be attributed to SI or to other interventions (or a combination).

In addition to the problem of self-selection bias, Dawson et al. (2014) conclude that there is a need for more robust research into the effectiveness of SI for social and transferrable outcomes. Although SI lacks a uniformly studied theoretical framework, it is often theorized to enhance performance because it can ensure additional time-on task. This time is spent effectively with the different types of active and collaborative learning that are ingrained in the session plans (International Centre for Supplemental Instruction, 2023). Most programs would additionally claim that the collaborative approach increases students' opportunity to meet other students and develop new friendships. A finding that was supported by studies from Dobbie and Joyce (2008) and Court and Molesworth (2008). Other studies found that SI includes students in a supportive environment and allowed students to discuss difficult material with other students, which increased mental wellbeing and reduced anxiety (Bronstein, 2008).

Findings since then confirmed positive effects of SI on non-academical outcomes. For example, by Stanich et al. (2018), who report positive effects of SI on students' sense of belonging in addition to academic outcomes, or by Hanson et al. (2016), who report that the intervention had a positive effect on liberal arts students' mental well-being.

Sense of belonging was conceptualized by Hurtado and Carter (1997) as the concept which "captures the individual's view of whether he or she feels included in the college community" (p. 327). This perspective differs from the traditional academic integration model in that it shifts the responsibility for integration from the student to the institution (Johnson et al., 2007). Sense of belonging can be improved by positive peer and faculty interactions because these make their a complex context feel more socially supportive (Hoffman et al., 2002; Johnson et al., 2007; Meeuwisse et al., 2010). Meeuwisse et al. (2010) found that interactive learning environments (measured with items such as "how often did you have to work cooperatively in small groups of students in the last year?") led to significantly higher informal and formal peer contact, which in turn led to higher sense of belonging. Meeuwisse et al. (2010) measured sense of belonging with a six item scale that included items such as "I feel at home here".

Well-being can and has been defined in many ways. The two traditional conceptions are subjective well-being, which can be described as feeling good, and psychological well-being, which can be described as optimal functioning (Ryan & Deci, 2001). Using social interdependence theory, Hanson et al. (2016), theorized that peer learning is expected to lead to well-being because interdependence and working together towards mutual goals stimulates cooperativeness and positive relations (Johnson & Johnson, 2009). Positive relations or relatedness is one of the sub-constructs of psychological well-being and one of the important predictors for subjective well-being as well (Ryan & Deci, 2001). Within positive psychology, psychological and subjective well-being were combined in the overarching construct of mental well-being in order to capture the strengths of both approaches and monitor well-being in non-clinical populations (Shah et al., 2021).

1.1. Present study

The present paper addresses the risk of self-selection bias in the literature by conducting a pre-registered randomized field experiment to test the effectiveness of being offered SI on performance measured in grades. Additionally, we address the need for more robust research into the effectiveness of SI on social or transferrable outcomes by testing whether mental well-being and sense of belonging are indeed positively influenced by SI (as stipulated by Hanson et al., 2016 and Stanich et al., 2018). We test the effectiveness with a sufficiently powered sample of students from 10 different courses of study within the educational domain.

We pre-registered the following hypotheses:

1) Students who receive access to SI for a specific course will obtain a higher grade for that course than their peers in the control condition
2) Students who receive access to SI will score higher on mental well-being than their peers in the control condition
3) Students who receive access to SI will score higher on sense of belonging than their peers in the control condition

2. Methods and materials

2.1. Participants

Participants were 493 undergraduate students (255 treatment, 238 control) enrolled in 10 different courses of study, and organized in 42 groups, from the education faculty of a large, public, Dutch university of applied sciences in an urban area at the start of 2022. All 852 first-year students from the participating courses of study were asked to participate in the study after receiving information about the procedures of the experiment, what data would be gathered and stored, and how. The 493 participating students were representative of the population of the education faculty. The majority of the sample (72.8%) was female. The educational background of the students consisted of 64.1% general academic track, 27.2% vocational track, 3.9% preparatory scientific track, and 4.8% used an admission test or an eligible international degree. We report and control for academic background, because previous education here is strongly related to central exam scores (similar to SAT scores) (Van der Zanden et al., 2018). The average age of the participants was 20.57 years old ($SD = 3.21$). Participant characteristics are shown in Table 1.

2.1.1. Final analytical sample

The demographic data contained no missing data for age or previous education, and one missing case for gender. Out of 493 students, 61 students did not participate in the exam and could not be graded (88% response rate). The treatment group did not contain significantly more or less missing grades or response on the survey than the control group (Table 2).

The survey data contained 22% missing cases at T0 (78% response rate) and 35% missing cases at T1 (65% response rate), leading to a final sample of 384 at T0 and 320 at T1.

Table 1
Descriptive comparison between control and treatment group, and baseline inferential balance checks with administrative and survey data.

Characteristic	Control Sample % or mean (*SD*)	Treatment group (*SD*)	χ^2 or t-value (*df*)	*p*-value	*N*
Male*	30%	25%	1.45 (1)	0.23	492
Vocational track*	24%	30%	2.43 (1)	0.12	493
Course of study*	N.a.	N.a.	2.71 (9)	0.98	493
Baseline survey (T0) missing (= 1)*	22%	22%	0.01 (1)	0.93	493
Post survey (T1) missing (= 1)*	38%	33%	1.50 (1)	0.22	493
Grade missing (= 1)*	13%	11%	0.49 (1)	0.49	493
Age	20.54	20.60	0.22 (491)	0.82	493
T0 Mental Well-being	3.44 (0.60)	3.40 (0.62)	−0.63 (382)	0.53	384
T0 Sense of Belonging**	3.92 (0.62)	3.91 (0.55)	0.10 (1)	0.75	384

Note. * = tested by means of χ^2 since variable is categorical. ** = tested with multilevel models. *Df* = degrees of freedom. The analytical strategy used for these tests is described in detail in paragraph 2.6.1.

