Reflective Goal-Setting Improves Academic Performance in Teacher and Business Education: A Large-Scale Field Experiment

Izaak Dekker[1, 2] *, Michaéla C. Schippers[1] & Erik Van Schooten[2, 3]

[1] Department of Technology and Operations Management, Rotterdam School of Management, Erasmus University, Rotterdam, Netherlands

[2] Research Centre Urban Talent, Rotterdam University of Applied Sciences, Rotterdam, Netherlands

[3] Kohnstamm Institute, University of Amsterdam, Amsterdam, Netherlands

*Corresponding author: dekker@rsm.nl

Author Note

Izaak Dekker https://orcid.org/0000-0002-6858-4001

Michaéla C. Schippers https://orcid.org/0000-0002-0795-5454

Erik van Schooten https://orcid.org/0000-0002-2401-9115

Reflective Goal-Setting Improves Academic Performance in Teacher and Business Education: A Large-Scale Field Experiment

Students often have trouble adjusting to higher education and this affects their performance, retention, and well-being. Scholars have suggested applying reflective goal-setting interventions, and most have found positive effects on academic performance and retention. However, one study found no effect at all, stressing the need for understanding the underlying mechanisms, as they could explain when the intervention works and why. Thus, we assessed these mechanisms through a rigorous effect test, using an experimental design and repeated measures. We measured engagement, self-regulated learning, resilience, grit, wellbeing, academic performance, and retention at three points in a large scale randomized controlled trial involving first-year teacher and business education students (N = 1,134). The treatment group earned significantly more course credits and had lower drop out rates. Contrary to previous findings, these effects were independent of gender or ethnicity. Grit, self-regulated learning, resilience, or engagement did not mediate the effects. This study confirmed reflective goal-setting's small and direct effect on academic performance, but no mediating or moderating effects. Differences in implementation fidelity could explain previous studies' varying effect-sizes.

Keywords: Academic performance; Academic achievement; Goal setting; Well-being; Intervention; Field Experiment; Self-regulated learning; Engagement.

Introduction

More than a quarter of all students leave western higher education without obtaining the degree for which they enrolled (OECD, 2019). The majority of those who drop out do so in their first year (Willcoxson, 2010), and ample evidence suggests that this might be due to students having trouble adjusting to higher education (Credé & Niehorster, 2012; Respondek et al., 2020). Difficulty in adjusting to a university and its specific features can lead to stress, poor mental well-being (Bayram & Bilgel, 2008; Catterall et al., 2014; Morosanu et al., 2010), and academic underachievement, manifested as low grades, reduced course credits, and high drop-out rates (Kuh et al., 2007; Reis & McCoach, 2000).

Academic performance is commonly defined as the extent to which students reach their short- and long-term educational goals. The United States and Canada use Grade Point Average (GPA), as an indicator for academic performance, while most European countries measure the number of standardized course credits (European Credits). Universities already invest in broad programs, such as peer coaching, supplementary tuition, mentoring, or summer schools, to improve academic performance and retention, but their effects are rarely tested with controlled experimental designs. Although several rigorous experimental studies have reported successful targeted interventions for specific at-risk students (Sherman et al., 2013; Walton & Cohen, 2011; Walton et al., 2015), these interventions cannot be generalized to a broader population.

Morisano et al. (2010) presented a reflective goal-setting intervention that was low-cost, scalable, and available to a broad student population, based on the principles of goal-setting theory (Locke & Latham, 2002). They reported that the intervention, in which students reflected their desired futures, prioritized goals, and developed strategies in an essay, improved both GPA and student retention. Since then, Dobronyi et al. (2019) and Schippers et al. (2015; 2020) tested the effects of reflective goal-setting interventions. Both Schippers et al. studies (2015; 2020) used a quasi-experimental design on multiple European business school student groups (*N* = 3,144 and 2,928, respectively). In the former study, the intervention enhanced retention rates and course credits by 20%, and although all students benefited, the performance of male students and ethic minorities improved the most (Schippers et al., 2015). The latter study found similar improvements in course credits and reported that participation was related to improved academic performance, regardless of the chosen goal (academic, social, etc.) (Schippers et al., 2020). On the other hand, Dobronyi et al. (2019) performed a large field experiment with first-year students from a Canadian university (*N* = 1,356), comparing the academic performance of a control group, an intervention group, and a group who received the intervention and a brief mindset intervention at the start of the year. Contrary to Morisano et al. (2010) and Schippers et al. (2015; 2020), they found no treatment effect. This raises a few questions: does reflective goal-setting truly have a significant effect, was Dobronyi et al.'s (2019) null effect due to the aforementioned studies' lack of generalizability, and is there a potential confounding factor that has not been taken into account? Regarding the latter, certain moderators that were not included in the previously mentioned studies might play a role and may account for the equivocal results.

Furthermore, prior research indicated the existence of four different types of factors that could shed light on the mechanism behind the intervention. Firstly, Schippers et al. (2015) suggested that gender and ethnicity moderate the effects, with the intervention being more effective for male students and ethnic minorities (Demographics). Secondly, Schippers et al. (2020) found that the number of words that the students write correlates with the intervention's effect, suggesting that the extent and earnestness of student participation, as well as their understanding of the purpose, might influence the results (Implementation fidelity). Thirdly, psychological constructs could explain the underlying mechanism, given that goal-setting aims to direct thoughts and behaviors that subsequently lead to performance (Self-regulation, engagement, grit, and resilience). Regarding self-regulation, goal-setting theory suggests that it mediates the effect of setting goals (Locke & Latham, 2002), but this mediating effect has not yet been tested with reflective goal-setting or in the educational domain. In Travers et al.'s (2015) qualitative diary study that explored the potential mechanism behind reflective goal-setting in higher education, students reported higher engagement. Additionally, Jachimowicz et al. (2018) suggested that reflecting on your passions and goals, and developing strategies improves grit and subsequently, performance. However, another potential explanation might be that the reflective goal-setting intervention boosts resilience, given that the latter is partly dependent on having a goal and particularly benefits struggling students (Azmitia et al., 2018; DeRosier et al., 2013; Windle et al., 2011). Fourth, within higher education, goal-setting interventions have almost exclusively been tested in business and economics courses. Thus, in order to generalize the results to higher education's broader domains and verify whether the intervention is domain-specific or not, samples should also include other types of university students. Policymakers, researchers, and practitioners need more conclusive evidence about the effects of reflective goal-setting interventions, and a definitive understanding of which contexts and under which conditions these effects can be expected.

Lastly, the failure to replicate effects is a widespread phenomenon. As only one-third of the related social psychology studies can be replicated, Maxwell et al. (2015) proposed using more rigorous designs with large power and Locke (2015) suggested aiming to replicate with variation. Replication with variation entails searching for moderators and mediators to inductively expand the theory's generality across different conditions. Accordingly, testing the aforementioned types of potential moderators and mediators can expand goal-setting theory in education, and help us explain when and why reflective goal-setting interventions are effective.

Based on these issues, we measured the four types of moderating and mediating effects in order to perform a replication with variation. We tested the potential treatment with a rigorous experimental design that had enough power to identify the true effects. To situate the results and implications, we divided the literature review into three sections: (1) an overview of goal-setting theory and the intervention's effects on academic performance in higher education, (2) why and how we expected the psychological constructs to mediate the treatment effects on performance, retention, and well-being, and (3) implementation fidelity's role in experimental studies and replications.

Literature Review and Hypothesis Development

Goal-Setting Theory and Interventions

Scholars have extensively studied goal-setting theory, which originated nearly 50 years ago in organizational psychology (Locke & Latham, 2002), and its unique behavioral effects in organizational contexts, sports, and healthcare (Epton et al., 2017). Goal setting, as an intervention, begins with establishing specific and ambitious goals in low complexity contexts. This process improves performance, because it (1) directs attention and efforts to goal-relevant tasks, (2) energizes the individual by separating current and desired states, (3) improves persistence, and (4) indirectly affects the individual's actions by contributing to the discovery and/or use of new strategies (Locke & Latham, 2002).

Although these mechanisms explain the effect of straightforward goal-setting exercises in low complexity contexts, an increasing amount of studies are modifying and applying goal-setting interventions to a first year higher education environment. The latter is a highly complex context, given that the tasks, environment, and expected high self-regulation are new concepts for first-year students. Within this context, three different types of goal-setting have thus far been experimentally or quasi-experimentally tested. These studies were not included in the goal-setting meta-analyses of Mento et al. (1987), Kleingeld et al. (2011), and Epton et al. (2017). Table A.1 in Appendix A offers an overview of all experimental studies examining the effects of goal-setting interventions on academic performance in higher education.