Table 2
Distribution of participants, missing grades, and amount of groups, over courses of study.

Course of Study	Treatment group *n* (missing grades)	Control group *n* (missing grades)	Total *n* (missing grades)	Group *N* (SI groups)
Biology	11	13 (2)	24 (2)	2 (1)*
Dutch language	4	4	8	1 (1)
Economics	10	6	16	2 (1)*
Elementary school teacher	24	24 (1)	48 (1)	4 (4)
English language	51 (11)	56 (13)	107 (24)	6 (6)
Geometry	10 (2)	10 (1)	20 (3)	2 (1)*
German language	5 (1)	3	8 (1)	1 (1)
History	34 (5)	32 (8)	66 (13)	4 (4)
Pedagogy	91 (9)	78 (6)	169 (15)	19 (19)
Mathematics	15 (1)	12 (1)	27 (2)	1 (1)
Total	255 (29)	238 (32)	493 (61)	42 (39)

Note. Missing grade means that the students participated in the experiment but did not participate in the exam at the end of the term. *The sample contained 42 groups combined into 39 SI groups. Invited students from two different classes within one course were sometimes invited for one SI group.

2.1.2. Power analyses

A sensitivity analysis in G*Power (*t*-test, difference between independent group means, two-tailed test, 0.80 power and an alpha threshold of 0.05) indicated that a sample size of 493 would detect group differences with an effect size of Cohen's $d = 0.25$ (Faul et al., 2009). With a sample size of 320 this would amount to $d = 0.27$. For the multilevel models (*F*-tests, linear multiple regression, fixed model, r^2 increase, using the same power and alpha threshold), a sample of 432 would detect small effect sizes of $f^2 = 0.02$. In multilevel modelling the effective *n* should be adjusted with the following formula (Hox et al., 2018, p. 223):

$$\text{Effective } n = n \,/\, [1 + (\text{mean cluster size} - 1) * \text{rho}] \quad (1)$$

With 42 clusters (the number of groups) with an average *n* of 11 and an intraclass correlation of 0.13, this leads to an effective *n* of 188. A sample of 188 would still be able to detect effect sized of Cohen's $f^2 = 0.04$.

2.2. Procedures

Before conducting the study, the research ethics committee of the university approved the procedures of the trial. The design of the study, sample, hypotheses, and analytical strategy were pre-registered (https://doi.org/10.17605/OSF.IO/96ZP4). Participants were recruited with emails and online in-class visits three weeks before the third quarter (the academic year is organized into four quarters or 'terms' at the university where this study took place). All participants signed online forms declaring their agreement with the setup of the experiment and the management and storage of the data. After all the students enrolled, they completed a baseline survey and were randomly allocated to the treatment or control condition. Students in the treatment condition received an invitation to attend the SI sessions one week before the start of the third quarter. Students in the control group received an email informing them that they were not allowed to attend SI this quarter but would be able to do so in the fourth. After the last session, and before their exams, the students completed the second survey. At the end of the fourth quarter, all participants were debriefed and compensated with either a journal or a chocolate bar.

2.2.1. Intervention

For 7 weeks, all students in the treatment group were offered weekly 100-min peer-led sessions, attached to the supported course unit. The sessions were scheduled in their school calendar in classrooms that were close in proximity (both in terms of location as in time) to their other scheduled classes.

SI is a specific form of peer learning for which the International Center for Supplemental Instruction from the University of Missouri – Kansas City (UMKS) (https://info.umkc.edu/si/) prescribes the following requirements: (i) using model students/peers, called SI leaders, who recently successfully completed the courses for which SI is provided; (ii) the SI leaders have to be extensively trained and observed by a certified SI-supervisor who is trained in one of the official SI centers; (iii) SI leaders are required to prepare sessions by organizing opening, middle, and closing activities, and applying collaborative learning techniques; (iv) SI leaders are required to attend a minimum of 60% regularly scheduled class sessions for the supported unit.

The first three requirements were met in the current study, the fourth was not. It is common for the fourth requirement not to be met in Europe (J. Malm, personal communication, August 26, 2022). The calendar did not permit all SI leaders to attend the scheduled classes for the supported units in real life. In order to commensurate this, SI leaders watched recordings of the lectures, and the teaching staff reviewed the session plans of the SI leaders to check whether they were in line with the content of that week's lecture.

The SI leaders were trained with a Dutch translation and adaptation of the Texas State University SI leader program (Gutierrez et al., 2017). The SI leader training manual and other materials used for the current study can be found at: https://doi.org/10.17605/OSF.IO/ZWH2D. The SI leaders were required to upload their lesson plans at least one week before execution, which allowed the SI supervisors and the lecturers to provide timely feedback. Additionally, SI leaders were also required to register attendance, which was monitored weekly by the principal investigator. Two SI supervisors attended at least one session of every SI leader to provide them with feedback, most of which were done during the second and third week. During the first two weeks, SI leaders could attend six voluntary online supervision sessions in which they could ask questions and exchange experiences.

In line with the findings and advice from Allen et al. (2019), we advertised SI with a series of short, targeted messages delivered to students via multiple media in order to increase session attendance. Posters in the hallways displayed testimonials from attending students about the usefulness and pleasantness of the SI sessions. Televisions in the hallways and on the coffee machines displayed a subtitled video in which students were interviewed about SI. All content was made based on

brainstorm sessions with students to verify whether these messages would be delivered in the right tone of voice.

2.3. Measures

Grades were measured with the exams for the subjects to which SI was attached. Grades range from 0 to 10, and a 5.5 is required for a pass. These grades were obtained through the university student records.

Mental well-being and sense of belonging were measured with a survey.

Mental well-being. We used a validated Dutch translation of the Short Warwick-Edinburgh Mental Well-being Scale (S-WEMWBS) to measure mental well-being (Shah et al., 2021). The S-WEMWBS combines subjective well-being items with psychological well-being items, and is subsequently highly correlated with SWB and PWB and inversely correlated with anxiety and depression scales (Shah et al., 2021). Validation studies confirmed its uni-dimensionality (Anthony et al., 2022), and validity and appropriateness in different European contexts (Koushede et al., 2019). It contains 7 items with question about how often students experienced the described states during the past two weeks. It contains items such as 'I felt relaxed', and 'I felt connected to others'. Students answered the questions on a 5-point scale ranging from "never" to "all the time". The reliability (Cronbach's alpha) of the scale was 0.77 at T0 and 0.75 at T1.

Sense of belonging. We used the scale for sense of belonging that was developed by Meeuwisse et al. (2010). This unidimensional scale contains 6 items such as 'I feel at home here' and 'I try to show up to classes as least as possible' (reversed). Meeuwisse et al. (2010) and Van Herpen et al. (2020) reported Cronbach alpha's ranging between .70 and .84 with other samples of Dutch students. Students answered the questions on a 5-point scale ranging from "never" to "all the time". In this study the reliability (Cronbach's alpha) of the scale was 0.68 at T0 and 0.65 at T1. These reliability scores were somewhat low and indicate that this measure contains relative high standard errors, making it less suitable for applied research.

Participation. At the start of each session, SI leaders noted the attendance of the individual students with a list that included the students who were allowed to attend SI that quarter.

2.4. Datasets

We obtained and merged three datasets for this study. An administrative dataset, extracted from the university student records, provided the course grades, course credits, and demographic data. A second dataset contained the attendance of students during the SI sessions. The third dataset contained the survey responses of the students at the start and at the end of the quarter. We published an anonymized version of the final dataset online: https://10.17605/OSF.IO/EZAT4).

2.5. Analytical strategy

2.5.1. Testing randomization

We performed several baseline tests to estimate whether randomization was successful (Table 1). We used χ^2 tests to infer whether the treatment and control group differed significantly on categorical variables (course of study, previous education, gender) and we performed *t*-tests on interval (age), and ordinal variables (well-being, sense of belonging). When intraclass correlations were significantly different from zero, we used a multilevel regression analysis. This was only required for T0 Sense of belonging (Appendix A-C).

2.5.2. Estimating the treatment effect

The students in this sample were nested in groups, which were nested in courses of study. We tested whether the assumption of independence of observations was violated because of this, by comparing the (−2*loglikelihood) deviance of a model with one (student only) and two levels of variance (students and class). Whenever this led to significant improvement of the model, multilevel analyses were used to correctly estimate standard errors (Hox et al., 2018). Adding classes and courses of study led to significant improvements with grades as dependent variable (Appendix D) but did not lead to improved model fit for well-being (Appendix E) or sense of belonging (Appendix F). Effects of the treatment on grades were, therefore, estimated with multilevel regression analysis in MLwiN (Rasbash et al., 2020). We calculated effect sizes as proportions of explained variance on the student level. Effects on mental well-being and sense of belonging were estimated with one-way independent analyses of variance (ANOVA). After testing the general treatment effect, we tested whether adding the intervention to a model with demographic covariates (age, gender, previous education) still yielded a significant improvement with analyses of covariance (ANCOVA).

3. Results

3.1. Participation

Although 7 sessions were scheduled per course, most students were not able to attend all. Out of a total of 274 scheduled sessions, 16 sessions were cancelled because the SI leader had Covid, 13 were cancelled because of excursions and extra internship weeks, 10 were cancelled because the university campus was closed due to extreme weather conditions. On average, students attended 2.3 sessions ($SD = 2.07$), 29% never attended a single session, 17.6% attended only one session, and 35.7% of the students attended 4 or more sessions (Table 3). The average attendance per course ranged from $M = 1.52$ to $M = 4.50$ attended sessions per student.

3.2. Descriptive statistics

The treatment and control group did not differ significantly from each other at any of the baseline measures, indicating successful randomization (Table 1). Table 4 shows the correlations and descriptive statistics of the dependent variables.

3.3. Treatment effects on grades

Students invited to attend SI sessions obtained significantly higher grades ($p = 0.011$, $r^2 = 0.01$) (Table 5, model 2). When we controlled for demographic covariates (age, gender, previous education) the treatment explained 1.62% of the variance ($r^2 = 0.02$, $p = 0.008$), which equals a Cohen's *d* of 0.26, and a difference of 0.45 grade points (Table 5, model 4, Fig. 1). We could, therefore, reject the null-hypothesis and accept the alternative for hypothesis 1, which predicted that students in the treatment group would obtain higher grades.