With regards to the three different types of goal-setting applied in higher education, the first type asks students to set goals for the grades, or the number of course credits that students set out to achieve (Clark et al., 2019; Van Lent, 2019; Van Lent & Soeverijn, 2020). For example, van Lent and Soevereijn (2020) performed a field experiment with 1,092 Dutch economy students and instructed a random subset of mentors to encourage students to set grade goals. Within this subset, half of the mentors were further instructed to motivate students to raise their grade goal. Students in the grade-goal group performed significantly better, but those who were pushed to raise their grades performed significantly worse. Van Lent (2019) also conducted a field experiment with 2,100 Dutch economy students, asking half of them to set grade goals or optionally, other goals in a short survey. Compared to the control group, these students did not perform better on their exams. Similarly, in their field experiment with 1,967 American microeconomics students, Clark et al. (2019) reported an insignificant increase in the performance of those who set grade goals. Thus, the evidence shows that goal-setting produces little to no positive effects on academic performance.

The second type of goal-setting intervention targets the specific tasks one wants to complete. The Clark et al. (2019) study also included another field experiment with 2,004 American students enrolled in microeconomics. The students that were randomly allocated to the treatment group were encouraged to set task goals (e.g., the number of online practice exams they would complete before their final exam), while those in the control group received no goal setting encouragements. After the intervention, students in the treatment group reported significantly higher task completion levels and scored marginally higher on performance. Despite the modestly positive results, a placebo effect risk is possible, given that the control group did not receive a control intervention.

The third category allows students to reflect on and determine their own life goals (Dobronyi et al., 2019; Morisano et al., 2010; Schippers et al., 2015; Schippers et al., 2020;

Travers et al., 2015). Whether it be grade or task goals, students are encouraged to choose their most important life goals in any domain. Within this category, different variations exist. In a small-scale trial conducted on struggling students from a Canadian university (*N* = 85), Morisano et al. (2010) tested a version that combined expressive writing exercises (Pennebaker & Chung, 2011) with mental contrasting (Oettingen et al., 2010), implementation intentions (Gollwitzer, 1999), and goal-setting theory. Their results revealed that the treatment group obtained a significantly higher GPA than the control group.

The previously explained Schippers et al. (2015; 2020) and Dobronyi et al. (2019) studies used another version, based on the self-authoring program (selfauthoring.com), that involves similar exercises, but also draws on negative scenarios (e.g., what will happen if you do not change your habits?). Schippers and Ziegler (2019) reviewed the literature and described the different elements that reflective goal-setting interventions should ideally contain. Their proposed version, the life-crafting intervention, emphasizes finding purpose in life and passion during the reflective writing exercises, applies implementation intentions more extensively and includes a final stage in which students publicly communicate their goal.

Although these different versions offer slightly different experiences, they draw on similar mechanisms, and can be categorized as reflective goal-setting interventions, compared to the other categories. Both grade, task, and reflective goal-setting interventions in higher education share a common ground in goal-setting theory, but they differ in how directed and extensive they are. There is some evidence that grade and task goals might lead to small benefits, but other studies show larger or no effects. As Locke and Latham (2005) argued, employing the right moderators or mediators can expand goal-setting theory. Thus, the chosen underlying moderators, which may even be population dependent, may have caused previous studies' varying effects. Furthermore, the aforementioned studies only included small samples of struggling students and large samples of business or economics students. Their findings on the moderating effect of gender and ethnicity are also inconclusive. Therefore, given these quasi-experimental findings, we formulated the following hypotheses:

Hypothesis 1. Students in both business and teacher education, who received a reflective goal-setting intervention at the start of their study, will obtain more course credits and drop-out less than their peers in the control condition.

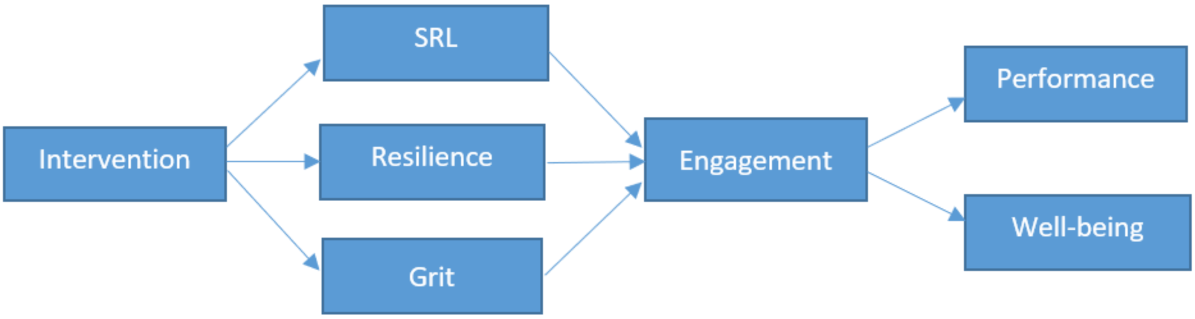
Hypothesis 2. Gender and ethnicity will moderate the intervention's effect on study credits and drop-out rates.

Potential Mediators: Self-Regulated Learning, Resilience, Grit, and Engagement

The recent diversification in the application of goal setting in the educational context has already led to proposed alterations in and additions to goal-setting theory that must be experimentally tested. For instance, Schippers et al. (2020) reported that only one out of five students that participated in the intervention chose an academic goal. Nevertheless, the intervention improved their academic performance, regardless of the subject of their goals. This finding differs from goal-setting theory that argues that task specificity is an essential criterion for success. Travers et al. (2015) studied 92 English university students and found that those who participated in the intervention observed the following mechanisms. After setting life-goals, they had to break these down into smaller goals, as stepping stones, and this enticed short-term action and an immediate increase in effort. Then, they sustained this increase through persistence and self-efficacy, and many reported that this effort led to subsequent engagement. This mechanism overlaps with Schippers' (2017) propositions. Given that a particular intervention can aim to increase students' goal-oriented behaviors, sense of purpose, and explication of their desired futures, Schippers (2017) suggested a focus on improving students' resilience and self-regulatory strategies, as these could lead to higher engagement, academic performance, and well-being (Figure 1).

Figure 1

Mediating Mechanisms between Goal-setting Intervention and Outcomes.



Note. SRL= Self-regulated learning. SRL is a multidimensional and modular construct (Pintrich & De Groot, 1990). For this study we used the modules effort regulation, attention, intrinsic goal orientation, self-efficacy, and metacognition.

In education, self-regulatory behavior is commonly defined as self-regulated learning (SRL), a multi-dimensional construct that includes "the cognitive, metacognitive, behavioral, motivational, and emotional/affective aspects of learning" (Panadero, 2017, p. 1). In their meta-analysis of SRL's effects on students and professionals, Sitzmann and Ely (2011) noted that "one commonality across all the theories is that goal-setting triggers self-regulation" (p. 422), but SRL also has a positive impact on educational attainment (Sitzmann & Ely, 2011). Depending on the goal's specificity, a person's commitment to the goal and his/her related task knowledge both lead to a focus on goal-related activities, effort regulation, persistence, and the use of task-relevant knowledge and strategies (Locke & Latham, 2002). In practice, SRL manifests itself in higher levels of academic initiative, such as active class participation, fewer absences, and less misbehaving in class (Hoyle & Sherrill, 2006; Oyserman et al., 2006). These practical implications are why we expect SRL to be a proxy for engagement and academic performance (Pintrich & De Groot, 1990; Sitzmann & Ely, 2011;).

A goal-setting intervention may also improve resilience, or the capacity to combat adversity, as having a clear goal and following it can enhance resilience (Connor & Davidson, 2003; Turner et al., 2011). Increased resilience might particularly benefit students who are more dependent on support systems (Azmitia et al., 2018; DeRosier et al., 2013). If a goal-setting intervention helps them set goals, the resilience they develop in the process further helps them persevere whenever they encounter a setback. As previous studies have found that resilience supports both academic performance and well-being (Johnson et al., 2015; Martin et al., 2015), resilience could also mediate a goal-setting intervention's influence on academic performance and well-being (see Figure 1).

Grit, related to SRL, engagement, and resilience, could also potentially explain why students, who have formulated their goals, persevere and perform well. Duckworth et al. (2007), who coined the term, defined it as a "perseverance and passion for long-term goals" (p. 1087). Studies have found that it can predict academic performance and engagement (Duckworth et al., 2007; Bowman et al., 2015; Hodge et al., 2017).