3.4. Treatment effects on mental well-being and sense of belonging

Students in the treatment and control group did not differ significantly in terms of mental well-being ($F = 0.62$, $df = 1, 319$, $p = 0.43$), or

Table 3
Session attendance.

Sessions	Students	%	Cumulative n	Cumulative %
0	74	29.0%	74	29.0%
1	45	17.6%	119	46.6%
2	24	9.4%	143	56.0%
3	21	8.2%	164	64.2%
4	40	15.7%	204	79.9%
5	34	13.3%	238	93.2%
6	15	5.9%	253	99.1%
7	2	0.8%	255	100%

Note. total amount of invited students is 255.

Table 4
Correlations and descriptive statistics of the measured variables.

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Grade	–								
2. Well-being T0	−0.45	–							
3. Belonging T0	−0.00	0.32**	–						
4. Well-being T1	−0.40	0.53**	0.23**	–					
5. Belonging T1	0.02	0.17*	0.55**	0.39**	–				
6. Attendance	0.05	−0.04	0.04	0.04	0.32**	–			
7. Age	−0.12*	0.01	−0.04	0.08	0.03	0.11	–		
8. Male	−0.04	−0.00	−0.14**	0.05	−0.07	0.15*	0.05	–	
9. Vocational track	−0.11*	0.08	0.06	−0.01	0.02	−0.07	0.18*	−0.17**	–
M	6.21	3.42	3.92	3.48	3.88	2.31	20.57		
SD	1.92	0.61	0.58	0.54	0.60	2.07	3.21		
n	432	384	384	320	320	255	493	492	493
%								27	27

Note. *$p < 0.05$. **$p < 0.01$; All variables skewness and kurtosis fell within −0.61 to 0.80, which is acceptable to assume normality (Cain et al., 2017).

Table 5
Multilevel regression analyses of treatment effects on grades after one quarter.

Effect	Parameter	Course credits			
		Model 1	Model 2	Model 3	Model 4
Fixed effects					
Intercept	γ_{000}	6.08 (0.24)	5.85 (0.26)	7.74 (0.59)	7.51 (0.59)
Intervention (= 1)	γ_{001}		0.43* (0.17)		0.45** (0.17)
Age	γ_{002}			−0.07** (0.03)	−0.07** (0.03)
Male (= 1)	γ_{003}			−0.33 (0.22)	−0.30 (0.22)
Vocational track (= 1)	γ_{004}			−0.28 (0.20)	−0.32 (0.20)
Random effects					
Course variance	e_{0ijk}	0.42 (0.25)	0.43 (0.26)	0.39 (0.23)	0.40 (0.24)
Group variance	μ_{0jk}	0.04 (0.09)	0.04 (0.09)	0.01 (0.08)	0.01 (0.08)
Student variance	ν_{0k}	3.15 (0.23)	3.11 (0.22)	3.09 (0.22)	3.04 (0.22)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	3.61	3.58	3.49	3.44
% expl. Var. course level			0	0.93	0
% expl. Var. group level			0	75.00	0
% expl. Var. student level			0.95	0.64	1.62
% expl. Var. total			0.83	2.51	1.15
Goodness of fit					
Deviance		1742.17	1735.94	1729.82	1722.95
Model of reference			Model 1	Model 1	Model 3
χ^2 fit improvement (Df)			$\chi^2_{(1)} =$ 6.23	$\chi^2_{(3)} =$ 6.12	$\chi^2_{(1)} =$ 6.87
P-value			$p < 0.05$	$p =$ n.s.	$p < 0.01$

Note. Standard errors are presented in parentheses. All *p* values in this table are two-tailed. Student $n = 432$; Group $n = 42$; Course of study $n = 10$. *$p < 0.05$ **$p < 0.01$.

sense of belonging ($F = 0.54$, $df = 1, 319$, $p = 0.59$) on the post-test. Controlling for age, gender, and previous education, did not impact the results for mental well-being ($F = 0.90$, $df = 4, 319$, $p = 0.46$), or sense of belonging ($F = 0.51$, $df = 4, 319$, $p = 0.73$). We, therefore, failed to reject the null-hypothesis for hypotheses 2 and 3.

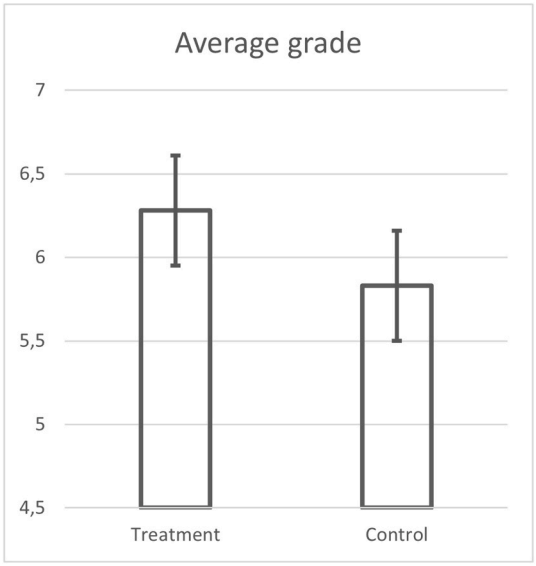


Fig. 1. Differences in average grade between control and treatment group CI 95% after one term of being offered SI while controlling for age, gender, and previous education as covariates.

3.5. Costs-effectiveness of the intervention

The time spend by the SI leaders, the involved coordinators, trainers, project manager, and schedular, times their respective salaries totaled €

Table 6
Costs of organizing and facilitating SI per quarter (Q).

	FTE	Budget Q3
Faculty		
SI Leaders	3.4	42,5
Coordination	1.0	26
Overhead		
Scheduler	0.2	2.9
Catering and communication		4
SI training	0.5	13
Projectmanagement	0.4	10
Total in fte's	5.8	
Total in € (x 1000)		98.4

Note: FTE = full time equivalent. Q3 = Quarter 3 (1/4th year). The following costs per fte per quarter were used for this calculation: SI-leaders = 12,500; Coordinator = 26,000; Scheduler = 14,300; Trainer = 26,000; Project manager = 26,000.

98,400 (Table 6). Divided by the 255 students who were granted access to the SI sessions is € 385.88 per student (Table 6). These costs per student are considered low for an educational intervention (Kraft, 2020). They should reasonably be expected to be lower given that no extra personnel would be required to also invite the control group to the same sessions (attendance per session would average 15 students, which is below the maximum of 20 that Dawson et al., [2014] describe), leading to a spending of € 252 per student. If the costs that could be used for subsequent quarters (training, project management, communication) are spread out among those quarters these costs are reduced to € 154 per student. Finally, it should be noted that the costs for wages (including taxes and social benefits) were relatively high in the context of this study (e.g., € 50,000 per year per full-time SI-leader). For reference, within the Dutch educational system universities receive around € 9,000 per year per student from the government and € 2,143 tuition from the students (€ 7,486 for non-EU citizens).

4. Discussion

This study estimated the effect of offering SI on grades, mental well-being, and sense of belonging of 493 first-year students with a pre-registered randomized field experiment. The results show that students who were invited to participate in SI sessions obtained significantly higher grades but did not score higher on mental well-being or belonging. The outcomes extend the literature on SI in several ways.

First, SI has been extensively studied for over 40 years, but this was the first sufficiently powered evaluation of SI that used random allocation (Allen et al., 2021; Dawson et al., 2014). These results indicate that SI can indeed be expected to improve the learning outcomes of students. The randomized design allowed this study to estimate the effectiveness of *offering* SI instead of *participating* in SI. The effect size (Cohen's d = 0.26) can be considered large because: (i) it includes all the students who never participated (29%), or showed up only once (17.6%). Based on previous findings in the literature (e.g., Paloyo, 2015) it is likely that the effect was driven mostly by the (35.7%) students who participated in >4 sessions; (ii) the costs per student are low (€ 154); (iii) the worldwide spread and accreditation of SI programs makes it feasible that this intervention is scalable; (iii) the experiment was relatively large[1] and pre-registered (Kraft, 2020).

Second, although some previous studies found effects of SI on well-being (Hanson et al., 2016), and sense of belonging (Stanich et al., 2018), we did not find evidence of such effects. The lack of effect on mental well-being could be due to the choice of the specific scale, we used a scale for mental well-being instead of psychological well-being. These scales are highly correlated, but do have a slightly different focus (Tennant et al., 2007) which could partly explain the difference. It could be that the effects found by Hanson et al. (2016) were caused by a confounding variable, given that the design of those studies is prone to self-selection bias. It is possible that we did not find effects on sense of belonging due to the relatively low reliability of the scale. It is also possible that positive effects on well-being or sense of belonging are more likely to occur when SI is offered in the first term of college (Van Herpen et al., 2020). This study offered SI in the third quarter, during which most dropout already occurred and students potentially already feel that they either do or do not belong. A fourth potential explanation might be that effects on well-being are secondary and influenced by increased performance, which could explain why they show up longitudinally but not in this trial. On a positive note, these null-effects could also be interpreted to suggest that SI improvement of learning outcomes did not cause negative side-effects on mental well-being or lower sense of belonging among those who were not attending or invited (Dekker & Meeter, 2022; Zhao, 2017).

4.1. Limitations and future studies

Although most of the quality standards for the execution of the SI program were met, we did not meet the requirement that SI-leaders should attend more than 60% of the lectures to which the SI sessions are attached. We compensated for this with video-recordings and meetings between lecturers and SI leaders. Nevertheless, SI leaders attending lectures potentially could have increased the attendance rates of students and quality of the session further, which, in turn, could have led to larger effects.

Attendance was slightly below the rates reported by Malm et al. (2011): with 7 sessions available, 21% did not attend, 10% attended 1 session, 47% attended 4 or more sessions. However, the participation rates were relatively good compared to those reported in most other studies (Allen et al., 2019). Dancer et al. (2014), for example, reported an average of 1.66 (SD = 3.3) over a 12 week semester, with only 18% of students attending more than 4 sessions. Paloyo et al. (2016) reported an average attendance of 2.67 sessions (SD = 3.98) over a 12 week quarter. In the studies from Bowman et al. (2021) respectively 74% and 78% of the students never attended a session.

All SI leaders in the current study were freshly recruited because there was no SI program in place before the experiment. The SI leaders were properly trained but not as experienced as the average SI leader at a university that has a SI program in place for several years. This could have negatively affected the quality of the sessions. However, this also indicates that SI can become effective directly after its introduction.