Engagement, characterized by dedication, vigor, and absorption, is "a persistent and pervasive affective–cognitive state that is not focused on any particular object, event, individual, or behaviour" (Schaufeli & Bakker, 2004, p. 295). Dedication is "a sense of significance, enthusiasm, inspiration, pride, and challenge," and to work with vigor means to have "high levels of energy and mental resilience […], the willingness to invest effort in one's work, and persistence also in the face of difficulties" (p. 295). Absorption refers to a state in which one loses track of the time by being highly concentrated and immersed in an activity. Accordingly, Travers et al. (2015) found that students who engaged in the reflective goal-setting intervention had higher vigor, dedication, and absorption levels. Overall, engagement relates to observed learning activities and course grades, and may be a mediating factor between SRL and academic performance (Bakker et al., 2014). Thus, reflective goal-setting could potentially improve SRL, resilience, grit, and engagement. If engagement is affected, this could, in turn, lead to improvements in performance and well-being (Schippers, 2017).

Well-being

Student well-being has recently become a concerning issues in academia (Auerbach et al., 2018). Specifically, policymakers and scientists argue that many measures that aim at improving academic performance do so at the cost of students' well-being. However, reflective goal-setting interventions aim to improve both academic performance and well-being, because they challenge students to set academic, social, and health-related goals (Schippers, 2017; Schippers & Ziegler, 2019). The action of setting a goal is not expected to increase well-being directly, but having the right priorities and strategies should help students engage in activities that allow them to pursue their goals in a healthy way. Therefore, we expect their engagement to lead to an increase in well-being. In line with Schippers (2017) and based on our expectations of a reflective goal-setting intervention's mechanisms, we propose the following hypotheses (following Figure 1's conceptual model).

Hypothesis 3. Students in the treatment condition will have a significantly higher growth in SRL (effort regulation, attention, intrinsic goal orientation, self-efficacy, and metacognition), resilience, grit, engagement, and well-being than their peers in the control condition.

Hypothesis 4. Gender (higher effect for males) and ethnicity (higher effect for ethnic minorities) will moderate the intervention's effect on SRL, resilience, grit, and engagement growth in both business and teacher education students.

Hypothesis 5. SRL, grit, resilience, and engagement will mediate the intervention's effect on course credits, drop-out rates, and well-being (Figure 1).

Implementation Fidelity

Implementation fidelity, or the degree to which an intervention is delivered as intended, is critical for successfully translating evidence-based interventions into practice. Previous studies' inconclusive results could be a result of the differences in intervention implementation. For instance, Durlak and DuPre (2008) revealed that careful implementation can result in larger effect sizes. In line with Dane and Schneider's (1998), and Carroll et al.'s (2007) models, Horowitz et al. (2018) applied their findings to the field of educational psychology and summarized the fidelity concerns into the following six broad categories: program differentiation, dosage, adherence, quality of delivery, student responsiveness and fidelity-of-receipt.

Program differentiation is the degree to which the tested intervention can be differentiated from the regular program. Using similar interventions with different names might disturb the potential effects and this is a particular risk for certain elements in goal-setting interventions, considering that goal-setting theory has been around for decades (Locke & Latham, 2002). Thus far, goal-setting educational literature has not reported program differentiation degrees. *Dosage* refers to 'how much' of the intervention was done, measured with completion rates, hours spent on the intervention, or output variables, such as the number of written words, as reported by the Morisano et al. (2010) and Schippers et al. (2020). *Adherence* refers to whether the treatment's parts were followed in the correct sequence. *Quality of delivery*, particularly important when teachers or non-researchers must deliver an intervention, is successful when participants experience the main points as easy to process, true, and emerging naturally (Horowitz et al., 2018). *Student responsiveness* involves students' responses to the adherence and quality of delivery. Lastly, *fidelity of receipt* refers to the degree to which students internalize the main points that the intervention aims to communicate. These dimensions require attention, as they provide conditional information that is expected to influence the results of an experimental study (Durlak, 2015; Durlak & DuPre, 2008).

Methodology

Research Design

We conducted a large-scale field experiment at the beginning of the 2018-2019 academic year to test hypotheses 1-5. The intervention consisted of two sets of assignments that were individually completed in computer rooms at a university. The participants, who were randomly and secretively assigned to a treatment or control group, were monitored during the assignments. The participants in the control group created control assignments that looked nearly identical to the intervention group's, but contained questions about the past instead of the future. We measured the intervention's effects on SRL, grit, resilience, and engagement at three points in time (T0, T1, and T2) with a survey. We conducted T0 at the start of the year and before the intervention, T1 at the end of the first semester, and T2 at the end of the second semester. We measured the intervention's effects on academic performance during T1 and T2 with the use of administrative data.

Participants

The sample consisted of first-year students enrolled in 13 courses of study[1] from two faculties within a large Dutch university of applied sciences, located in an urban environment. With regards to the student population in these universities in The Netherlands, 43% followed an academic track in high school and 31% have a vocational education background (The Netherlands Association of Universities of Applied Sciences, 2020). We controlled for this

[1] The Dutch higher education system differs from the Anglo-American system in that students have to enroll for a specific course of study (comparable to choosing a major) that consists of a standard curriculum with few or no electives in the first year. Dropping out in this context means abandoning a complete course of study with all of the courses that it contains. Under the current Dutch law, students are not allowed to re-enroll for a course of study at the same university if they fail to successfully complete all their first year courses within two years.

sample characteristic in our analysis, because it differs slightly from the samples of previous studies (Dobronyi et al., 2019; Schippers et al. 2015; 2020) and because previous education in The Netherlands is strongly related to central exam scores (similar to SAT scores), which is a predictor for performance (Van der Zande et al., 2018).

The sample was taken from teacher education and business studies faculties. Within the business faculty, two out of five courses participated with all their 302 first-year students. In the teacher education faculty, 11 out of 13 courses participated with a total of 832 first-year students. Table 1 shows an overview of the participant characteristics. During our interactions with teachers and managers, we compared the existing program to all parts of the reflective goal-setting intervention to determine program differentiation. As no courses used any parts of the intervention, we could include all parts in the experiment.

Table 1

Sample Characteristics of the Freshmen per Faculty and Condition

	Business		Education		Treatment		Control	
	N	%	*N*	%	*N*	%	*N*	%
Participants	302	27	832	73	571	50	563	50
Male	208	69	333	40	268	47	276	49
Ethnic minority	73	24	275	33	177	31	175	31
Vocational background	85	28	225	27	154	27	158	28

The internal review board of the researchers’ affiliated university approved the experiment before execution. All participants signed informed consent forms before being included. The procedure in the data management plan ensured the use of pseudonyms before datasets were merged, and anonymous and save storage afterwards. After the experiment, all the participants were debriefed and received a book about classroom management (teacher education) or a business journal (business education).

In total, 942 (81%) finished both parts of the treatment. We did not find any significant differences in participation rates between the groups. Out of the total of 1,134 students, 1,060 completed every item of the T0 survey and 504 finished the T1 survey online. To secure enough response for the third survey, we distributed the T2 survey in paper format during the classes (653 responses). To assess whether missing responses had potentially led to a non-response bias, we performed several non-response analyses. Specifically, we used a multilevel logistic regression analysis to test whether participation in one of the surveys significantly correlated with being part of the treatment group or relevant control variables (gender, ethnicity, and previous education). The response did not significantly differ from the sample based on assignment to the treatment group, gender or previous education. However, significantly less students from an ethnic minority responded to the survey, although this difference was relatively small (For survey T0 r^2 = [1, N = 1,134] .036, $p < .001$; for survey T1 r^2 = [1, N = 1,134] .010, $p < .05$; for survey T2, r^2 = [1, N = 1,134] .007, $p < .05$).

After screening, we removed 104 cases in the T0 survey, 21 cases in the T1 survey, and 23 cases in the T2 survey (those who responded the same answer to all questions, or did not clearly write their identification number in the T2 survey). The final dataset contained 1,134 cases with demographic data, study credit, and drop-out status, of whom 956 had T0 survey scores, 483 had T1 scores, and 630 had T2 scores. As we used repeated measures, we could apply full information estimation in MLwiN (Rasbash et al., 2020). This led to a sample of 1,045 students in the repeated measures growth model.

We calculated power with the G*Power 3 program (Faul et al., 2007). For linear regression that we used to measure effects on study credits or drop-out rates, a sample of 90 was required to find a small (*f* .15) effect size at the 5% confidence level with a power of .95. The sample in this study contained nearly 13 times as many cases. For the growth models that we employed to study potential mediating mechanisms, obtaining two groups with three repeated measures, a .5 correlation between repeated measures, and a .9 correction for non-sphericity required a sample of 230 at a .9 power level. We corrected this for multilevel structure (Hox et al., 2010): Neff = N / [1 + (nclus - 1) rho]. Neff = 230 leads to a required *N* = 230 * 4.45 = 1,023.50 to find a *f* .1 effect size, and *N* = 556 to find a *f* .15 effect size (both small).