Future studies could explore the mechanism behind the effects of SI: what process causes increased performance? This could be studied qualitatively to explore what SI means for students who attend, and how SI might have changed their behaviour within and potentially beyond the attached course. Additionally, future studies could further explore and test the essential elements required for sufficient implementation fidelity. What is the influence of the relationship between the SI leaders and the faculty on the perceived quality of sessions and performance? The observations that are required for accreditation of a SI program could be used to systematically evaluate to which degree different session qualities influence performance. Finally, the field could use a new systematic (if possible meta-analytic) overview of findings that appeared after the scope of the Dawson et al. (2014) review: 2001–2010. Recent studies with advanced statistical approaches such as Allen et al. (2021), Bowman et al. (2021), Paloyo et al. (2016), and the current study with its randomized field experiment design, yielded substantial new insights about the effectiveness of supplemental instruction.

Statement of credit

Izaak Dekker: Conceptualization; Investigation; Formal analysis; Data curation; Funding acquisition; Methodology; Project administration; Writing - original draft; Writing - review & editing. Merel Luberti: Writing – review & editing. Jantien Stam: Investigation.

Funding

This experiment was funded by the National Program for Education from the Dutch Ministry of Education, Culture, and Sciences.

Acknowledgements

We acknowledge Beth Bedinotti, Alessandra Corda, Albert de Voogd, Martijn Koek, Marie-Jose Koerhuis-Pasanisi, Nienke Liekelema, Rosja Mastop, Jet van Dam, Jenny van Dijk, Dani van Keuk, Dick van Straaten, Armin Viergever, Jess Witte, and all the SI leaders for their valuable contributions to the execution of the experiment and Martijn Koek, Joakim Malm, Marieke Thurlings, and Marij Veldman for their feedback

[1] Kraft (2020) describes that RCTs with a sample size <100 report an average effect size of 0.29, compared to 0.16 for samples between 251 and 500, and 0.10 for samples between 501–2000.

on the manuscript.

Appendices.

Appendix A

Table 1
Establishing random part with multilevel analyses for baseline mental well-being.

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	3.48 (0.03)	3.48 (0.03)	3.48 (0.03)
Random effects				
Student variance	e_{0ijk}	0.30 (0.02)	0.29 (0.03)	0.29 (0.13)
Group variance	μ_{0jk}		0.00 (0.01)	0.00 (0.01)
Course variance	ν_{0k}			0.00 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.30	0.29	0.29
Goodness of fit				
Deviance		517.00	517.00	517.00
Model of reference			Model 1	Model 2
χ^2 fit improvement			$\chi^2 = 0.00$ $df = 1$	$\chi^2 = 0.00$ $df = 1$
P-value			$p =$ n.s.	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is T1 Mental Well-being. Student $n = 320$; group $n = 41$; course of study $n = 10$. N.s. = non-significant; df = degrees of freedom.

Appendix B

Table 2
Establishing random part with multilevel analyses for baseline sense of belonging.

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	3.91 (0.03)	3.93 (0.04)	3.91 (0.05)
Random effects				
Student variance	e_{0ijk}	0.34 (0.02)	0.32 (0.02)	0.31 (0.02)
Group variance	μ_{0jk}		0.02 (0.01)	0.02 (0.01)
Course variance	ν_{0k}			0.00 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.34	0.34	0.33
Goodness of fit				
Deviance		671.03	665.26	664.71
Model of reference			Model 1	Model 2
χ^2 fit improvement			$\chi^2 = 5.77$ $df = 1$	$\chi^2 = 0.55$ $df = 1$
P-value			$p < 0.05$	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is Grade. Student $n = 384$; group $n = 42$; course of study $n = 10$. N.s. = non-significant; df = degrees of freedom.

Appendix C

Table 3
Establishing randomization success with baseline sense of belonging

Effect	Parameter		
		Model 1	Model 2
Fixed effects			
Intercept	γ_{00}	3.93 (0.04)	3.94 (0.05)
Intervention (= 1)	γ_{01}		−0.02 (0.06)
Student variance	e_{0ijk}	0.32 (0.02)	0.31 (0.02)
Group variance	μ_{0jk}	0.02 (0.01)	0.02 (0.01)
Total variance	$e_{0ijk} + \mu_{0jk}$	0.34	0.33
Deviance		665.26	665.16
Model of reference			Model 1
χ^2 fit improvement			$\chi^2 = 0.10$ $df = 1$
P-value			$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is Grade. Student $n = 384$; group $n = 42$; course of study $n = 10$. N.s. = non-significant; *df* = degrees of freedom.

Appendix D

Table 4
Establishing random part with multilevel analyses for grades.

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	6.21 (0.09)	6.14 (0.14)	6.08 (0.24)
Random effects				
Student variance	e_{0ijk}	3.69 (0.25)	3.20 (0.23)	3.15 (0.23)
Group variance	μ_{0jk}		0.47 (0.18)	0.04 (0.09)
Course variance	ν_{0k}			0.42 (0.25)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	3.69	3.67	3.61
Goodness of fit				
Deviance		1790.07	1765.88	1742.17
Model of reference			Model 1	Model 2
χ^2 fit improvement			$\chi^2 = 24.19$	$\chi^2 = 23.71$
			$df = 1$	$df = 1$
P-value			$p < 0.001$	$p < 0.001$

Note. Standard errors are in parentheses. Dependent variable is Grade. Student $n = 432$; group $n = 42$; course of study $n = 10$. N.s. = non-significant; *df* = degrees of freedom.