At the end of the year, we asked a random selection of 20 students from the treatment group to partake in qualitative focus groups for evaluation purposes and 14 of these students attended. We asked them to evaluate the two parts of the intervention, describe if they had learned anything, and if they had applied what they had learned beyond the intervention. All study programs, except pre-service economics teachers, were represented in this group. Eight of the participating students were female, four were ethnic minorities, and seven had a vocational education background.

Data Analysis

Measuring Fidelity

We recorded and transcribed the two focus group conversations, and followed a particular protocol to ensure that we evaluated all parts of the intervention, the students' experiences, and the degree to which they had internalized the main points. Specifically, we used axial coding to form categories from the answers, and asked the students, through an email member check, whether they agreed with the derived summary and answer categories.

Testing Randomization

We conducted independent sample t-tests and χ^2 tests to verify the success of the randomization. This involved assuring that there were no significant differences in the dependent variables (SRL, grit, resilience, engagement, and well-being), demographics, and high school GPA (previous performance is a strong predictor of future performance) between the control and treatment groups before the intervention (T0). As Levene's test indicated unequal variances for metacognition (*F* = [1, 950] 4.37, *p* = .04) and resilience (*F* = [1, 950] 5.86, *p* = .02), we adjusted the degrees of freedom accordingly (Table 2). The T0 survey scores showed no significant variable differences between the treatment and control groups (Table 2), confirming that the randomization was successful. Table B.2 to B.4 in the Appendix present the intercorrelations between the latent traits in the confirmatory factor analyses (CFA).

Table 2

Descriptive Statistics with Administrative and Survey Data and Results χ^2 or Independent T-tests

	Control Sample mean (*SD*)	Difference with treatment group *(SE)*	χ^2 or t-value (*df*)	*p*-value	*N*
Male*	.49 (.50)	.02 (.02)	.582 (1)	.45	1,134
Ethnic minority background*	.30 (.46)	.01 (.02)	.010 (1)	.92	1,134
Vocational background*	.28 (.45)	-.01 (.05)	.01 (1)	.94	1,134

GPA High School[2]	6.50 (.44)	-.48 (.24)	-1.56 (70)	.12	701
T0 Self-efficacy	3.92 (.56)	.01 (.03)	-.14 (96)	.89	958
T0 intrinsic g. orient.	4.21 (.50)	.05 (.02)	1.43 (95)	.15	956
T0 Metacognition	3.42 (.62)	.03 (.03)	.67 (947.23)	.50	952
T0 Attention	3.46 (.67)	.05 (.03)	1,057 (947)	.29	949
T0 Effort regulation	3.73 (.52)	.05 (.03)	1,474 (958)	.14	960
T0 Resilience	3.93 (.48)	.00 (.03)	.010 (948.93)	.99	956
T0 Grit	3.65 (.52)	.05 (.03)	1,370 (958)	.17	960
T0 Engagement	3.32 (.66)	.01 (.03)	.34 (954)	.73	956
T0 Well-being	4.55 (.73)	-.04 (.03)	-.75 (954)	.46	956

**= tested by means of* χ^2 *since variable is dichotomous. df* = degrees of freedom

Note. Analysis done unilevel because the students did not yet belong to natural groups upon entry.

Measuring Treatment Effects on Performance and Behavior

As the sample consists of natural groups (courses and faculties), we conducted multilevel regression analyses when the intra-class correlations of the study program or faculty appeared to be significant. The intervention's effect on the social-cognitive variables was estimated with multilevel growth models through three repeated measures in MLwiN (Rasbash et al., 2020). We verified if the growth was non-linear by testing whether adding time-squared to the equation significantly improved the model fit. This allowed us to infer if the treatment was related to higher scores at both points in time. In these models, we estimated the treatment's effect on growth, as the interaction between time and condition. We

[2] GPA in Dutch High Schools is measured on a 10-point scale, 6 is the threshold for passing. Students with a Dutch tertiary vocational education degree are admissible to a university of applied sciences without having a GPA score.

also estimated the hypothesized moderation effects of gender, previous education, and ethnicity (hypothesis 4) through these growth models. The tested models included condition (intervention), gender, previous education, time (and time-squared), ethnicity, and the these variables' interaction terms as fixed effects. We included faculty (*N* = 2) and course (*N* = 13) as variance levels in the random effects whenever this led to a significant model fit improvement. Testing for non-linearity was relevant, because performance in credits accumulates, while behavior (in a literal sense) does not, and our theory predicted that the intervention would have particular time-dependent effects at the start of the study.

First, we looked for the intervention's direct effects on growth on every construct separately (hypothesis 3). Second, we tested whether any effects might be moderated (hypothesis 4). When no direct effect was found, we could also exclude a mediated effect (hypothesis 5) (Fairchild & MacKinnon, 2009). We tested the models' fit improvements by means of the difference in deviance (-2*loglikelihood) between nested models. This difference has a chi-square distribution with the difference in the number of parameters estimated as degrees of freedom. Effect sizes are calculated as the proportions of explained variance between the nested models, both for total variance and for variances per level. After fitting the growth models, we also performed an ordinary multilevel analysis for every psychological construct separately to verify if this resulted in different outcomes.

For the dependent variable 'course credits' (hypothesis 1), we did not use growth models, but ordinary multilevel modeling, as all the students started with zero course credits (T0). Therefore, we only have two measurements for course credits (T1 and T2). Using a RCT as the study design, the condition's effects on T1 or T2 reflects the goal-setting treatment's effects. In the analyses, with T1 as the dependent variable, we had to use two variance levels (student and course) (see Appendix Table B.5.0). For the obtained credits after a year (T2), adding a course or faculty level to the student level did not significantly improve the model fit (Appendix Table B.6.0.). Therefore, we conducted unilevel analyses with this dependent variable.

Measuring Treatment Effects on Drop-out Rates using Multilevel Logistic Regression

As dropping out of a study program is a binary variable (1 = drop-out, 0 = not), we used logistic regression analyses for this dependent variable and verified whether a multilevel logistic regression was needed. We obtained the starting values for this analysis using first order marginal quasi-likelihood and the final model fit with second order predictive quasi-likelihood (Rasbash et al., 2020). Adding the course level to a logistic regression model did not significantly improve the model fit (χ^2 = [1] .18, *p* = n.s.). It can be inferred that the faculty level is not needed either, because courses are nested in the faculties. Therefore, we conducted a binary logistic regression in SPSS to measure the treatment's effect on drop-out rates, with and without controlling for gender, ethnicity, and previous education. We used Nagelkerke's r-square to estimate the proportion of explained variance per model, and the difference in Nagelkerke's r-square for the fit improvement between nested models. Also, we calculated the log odds, as an indication of the independent variables' effects.

Instruments

We measured dosage fidelity by tracking the completion rates and the number of words that students wrote in both parts of the intervention (Table 1). Three items at the end of the intervention and control group tested student responsiveness to the intervention on a five-point Likert scale, ranging from disagree or agree: serious participation, if they learned something, and if the intervention shaped their thoughts about their future. We also qualitatively assessed both student responsiveness and receipt fidelity at the end of the year with two focus groups (*N* = 14, intervention only).

The selected university used the European Credit Transfer and Accumulation System (ECTS). Within a year, students are expected, when successful, to obtain 60 ECTS course credits that stand for 1,680 study hours (1 credit amounts to 28 study hours). In their first-year, students need to obtain a minimum of 42 out of 60 ECTS to be allowed to continue studying. Thus, we measured academic performance by tracking the participants' obtained ECTS credits and drop-out/retention rates, supplied by the university administration.

The following standardized scales measured SRL (self-efficacy, intrinsic goal orientation, metacognition, effort regulation, and attention), resilience, grit, engagement, and general psychological well-being (PGWB). The modular subscales for effort regulation, metacognition, attention, intrinsic goal orientation, and self-efficacy stem from the Motivated Strategies for Learning Questionnaire (MSLQ) (Duncan & McKeachie, 2005; Pintrich et al., 1993). Both subscale selection and Dutch translation were based on a previous study that tested the instruments on Dutch professional higher education students (De Bruijn-Smolders, 2017). We measured resilience with a Dutch translation of the 10-item Connor-Davidson Resilience Scale (Campbell-Sills & Stein, 2007), grit with a Dutch translation of the 10-item GRIT-S scale (Duckworth & Quinn, 2009), and well-being with a Dutch translation of the six-item PGWB scale (Grossi et al., 2006). Schaufeli et al.'s (2006) nine-item UWES scale served to measure student engagement.

Most subjective and psychological well-being scales include items that are closely related to having a goal or purpose (Klug & Maier, 2015; Ryff & Singer, 1996). This could cloud conceptual clarity and make the correlation between goal pursuit and subjective well-being spurious. The short PGWB scale covers six health-related quality of life domains and none of the items overlap with setting or having a goal: anxiety, depressed mood, positive well-being, self-control, general health, and vitality. Therefore, using this scale allows for a more valid testing of goal setting's effect on well-being.

Half a year before the experiment, we pre-tested all the scales on a small sample of students from a different cohort with the think-aloud method (Ryan et al., 2012). After this assessment, we made minor language adjustments to replace complicated words and ambiguous formulations.

Psychometrics

We performed a CFA with the Mplus program (Muthén & Muthén, 1998-2006) on the questionnaire items to verify the self-efficacy, intrinsic goal orientation, metacognition, effort regulation, attention, resilience, grit, engagement, and well-being scales' validity. We calculated the covariance structures using weighted least squares with means and variances (WLSMV), because the scores are categorical (Likert scales). For each measurement moment, we conducted a separate CFA. After the initial CFA, we used modification indices and factor loadings to identify problematic items. As the variables were summed per used scale in the repeated measures' multilevel regression analyses, the models for each of the three measurement moments must contain the same items. Based on the modification indices, only two items had to be removed. Table 3 shows the results of the CFA before and after this removal from all repeated measures. Table 4 depicts the reliability of the scales at every repeated measure and after the two item removal. The scales' Cronbach's alpha reliabilities range from moderate (.65) to robust (.86) (Taber, 2018). All scales have alphas above .7, except for effort regulation and intrinsic goal orientation that are slightly under.[3]

[3] The authors of the final validated MSLQ version reported similar (.69 - .74) alpha coefficient's for these subscales (Duncan & McKeachie, 2005).

Table 3

Results CFA (WLSMV)

	T0	T1	T2
χ^2	5,388.69	4,359.32	5,496.47
df	1,793	1,793	1,793
p	.000	.000	.000
RMSEA (90% CI)	.05 (.04-.05)	.05 (.05-.06)	.06 (.05-.06)
CFI	.89	.86	.81
TLI	.89	.85	.80

Note. CFA performed with 62 items (after removal of 2 items). For an extended table with the results before removal of see Table B.1 (in the appendix). Sample sizes: T0 *N* = 960; T1 *N* = 505; T2 *N* = 666.

Table 4

Reliability of the Item Sums per Construct at T0, T1 and T2 (after removal of 2 items)

Scale	*N*	Cronbach's *α*	*N* -items	Range c-i-t-c	items removed
T0 selfeff	958	.75	5	.43 - .62	-
T1 selfeff	499	.75	5	.41 - .65	-
T2 selfeff	617	.75	5	.41 - .55	-
T0 intrins	956	.70	5	.35 - .56	-
T1 intrins	497	.73	5	.37 - .59	-
T2 intrins	624	.68	5	.40 - .49	-
T0 meta	952	.77	7	.43 - .58	-
T1 meta	497	.75	7	.28 - .57	-
T2 meta	607	.77	7	.41 - .53	-
T0 attent	947	.78	6	.40 - .65	-
T1 attent	496	.79	6	.38 - .68	-
T2 attent	641	.78	6	.45 – .60	-
T0 effort	953	.65	5	.30 - .53	1
T1 effort	500	.67	5	.31 - .58	1

T2 effort	654	.66	5	.35 - .55	1
T0 resil	944	.82	10	.36 - .58	-
T1 resil	481	.86	10	.41 - .63	-
T2 resil	611	.81	10	.30 - .56	-
T0 grit	937	.78	10	.26 - .56	-
T1 grit	494	.75	10	.25 - .55	-
T2 grit	592	.72	10	.24 - .53	-
T0 engag	951	.83	8	.32 - .70	1
T1 engag	485	.85	8	.46 - .72	1
T2 engag	617	.80	8	.37 - .66	1
T0 wellb	956	.79	6	.49 - .64	-
T1 wellb	483	.85	6	.56 - .71	-
T2 wellb	614	.86	6	.52 - .73	-

c-i-t-c= corrected item total correlation

We used several fit indices to evaluate the model fit. As the χ^2 statistic is highly sensitive to sample size and tests exact fit, which is too strict a criterion for the social sciences, we also used the Comparative Fit Index (CFI), Tucker Lewis index (TLI), and Root Mean Square Error of Approximation (RMSEA). Generally, a model is considered fair when CFI and TLI ≥ .90, and good when CFI and TLI ≥ .95 (Hu & Bentler, 1999). In addition, RMSEA-values (upper estimate of the 90% confidence interval) of ≤ .05 are considered a close (good) fit, between .05 and .08 a fair fit, between .08 and .10 a mediocre fit, and > .10 a poor fit (Hu & Bentler, 1999; MacCallum et al., 1996). The χ^2 of the three models indicate no exact fit and all the RMSEA values of the models indicate a good or fair fit, but the CFI and TLI range between .80 and .89, which is slightly below the fair fit value. All items load significantly on the factor they are supposed to measure, and we also did not find perfect correlations between factors. Therefore, the overall validity of the instruments seems reasonable, the different constructs show good discriminant validity, and the reliabilities are moderate to robust.

Results

Implementation Fidelity

We assessed implementation fidelity using Horowitz et al.'s (2018) six categories. Regarding the dosage fidelity, 536 students (94%) finished part one of the intervention and 470 (82%) finished both parts. We ensured that every student completed parts 1 and 2 in the right sequence by closing the access to part 1 before sending part 2 to the students (adherence). We are able to cover quality of delivery, because the intervention was delivered online and the conditions were controlled in surveilled computer classrooms. The items that measured responsiveness indicate that 69.9% of the participants in the treatment condition, who completely finished both parts, agreed that they took the assignments seriously. One in five (20.1%) neither agreed nor disagreed and 9.2% disagreed. The degree to which the students took the assignment seriously correlated significantly with the number of written words (r = .36, p < .001).

During the focus groups, two students reported that they did not take the assignment seriously, because "it was part of an experiment" and "because I don't like writing so much." A few students reported that the intervention had influenced their behavior, three of which noted its influence in other domains as well as academia. One student stated that the intervention had helped him combat both his planning and financial issues right at the start of his studies. Another student noted remembering writing down a social and academic goal: "the intervention made me realize that I should stop my loner behavior and try to fit in socially […] the academic goal made me ask for help sooner whenever I got stuck."

Half of the students in the focus group, seven of 14, initially did not remember taking part in the intervention, similar to what other researchers reported (Walton & Cohen, 2011). However, some did remember it later on in the conversation: "It was right at the start of the study, it was a chaotic period, and I've forgotten nearly everything that happened." Some of these students later stated that they did think it brought them more focus at the start of their study. When we discussed potential intervention improvements, all the students in the focus group agreed that a more personalized follow-up would aid them internalize and utilize the intervention throughout the course of the year. As one student put it: “One’s teacher or coach should recall the intervention one period later. You write down your goals then, but now you are here in this point in time. What about these goals now?” When asked if email reminders would suffice, the students reported that they already received too many emails and would perceive this as a burden rather than helpful. Overall, these results indicate moderate implementation quality. Therefore, we expect to still find a (suboptimal) effect of the intervention.

Hypotheses

Students received an average of 17.24 course credits in the first semester. Those in the treatment group, on average, earned 1.04 study credits more than their peers in the control group during the first semester, which is a significant difference (Table 5, models 1 and 2). This advantage becomes slightly larger and remains significant when we first control for previous education, ethnicity, and gender (Table 5, models 3 and 4). To test whether the intervention works better for subgroups, as determined by Schippers et al. (2015), we added the interaction effects between condition and previous education, ethnicity, and gender, respectively, to a model, with the main effects being condition, previous education, and gender. However, none of these moderator effects proved a significant improvement to the model (Table B.5.1, Appendix). This suggests that the intervention did not work differently for males, ethnic minorities, or those holding a vocational education background.