Appendix E

Table 5
Establishing random part with multilevel analyses for T1 mental well-being.

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	3.48 (0.03)	3.48 (0.03)	3.48 (0.03)
Random effects				
Student variance	e_{0ijk}	0.30 (0.02)	0.29 (0.03)	0.29 (0.13)
Group variance	μ_{0jk}		0.00 (0.01)	0.00 (0.01)
Course variance	ν_{0k}			0.00 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.30	0.29	0.29
Goodness of fit				
Deviance		517.00	517.00	517.00
Model of reference			Model 1	Model 2
χ^2 fit improvement			$\chi^2 = 0.00$	$\chi^2 = 0.00$
			$df = 1$	$df = 1$
P-value			$p =$ n.s.	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is T1 Mental Well-being. Student $n = 320$; group $n = 41$; course of study $n = 10$. N.s. = non-significant; *df* = degrees of freedom.

Appendix F

Table 6
Establishing random part with multilevel analyses for T1 sense of belonging.

Effect	Parameter	Course credits		
		Model 1	Model 2	Model 3
Fixed effects				
Intercept	γ_{000}	3.88 (0.03)	3.88 (0.04)	3.88 (0.04)
Random effects				
Student variance	e_{0ijk}	0.35 (0.03)	0.34 (0.03)	0.34 (0.03)
Group variance	μ_{0jk}		0.02 (0.01)	0.02 (0.01)
Course variance	ν_{0k}			0.00 (0.00)
Total variance	$e_{0ijk} + \mu_{0jk} + \nu_{0k}$	0.35	0.36	0.36
Goodness of fit				
Deviance		575.81	573.11	573.11
Model of reference			Model 1	Model 2
χ^2 fit improvement			$\chi^2 = 2.70$	$\chi^2 = 0.00$
			$df = 1$	$df = 1$
P-value			$p =$ n.s.	$p =$ n.s.

Note. Standard errors are in parentheses. Dependent variable is Grade. Student $n = 320$; group $n = 41$; course of study $n = 10$. N.s. = non-significant; *df* = degrees of freedom.

References

Allen, P. J., de Freitas, S., Marriott, R. J., Pereira, R. M., Williams, C., Cunningham, C. J., & Fletcher, D. (2021). Evaluating the effectiveness of supplemental instruction using a multivariable analytic approach. *Learning and Instruction, 75*, 1–10. https://doi.org/10.1016/j.learninstruc.2021.101481, 101481.

Allen, P. J., Tonta, K. E., Haywood, S. B., Pereira, R. M., & Roberts, L. D. (2019). Predicting peer-assisted study session attendance. *Active Learning in Higher Education, 20*(3), 249–262. https://doi.org/10.1177/1469787417735613

Anthony, R., Moore, G., Page, N., Hewitt, G., Murphy, S., & Melendez-Torres, G. J. (2022). Measurement invariance of the short Warwick-Edinburgh Mental Wellbeing Scale and latent mean differences (SWEMWBS) in young people by current care status. *Quality of Life Research, 31*(1), 205–213. https://doi.org/10.1007/s11136-021-02896-0

Ashwin, P. (2003). Peer support: Relations between the context, process and outcomes for the students who are supported. *Instructional Science, 31*(3), 159–173. https://doi.org/10.1023/A:1023227532029

Bowman, N. A., Preschel, S., & Martinez, D. (2021). Does supplemental instruction improve grades and retention? A propensity score analysis approach. *The Journal of Experimental Education*, 1–25. https://doi.org/10.1080/00220973.2021.1891010

Bronstein, S. B. (2008). Supplemental instruction: Supporting persistence in barrier courses. *Teaching and Learning Assistance Review, 13*, 31–45.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods, 49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

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

Dancer, D., Morrison, K., & Tarr, G. (2014). Measuring the effects of peer learning on students' academic achievement in first-year business statistics. *Studies in Higher Education, 40*(10), 1808–1828. https://doi.org/10.1080/03075079.2014.916671

Dawson, P. (2014). Beyond a definition: Toward a framework for designing and specifying mentoring models. *Educational Researcher, 43*(3), 137–145. https://doi.org/10.3102/0013189X14528751

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

Dekker, I., & Meeter, M. (2022). Evidence-based education: Objections and future directions. *Frontiers in Education, 7*, 1–9. https://doi.org/10.3389/feduc.2022.941410, 941410.

Dobbie, M., & Joyce, S. (2008). Peer-assisted learning in accounting: A qualitative assessment. *Asian Social Science, 4*(3), 18–25. https://doi.org/10.5539/ass.v4n3p18

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gutierrez, A. A., Horner, J. L., & Workman Alyea, L. (2017). *Supplemental instruction training manual. Student learning assistance center*. Texas State University. https://gato-docs.its.txst.edu/jcr:e446fac3-e741-4840-a9b1-bd754bd6514d/SI-Training-Manual.pdf.