Table 5

Treatment Effects on Course Credits After One Semester

Model	1	2	3	4
Fixed part				
Intercept	17.24 (.94)	16.73 (.97)	20.86 (.96)	20.33 (.99)
Intervention (= 1)		1.04* (.53)		1.09* (.50)
Vocational background (= 1)			-3.59*** (.60)	-3.60*** (.59)
Ethnic minority backg. (= 1)			-3.52*** (.59)	-3.54*** (.59)
Male (= 1)			-3.21*** (.55)	-3.20*** (.55)
Random part				
Student variance	77.04 (3.27)	76.77 (3.26)	70.03 (2.97)	69.73 (2.96)
Course variance	10.13 (4.46)	10.12 (4.45)	9.00 (3.98)	9.00 (4.00)
Total variance	87.17	86.88	78.99	78.72
Deviance	8,102.86	8,098.92	7,995.29	7,990.59
% expl. var. student level		.35	9.10	.42
% expl. var. study program level		.17	11.58	-
% expl. var. total		.33	9.39	.34
Sig. difference of fit		model 1	model 1	model 3
compared to …		$\chi^2_{(1)} = 3.94$	$\chi^2_{(3)} = 107.58$	$\chi^2_{(1)} = 4.70$
		$p < .05$	$p < .001$	$p < .05$

*=sig. at 5%; ** sig. at 1%; ***=sig. at 0.1%. (n.s.=non significant)

Note. Standard errors are presented in parentheses. Student $N = 1{,}134$; study program $N = 13$; faculty $N = 2$.

At the end of the first year, the students earned an average of 42 course credits. Students assigned to the treatment group earned 2.7 credits more than their peers in the control group. After controlling for previous education, ethnicity, and gender (Table 6, models 3 and 4), the difference between the treatment and control groups decreases to 2.5 credits, but remains significant ($p < .05$). As with the study credits at T1, there are no significant interaction effects: the intervention seems equally beneficial for all sub-groups and the effect is not dependent on gender, background, or ethnicity. However, the intervention's effect sizes on course credits are small. After controlling for previous education, ethnicity, and gender, the intervention explains 0.34% of the variation in credits at T1 and 0.35% at T2. However, students on average only invested two hours in the intervention and one study credit amounts to 28 study hours.

Table 6

Treatment Effects on Course Credits After One Year

Model	1	2	3	4
Fixed part				
Intercept	42.01 (.67)	40.65 (.95)	50.52 (1.09)	49.21 (1.27)
Intervention (= 1)		2.70* (1.34)		2.53* (1.28)
Vocational backgr. (= 1)			-9.96*** (1.50)	-9.95*** (1.49)
Ethnic minority b. (= 1)			-7.00*** (1.46)	-7.01*** (1.46)
Male (= 1)			-7.56*** (1.30)	-7.50*** (1.30)
Random part				
Student variance	508.26 (21.35)	506.44 (21.27)	463.86 (19.50)	462.26 (19.41)
Deviance	10,284.11	10,280.02	10,180.50	10,174.68
% expl. var. student (= total) level		.36	8.74	.35
Sig. difference of fit compared to model		Model 1 $\chi^2_{(1)} = 4.08$ $p < 0.05$	Model 1 $\chi^2_{(3)} = 103.66$ $p < 0.001$	Model 3 $\chi^2_{(1)} = 5.77$ $p < .05$

*sig. at 5%; **sig. at 1%; ***sig. at 0.1%. (n.s.=not significant)

Note. Standard errors are presented in parentheses. Student *N* = 1,134; study program *N* = 13; faculty *N* = 2.

With respect to drop-out rates, the results were similar: 39% of all students in the control group dropped out of their study program during the first year, compared to 33% in the treatment group. The logistic regression shows that the intervention significantly predicts drop-out rates (*p* = .036), but the proportion of explained variance is small. The log odd is .772, meaning that a student in the control group has a 1.3 times higher chance of dropping out than one in the intervention group. After controlling for previous education, ethnicity, and gender, the intervention's effect is still significant (*p* = .042) and the three covariates together are highly significant (*p* = .000). We may also conclude that after controlling for the three covariates, the intervention explains a proportion of .5% extra variance in drop-out rates. Therefore, hypothesis 1 is accepted, while hypothesis 2 is rejected. The cost-benefit ratio can be considered good, because the treatment has a time investment of about two hours per student, resulting in an average 2.5 extra credits (approximately 70 study hours) and 6 percentage point less drop-outs at the end of the year.

Our third hypothesis predicted a treatment effect on growth in SRL, resilience, grit, engagement, and well-being. Contrary to expectations, both multilevel growth and regression models that measured treatment effects after one and two semesters showed no direct significant treatment effects on effort regulation, metacognition, attention, intrinsic goal orientation, self-efficacy, grit, resilience, engagement, or well-being (Table B.7-B.15, Appendix). Therefore, hypothesis 3 is rejected.

Although it is unlikely to find a moderator effect without a direct effect, it is still potentially possible. Thus, we continued testing whether significant treatment effects could be found if we added gender and ethnic minority as moderators (hypothesis 4). None of these models proved significant, rejecting hypothesis 4.

Hypothesis 5 supposed that the selected SRL modules, grit, resilience, and engagement would mediate the treatment effect on performance and well-being. However, no mediation can occur, because we did not find a direct effect of the intervention on well-being (Fairchild & MacKinnon, 2009), rejecting hypothesis 5.

Discussion

As universities are looking for scalable and low-cost interventions that could aid a broad population, a reflective goal-setting intervention could potentially provide a solution. However, thus far, the evidence about its effectiveness is divided, the mechanism that could explain why and when it works is still underexplored, and the domains in which it is tested are relatively limited. The reflective goal-setting intervention in this study yielded a significant positive effect on course credits and retention. In contrast to earlier results (Schippers et al., 2015), the effect was independent of domain, gender, ethnicity, or educational background. Also contrary to expectations, the treatment group did not differ significantly in SRL, grit, resilience, and engagement growth, these constructs do not appear to be mediators between the intervention and academic outcomes.

Our findings expand the literature on reflective goal-setting and life crafting’s effects on academic performance in several ways. First, we bridged the conflicting findings on its effectiveness, as noted in the literature review, showing a smaller effect size than the small-scale quasi-experimental studies, but a significant positive effect, contrary to Dobronyi et al. (2019). Previous studies did not monitor implementation fidelity or only partially. Thus, to our knowledge, this was the first goal-setting intervention study to assess implementation fidelity as part of the design. Due to its moderate fidelity, we expect that the intervention’s effect may have been suppressed. The degree to which the intervention has been successfully implemented thus far could potentially explain the differences we found in effect sizes. For instance, in terms of student responsiveness, 70.1% reported taking the intervention seriously. Among the reasons were a lack of communication and being part of an experiment. These issues are particular to the design of large-scale experiments and could explain smaller effect sizes. A second example is the intervention’s dosage fidelity. Prior research showed the number of written words to be a significant predictor of academic performance (Schippers et al., 2015; 2020). Students in the current study wrote nearly three times less than the average of around 3,000 words in Morisano et al. (2010) and Schippers et al. (2020).[4] Writing more can be an indicator of more extensive reflections and more specific goal achievement plans. Thus, part of the intervention’s effect could potentially be attributed to dosage fidelity. Future studies can build on this approach to ensure that implementation fidelity is closely monitored and taken into account in a meta-analysis. Practitioners could monitor this variable as a potential condition for success.

Second, the intervention did not improve the SRL modules, grit, resilience, engagement, or general psychological well-being. Thus, the constructs did not mediate the treatment effect, contrary to Schippers (2017) expectations, nor did the intervention lead to expected significant benefits on well-being, as suggested by Schippers and Ziegler (2019). This improves the accuracy of our knowledge by rejecting hypotheses that previous studies supported based on correlational evidence (e.g., Sitzmann & Ely, 2011; Travers et al., 2015). It is particularly striking that we found no intervention effects on SRL or specifically effort regulation, given all the previous findings on this effect in other contexts (Locke & Latham, 2002). This might suggest that either the first year of higher education is substantially different from the contexts in which goal-setting interventions have thus far been tested, or that reflective goal-setting has a distinctly different effect from other types of goal-setting interventions.

[4] Dobronyi et al. (2019) did not report the number of words.

Third, we expanded the intervention to a new domain. Specifically, reflective goal-setting interventions have only been applied to students studying business or economics, and we showed that their effects can also be reproduced in the context of teacher education.

Fourth, we specified the degree to which reflective goal-setting interventions can improve equal opportunities in college. Thus far, quasi-experimental studies have indicated that such interventions could close the achievement gap, suggesting that underperforming male students and ethnic minorities would benefit more from the intervention (Schippers et al., 2015). However, we found no significant interaction effects between these variables and the intervention on course credits, highlighting that the intervention affected performance irrespective of gender, ethnicity, or previous education. Given the high power and large sample size of this study and the spread among gender and ethnicity, a type II error is unlikely.

Finally, we found positive treatment effects both after a semester and at the end of the year. As the treatment effect on obtained course credits grew proportionately, the intervention had a durable benefit that improves over time. This finding is in line with Walton (2014) as well as Schippers and Ziegler (2019), who argued that a well-timed intervention at the start of one's studies can create a positive recursive spiral or stop a negative spiral. It might well be that the intervention aided students to organize and prioritize their studies during a crucial period. Those in the focus group indeed mentioned that participating in the intervention aided them in organizing their studies, and even their finances and social lives.

Limitations and Future Directions

Due to the rigorous controlled experimental design, the students and teachers received limited information about the intervention and none about its expected benefits. In the focus-group interviews, students mentioned that this made them somewhat skeptical about participating. They reported that integrating the intervention in the regular curriculum and having a mentor follow-up during regular coaching sessions would increase the positive effect. Some students remarked experiencing too little of a follow-up, except for the emails that they perceived as bothersome. Future studies could look into new innovative and personalized ways of organizing follow-ups, using, for example, a chatbot-coach that personally reminds and helps them to work on their goals (Dekker et al., 2020). In this way, reflective goal-setting interventions might yield a larger effect.

In line with the principles of replication with variation (Locke, 2015), the current study examined grit, engagement, resilience, and several modules of SRL, as mediators for the goal-setting intervention's effect to expand the related literature's generality. Given that these constructs did not prove to be a part of the core mechanisms in this context, future studies could also explore the mediating or moderating effects of other potential constructs, such as procrastination, or other variables that do not require self-reported measures, such as time spent on study and attendance. Further information on mediating constructs can aid the effective directed implementation in the right conditions and contexts.

Although we carefully considered all the aspects for implementation fidelity, we still cannot compare the results to other studies, as they did not report on these aspects and this study appears to be the first to examine implementation fidelity. Future studies should include transparent measures on the different aspects of implementation fidelity to compare and weigh its impact.

As mentioned before, several types of reflective goal-setting exercises are available. In the current study, they are categorized as the same type, because they share several working mechanisms that distinguish them from other types of goal-setting interventions, but there are differences (Schippers & Ziegler, 2019). These different versions might be altered and improved over time. Thus, future research should carefully document which version they use and describe its different individual parts. Finally, we found that gender, previous education, and ethnicity were strong predictors of academic performance and retention during the first year of college. Studying interventions that could potentially mitigate these negative effects, both in the first year and during the rest of the course, remains a relevant topic.

Conclusion

The teacher and business education students who received a reflective goal-setting intervention at the beginning of their study obtained significantly more course credits and dropped-out significantly less than those who received a control assignment. The treatment effects were independent of gender, ethnicity, or previous education, while growth in grit, resilience, engagement, or SRL did not mediate the direct effects. The intervention also did not significantly influence the students' general psychological well-being, and its implementation fidelity was moderate, suggesting that the latter may have suppressed the treatment's effects. These findings indicate that reflective goal-setting has a small, but significant effect on academic performance when it is implemented at a moderate level. As the intervention only took students two hours to complete and their gains equaled to 70 study hours (2.5 study credits) and 6 percentage point less dropout, this is good news for educators seeking to improve academic performance. A marginal addition of credits may especially make a difference for low performing students. Carefully implementing a scalable online intervention can also ensure that more students benefit from the intervention's positive effects.

CRediT roles

Izaak Dekker: Conceptualization; Investigation; Formal analysis; Data curation; Funding acquisition; Methodology; Project administration; Writing - original draft **Michaéla Schippers:** Resources; Writing - review & editing **Erik van Schooten:** Formal analysis; Methodology; Writing - review & editing

Acknowledgements

We wish to thank Ellen Klatter, Bettina de Jong, Monique de Bruijn-Smolders, Claudia Gomes, Dominique Morisano, Frans Spierings, Monique Volman, Martijn Meeter, Clementine Degener, Conrad Heilmann, Laurie Delnoij, Louise Elffers, Anders Schinkel, and Geert ten Dam for their helpful feedback.

We thank Justin Slabbekoorn, Mehmet Sahan, Janine Slootmaekers, Chanté Zuidgeest, Fleur Doolaard, Dian van Arkel, Anneke Delhaas, Els de Bock, Erwin van Braam, Annet Brugel, Henriett Duijs, Nina Adriaanse, Ishany Gaffar, Annemarie van Efferink, Emiel de Ruiter, Bert van Wijk, Lex Freund, and Aad Schippers for their aid in the execution of the experiment.

Funding sources

This research was partly supported with a PhD research grant awarded to the first author by an independent grant committee from the Rotterdam University of Applied Sciences.

References

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Greif Green, J., Hasking, P., Murray, E., Nock, M. K., Pinder-Amaker, S., Sampson. N. A., Stein, D. J., Vilagut, G., Zaslavsky, A. M., & Kessler, R. C. (2018). WHO world mental health surveys international college student project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology, 127*(7), 623-638. https://doi:10.1037/abn0000362

Azmitia, M., Sumabat-Estrada, G., Cheong, Y., & Covarrubias, R. (2018). "Dropping out is not an option": How educationally resilient first-generation students see the future. In C. R. Cooper & R. Seginer (Eds.), Navigating pathways in multicultural nations: Identities, future orientation, schooling, and careers. *New Directions for Child and Adolescent Development*, *160*, 89– 100. https://doi.org/10.1002/cad.20240

Bakker, A. B., Sanz Vergel, A. I., & Kuntze, J. (2014). Student engagement and performance: A weekly diary study on the role of openness. *Motivation and Emotion, 39*(1), 49-62. https://doi:10.1007/s11031-014-9422-5

Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology, 43*(8), 667-672. https://doi.org/10.1007/s00127-008-0345-x

Bettinger, E. P., & Baker, R. B. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis*, *36*(1), 3-19. https://doi.org/10.3102/0162373713500523

Bipp, T., Kleingeld, A., Van Den Tooren, M., & Schinkel, S. (2015). The effect of self-set grade goals and core self-evaluations on academic performance: A diary study. *Psychological Reports, 117*(3), 917-930. https://doi.org/10.2466/11.07.PR0.117c26z0

Bowman, N. A. H., Patrick, L., Denson, Nida, & Bronkema, R. (2015). Keep on truckin' or stay the course? Exploring grit dimensions as differential predictors of educational achievement, satisfaction, and intentions. *Social Psychological and Personality Science, 6*(6), 639–645. https://doi.org/10.1177/1948550615574300

Campbell-Sills, L., & Stein, M. B. (2007). Psychometric analysis and refinement of the Connor–Davidson resilience scale (CD-RISC): Validation of a 10-item measure of resilience. *Journal of Traumatic Stress*, *20,* 1019–1028. https://doi:10.1002/jts.20271

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation science*, *2*(1), 40. https://doi.org/10.1186/1748-5908-2-40

Catterall, J., Davis, J., & Yang, D. F. (2014). Facilitating the learning journey from vocational education and training to higher education. *Higher Education Research & Development*, 33(2), 242-255. https://doi.org/10.1080/07294360.2013.832156

Clark, D., Gill, D., Prowse, V., & Rush, M. (2019). Using goals to motivate college students: Theory and evidence from field experiments, *Review of Economics and Statistics*, 1-45. https://doi.org/10.1162/rest_a_00864

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety, 18*(2), 76-82. https://doi:10.1002/da.10113

Credé, M., & Niehorster, S. (2012). Adjustment to college as measured by the student adaptation to college questionnaire: A quantitative review of its structure and relationships with correlates and consequences. *Educational Psychology Review*, *24*, 133–165. https://doi.org/10.1007/s10648-011-9184-5

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23-45. https://doi.org/10.1016/S0272-7358(97)00043-3

De Bruijn-Smolders, M., (2017). *Self-regulated learning and academic performance; a study among freshmen*. [Doctoral dissertation, Erasmus University Rotterdam]. hdl.handle.net/1765/102845

Dekker, I., Schippers, M. C., De Jong, E. M., De Bruijn-Smolders, M., Alexiou, A., & Giesbers, B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Front. Psychol. 11:*1063. https://doi:10.3389/fpsyg.2020.01063

DeRosier, M. E., Frank, E., Schwartz, V., & Leary, K. A. (2013). The potential role of resilience education for preventing mental health problems for college students. *Psychiatric Annals*, *43*(12), 538-544. https://doi.org/10.3928/00485713-20131206-05

Dobronyi, C. R., Oreopoulos, P., & Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. *Journal of Research on Educational Effectiveness, 12*(1), 38-66. https://doi:10.1080/19345747.2018.1517849

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166-174. https://doi.org/10.1080/00223890802634290

Duncan, T., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist*, *40,* 117–128. https://doi.org/10.1207/s15326985ep4002_6

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*, *16*(8), 1123-1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327-350. https://doi.org/10.1007/s10464-008-9165-0