Hanson, J. M., Trolian, T. L., Paulsen, M. B., & Pascarella, E. T. (2016). Evaluating the influence of peer learning on psychological well-being. *Teaching in Higher Education, 21*(2), 191–206. https://doi.org/10.1080/13562517.2015.1136274

Herrmann-Werner, A., Gramer, R., Erschens, R., Nikendei, C., Wosnik, A., Griewatz, J., Zipfel, S., & Junne, F. (2017). Peer-assisted learning (PAL) in undergraduate medical education: An overview. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 121*, 74–81. https://doi.org/10.1016/j.zefq.2017.01.001

Hoffman, M., Richmond, J., Morrow, J., & Salomone, K. (2002). Investigating "sense of belonging" in first-year college students. *Journal of College Student Retention: Research, Theory & Practice, 4*(3), 227–256. https://doi.org/10.2190/DRYC-CXQ9-JQ8V-HT4V

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Hurtado, S., & Carter, D. F. (1997). Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. *Sociology of Education*, 324–345. https://doi.org/10.2307/2673270

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*(5), 365–379. https://doi.org/10.3102/0013189X09339057

Johnson, D. R., Soldner, M., Leonard, J. B., Alvarez, P., Inkelas, K. K., Rowan-Kenyon, H. T., & Longerbeam, S. D. (2007). Examining sense of belonging among first-year undergraduates from different racial/ethnic groups. *Journal of College Student Development, 48*(5), 525–542. https://doi.org/10.1353/csd.2007.0054

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

Koushede, V., Lasgaard, M., Hinrichsen, C., Meilstrup, C., Nielsen, L., Rayce, S. B., Torres-Sahli, M., Gudmundsdottir, D. G., Stewart-Brown, S., & Santini, Z. I. (2019). Measuring mental well-being in Denmark: Validation of the original and short version of the Warwick-Edinburgh mental well-being scale (WEMWBS and SWEMWBS) and cross-cultural comparison across four European settings. *Psychiatry Research, 271*, 502–509. https://doi.org/10.1016/j.psychres.2018.12.003

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

U.S. Department of Education. (1995). Supplemental instruction (SI): Improving student performance and reducing attrition. In G. Lang (Ed.), *Educational programs that work* (21st ed.) 14.4). Longmont, CO: Sopris West.

Malm, J., Bryngfors, L., & Mörner, L. L. (2011). Improving student success in difficult engineering education courses through Supplemental Instruction (SI): What is the impact of the degree of SI attendance? *Journal of Peer Learning, 4*, 16–23. http://ro.uow.edu.au/ajplhttp://ro.uow.edu.au/ajpl/vol4/iss1/4.

Martin, D. (2008). Foreword. *Journal of Peer Learning, 1*(1), 3–5. https://ro.uow.edu.au/ajpl/vol1/iss1/2/.

Martin, D., & Arendale, D. (1993). Supplemental instruction: Improving first-year student success in high-risk courses. In *Columbia: National resource center for the first year experience and students in transition* (2nd ed.). 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*(2), 221–231. https://doi.org/10.1080/03075079712331381054

Meeuwisse, M., Severiens, S. E., & Born, M. P. (2010). Learning environment, interaction, sense of belonging and study success in ethnically diverse student groups. *Research in Higher Education, 51*(6), 528–545. https://doi.org/10.1007/s11162-010-9168-1

Ning, H. K., & Downing, K. (2010). The impact of supplemental instruction on learning competence and academic performance. *Studies in Higher Education, 35*(8), 921–939. https://doi.org/10.1080/03075070903390786

Paloyo, A. R. (2015). A note on evaluating Supplemental Instruction. *Journal of Peer Learning, 8*(1), 1–4. http://ro.uow.edu.au/ajpl/vol8/iss1/2.

Paloyo, A. R., 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

Parkinson, M. (2009). The effect of peer assisted learning support (PALS) on performance in mathematics and chemistry. *Innovations in Education & Teaching International, 46*(4), 381–392. https://doi.org/10.1080/14703290903301784

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2020). *A user's guide to MLwiN-Version 3.05*. University of Bristol.

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52*(1), 141–166. https://doi.org/10.1146/annurev.psych.52.1.141

Shah, N., Cader, M., Andrews, B., McCabe, R., & Stewart-Brown, S. L. (2021). Short warwick-edinburgh mental well-being scale (SWEMWBS): Performance in a clinical sample in relation to PHQ-9 and GAD-7. *Health and Quality of Life Outcomes, 19*(1), 1–9. https://doi.org/10.1186/s12955-021-01882-x

Stanich, C. A., Pelch, M. A., Theobald, E. J., & Freeman, S. (2018). A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women. *Chemistry Education: Research and Practice, 19*(3), 846–866. https://doi.org/10.1039/C8RP00044A

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-brown, S. (2007). The warwick-edinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes, 5*(1), 1–13. https://doi.org/10.1186/1477-7525-5-63

Van Herpen, S. G., Meeuwisse, M., Hofman, W. A., & Severiens, S. E. (2020). A head start in higher education: The effect of a transition intervention on interaction, sense of belonging, and academic performance. *Studies in Higher Education, 45*(4), 862–877. https://doi.org/10.1080/03075079.2019.1572088

Van der Zanden, P. J., Denessen, E., Cillessen, A. H., & Meijer, P. C. (2018). Domains and predictors of first-year student success: A systematic review. *Educational Research Review, 23*, 57–77. https://doi.org/10.1016/j.edurev.2018.01.001

Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change, 18*(1), 1–19. https://doi.org/10.1007/s10833-016-9294-4

Izaak Dekker is researcher at the Amsterdam University of Applied Sciences (AUAS). His research interests include educational effectiveness in higher education.

Merel Luberti is a research master student at the University of Amsterdam and (co-)teaches at the AUAS. Her research interests include linguistics and peer learning.

Jantien Stam is the coordinator of the mathematics teacher education bachelor program at the AUAS. Her research interests include educational effectiveness.