Epton, T., Currie, S., & Armitage, C. J. (2017). Unique effects of setting goals on behavior change: Systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*(12), 1182. https://doi.org/10.1037/ccp0000260

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention science*, *10*(2), 87–99. https://doi:10.1007/s11121-008-0109-6

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *Am. Psychol.* 54, 493–503. doi: 10.1037/0003-066X.54.7.493

Grossi, E., Groth, N., Mosconi, P., Cerutti, R., Pace, F., Compare, A., & Apolone, G. (2006). Development and validation of the short version of the Psychological General Well-Being Index (PGWB-S). *Health and Quality of Life Outcomes*, *4*, 88. https://doi:10.1186/1477-7525-4-88

Hodge, B., Wright, B., & Bennett, P. (2018). The role of grit in determining engagement and academic outcomes for university students. *Research in Higher Education*, *59*(4), 448-460. https://doi.org/10.1007/s11162-017-9474-y

Horowitz, E., Sorensen, N., Yoder, N., & Oyserman, D. (2018). Teachers can do it: Scalable identity-based motivation intervention in the classroom. *Contemporary Educational Psychology*, *54,* 12-28. https://doi.org/10.1016/j.cedpsych.2018.04.004

Hoyle, R. H., & Sherrill, M. R. (2006). Future orientation in the self-system: Possible selves, self-regulation, and behavior. *Journal of Personality*, *74*(6), 1673-1696. https://doi.org/10.1111/j.1467-6494.2006.00424.x

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. London: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55. https://doi.org/10.1080/10705519909540118

Jachimowicz, J. M., Wihler, A., Bailey, E. R., & Galinsky, A. D. (2018). Why grit requires perseverance and passion to positively predict performance. *Proceedings of the National Academy of Sciences*, *115*(40), 9980-9985. https://doi:10.1073/pnas.1803561115

Johnson, M. L., Taasoobshirazi, G., Kestler, J. L., & Cordova, J. R. (2015). Models and messengers of resilience: a theoretical model of college students' resilience, regulatory strategy use, and academic achievement. *Educational Psychology*, *35*(7), 869-885. https://doi.org/10.1080/01443410.2014.893560

Kleingeld, A., Van Mierlo, H., & Arends, L. (2011). The effect of goal setting on group performance: A meta-analysis. *Journal of Applied Psychology, 96*(6), 1289–1304. https://doi.org/10.1037/a0024315

Klug, H. J., & Maier, G. W. (2015). Linking goal progress and subjective well-being: A meta-analysis. *Journal of Happiness Studies*, *16*(1), 37-65. https://doi.org/10.1007/s10902-013-9493-0

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2007). *Piecing together the student success puzzle*: *Research, propositions, and recommendations* (ASHE Higher Education Report, Vol. 32). San Francisco, CA: Jossey-Bass.

Locke, E. A. (2015). Theory building, replication, and behavioral priming: Where do we need to go from here?. *Perspectives on Psychological Science*, *10*(3), 408-414. https://doi.org/10.1177/1745691614567231

Locke, E. A., & Latham, G. (2002). Building a practically useful theory of goal-setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705-717. https://doi.org/10.1037/0003-066X.57.9.705

Locke, E. A., & Latham, G. (2005). New directions in goal-setting theory. *Current directions in psychological science*, *15*(5), 265-268. https://doi.org/10.1111/j.1467-8721.2006.00449.x

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Martin, A. J., Bottrell, D., Armstrong, D., Mansour, M., Ungar, M., Liebenberg, L., & Collie, R.J. (2015). The role of resilience in assisting the educational connectedness of at-risk youth: A study of service users and non-users. *International Journal of Educational Research*, *74*, 1–12. https://doi.org/10.1016/j.ijer.2015.09.004

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean?. *American Psychologist*, *70*(6), 487. https://doi.org/10.1037/a0039400

Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes*, *39*(1), 52-83. https://doi.org/10.1016/0749-5978(87)90045-8

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, *95*(2), 255-264. https://doi.org/10.1037/a0018478

Morosanu, L., Handley, K., & O’Donovan, B. (2010). Seeking support: researching first-year students’ experiences of coping with academic life. *Higher Education Research & Development*, *29*(6), 665-678. https://doi.org/10.1080/07294360.2010.487200

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Organization for Economic Co-operation and Development. (2019). *Education at a glance 2019: OECD indicators*. Paris: OECD.

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology*, *91*(1), 188. https://doi.org/10.1037/0022-3514.91.1.188

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Pennebaker, J. W., & Chung, C. K. (2011). *Expressive writing: Connections to physical and mental health.* In H. S. Friedman (Ed.), *Oxford library of psychology. The Oxford handbook of health psychology* (p. 417–437). Oxford University Press.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33. https://doi.org/10.1037/0022-0663.82.1.33

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

Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly*, *44*, 152–170. https://doi.org/10.1177/001698620004400302

Respondek, L., Seufert, T., Hamm, J. M., & Nett, U. E. (2020). Linking changes in perceived academic control to university drop-out and university grades: A longitudinal approach. *Journal of Educational Psychology, 112*(5), 987-1002. http://dx.doi.org/10.1037/edu0000388

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414-430. https://doi.org/10.1177/1098214012441499

Ryff, C. D., & Singer, B. (1996). Psychological well-being: Meaning, measurement, and implications for psychotherapy research. *Psychotherapy and psychosomatics*, *65*(1), 14-23. https://doi.org/10.1159/000289026

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. *Journal of Organizational Behavior, 25*(3), 293-315. https://doi:10.1002/job.248

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement. *Educational and Psychological Measurement, 66*(4). https://doi.org/10.1177/0013164405282471

Schippers, M. C. (2017). *Ikigai: Reflection on life goals optimizes performance and happiness.* (Inaugural address). https://repub.eur.nl/pub/100484/27710_Oratie_Boekje_Micheala_Schippers_ONLINE.PDF

Schippers, M. C., Morisano, D., Locke, E. A., Scheepers, A. W. A., Latham, G. P., de Jong, E. M., (2020). Writing about personal goals and plans regardless of goal type boosts academic performance, *Contemporary Educational Psychology*, *60*, 101823. https://doi.org/10.1016/j.cedpsych.2019.101823

Schippers, M. C., Scheepers, W. A. & Peterson, J. B. (2015). A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. *Palgrave Communications, 1*(1), 1-12. https://doi:10.1057/palcomms.2015.14

Schippers, M. C., & Ziegler, N. (2019). Life crafting as a way to find purpose and meaning in life. *Frontiers in Psychology, 10,* 2778. https://doi.org/10.3389/fpsyg.2019.02778

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., Tomassetti, S., Nussbaum, A. D., & Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology, 104*(4), 591–618. https://doi.org/10.1037/a0031495

Sitzmann, T. & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go, *Psychological Bulletin*, *137*(3), 421–442. https://doi.org/10.1037/a0022777

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

The Netherlands Association of Universities of Applied Sciences 'Vereniging Hogescholen' (2020). *Cijfers en feiten databank.* https://www.vereniginghogescholen.nl/kennisbank/feiten-en-cijfers

Travers, C. J., Morisano, D., & Locke, E. A. (2015). Self-reflection, growth goals, and academic outcomes: A qualitative study. *British Journal of Educational Psychology*, *85*(2), 224-241. https://doi.org/10.1111/bjep.12059

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

Van Lent, M. (2019). Goal Setting, information, and goal revision: A field experiment. *German Economic Review, 20*(4), e949-e972. https://doi.org/10.1111/geer.12199

Van Lent, M., & Souverijn, M. (2020). Goal setting and raising the bar: A field experiment. *Journal of Behavioral and Experimental Economics*, *87*, 101570. https://doi.org/10.1016/j.socec.2020.101570

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 18;331*(6023):1447-1451. https://doi:10.1126/science.1198364

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a “chilly climate” transform women’s experience, relationships, and achievement in engineering. *Journal of Educational Psychology, 107*(2), 468. https://doi.org/10.1037/a0037461

Walton, G. M. (2014). The new science of wise psychological interventions. *Curr. Dir. Psychol. Sci.* 23, 73–82. doi: 10.1177/0963721413512856

Willcoxson, L. (2010). Factors affecting intention to leave in the first, second and third year of university studies: a semester-by-semester investigation. *High. Educ. Res. Dev.* 29, 623–639. doi: 10.1080/07294360.2010.501071

Windle, G., Bennett, K., & Noyes, J. (2011). A methodological review of resilience measurement scales. *Health and Quality of Life Outcomes*, *9*(8), 1–18. https://doi.org/10.1186/1477-7525-9-8