Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

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

Izaak Dekker, Michaéla Schippers & Erik Van Schooten

To cite this article: Izaak Dekker, Michaéla Schippers & Erik Van Schooten (2023): Reflective Goal-setting Improves Academic Performance in Teacher and Business Education: A Large-scale Field Experiment, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2023.2231440

To link to this article: https://doi.org/10.1080/19345747.2023.2231440

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2023.2231440

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

Izaak Dekker[a,b]*, Michaéla Schippers[a], and Erik Van Schooten[b,c]

[a]Department of Technology and Operations Management, Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands; [b]Research Centre Urban Talent, Rotterdam University of Applied Sciences, Rotterdam, the Netherlands; [c]Kohnstamm Institute, University of Amsterdam, Amsterdam, the Netherlands

ABSTRACT

A reflective goal-setting intervention could help students adjust to higher education, and improve their performance and well-being, as has been shown by small-scale and quasi-experimental studies conducted so far. However, a large experimental study found no effects, highlighting the importance of replication, and a better understanding of the mechanisms that explain when and why the intervention works. This replication study tested the effects of such a goal-setting intervention on the academic performance of 1,134 first-year business and teacher education students, with a randomized control trial. The treatment group earned significantly more course credits, and had a 15% lower risk of dropping out of college, compared to the control group. Contrary to the findings of previous studies, this study found no evidence that these effects are larger for men, or ethnic minorities. Additionally, we found no effect of the intervention on self-regulated learning, resilience, grit, engagement, or well-being.

ARTICLE HISTORY
Received 8 January 2021
Revised 15 May 2023
Accepted 17 June 2023

KEYWORDS
Goal setting; academic performance; teacher education; self-regulated learning; well-being

Introduction

More than a quarter of all students leave western higher education without obtaining the degree for which they enrolled (Organisation for Economic Co-operation and Development, 2019). The majority of the dropouts happen in the first year (Willcoxson, 2010), and ample evidence exists that this might be due to students having trouble adjusting to higher education (e.g., 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; Morosanu et al., 2010), and academic underachievement, manifested in low

CONTACT Izaak Dekker i.dekker@hva.nl Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands
*Present address: Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19345747.2023.2231440.

grades, reduced course credits, and high dropout rates (Kuh et al., 2007; Reis & McCoach, 2000).

Several rigorous experimental studies have reported that targeted interventions can improve the performance of at-risk students (e.g., Sherman et al., 2013; Walton & Cohen, 2011; Walton et al., 2015). However, universal interventions (Greenberg & Abenavoli, 2017) that target a broad student population are rarely tested with controlled experimental designs.

Morisano et al. (2010) trialed a goal-setting intervention that was low-cost, and potentially scalable, on a small sample of students. They reported that the intervention, in which students reflected on their desired futures, prioritized goals in line with goal-setting theory (Locke & Latham, 2002), and developed strategies, helped improve both GPA and student retention. Dobronyi et al. (2019) and Schippers et al. (2015, 2020) tested the effects of similar goal-setting interventions on larger samples. The studies by Schippers et al. (2015, 2020) used a quasi-experimental design on multiple European business school student groups ($n = 3{,}144$ and 2,928, respectively). In the 2015 study, the intervention enhanced retention rates and course credits by 22%, and the performance of male students and ethnic minorities improved the most (Schippers et al., 2015). The Schippers et al.'s (2020) study reported that participation was related to improved academic performance, regardless of the chosen goal (academic, social, etc.). Dobronyi et al.'s (2019) large field experiment with first-year students from a Canadian university ($n = 1{,}356$) compared the academic performance of a control group, an intervention group, and a group who received a shortened version of 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. These contradictory results call for more rigorous replication studies, in order to find out whether this intervention can reliably work across settings. Ideally, these replications should also extend our knowledge of the underlying mechanism, and analyze why and when the intervention works and for which students (Dekker & Meeter, 2022). Locke (2015) wrote that further development of the goal-setting theory calls for "replication with variation" (p. 410). Replication with variation entails searching for moderators and mediators, to inductively expand the theory's generality under different conditions.

Research indicated the existence of three different types of factors that could shed light on the mechanism behind the intervention. First, Schippers et al. (2015) suggested that gender and ethnicity moderated the effects, with the intervention being more effective for male students and ethnic minorities (demographics). Second, Schippers et al. (2020) found that the number of words the students wrote correlated with the intervention's effect, suggesting that the extent and earnestness of student participation, as well as their understanding of the purpose, might influence results (implementation fidelity). Third, psychological constructs could mediate the effect of goal-setting on performance, given that goal-setting aims to direct thoughts and behaviors (self-regulation, engagement, grit, and resilience) that subsequently lead to performance. Measuring the impact of goal-setting on both performance and psychological constructs simultaneously could make it possible to test whether the psychological constructs mediate the effect of goal-setting on performance.

Additionally, the Morisano et al. (2010) study, was replicated exclusively in business and economics courses. To generalize the results to different higher education domains and verify whether the intervention is domain-specific or not, replications should also include other types of university students.

Accordingly, we performed a "replication with variation" study, in which we measured the effect of the intervention of Morisano et al. (2010) on academic performance, and added variations with additional analyses, into the underlying mechanism. Specifically, we tested whether gender, ethnicity, or domain (teacher versus business education) moderated, and whether engagement, grit, SRL, and resilience mediated the treatment effect.

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 expect several psychological constructs to mediate the treatment effects on performance and well-being, and (3) the role of implementation fidelity in experimental studies and replications.

Literature Review and Hypothesis Development

Goal-setting Theory and Interventions

Scholars have extensively studied the goal-setting theory and its behavioral effects in organizational contexts, sports, healthcare, and education (Epton et al., 2017; Locke & Latham, 2002; Morisano, 2013). Goal-setting intervention studies began with establishing specific and ambitious goals in low-complexity contexts, such as setting targets for optimizing truck loads, e.g., trying to increase trees that can be loaded onto a truck. An increasing number of studies are modifying and applying goal-setting interventions to the first-year higher education environment, which is a highly complex context, given that the tasks, environment, and the high expected self-regulation are new for first-year students. Experimental goal-setting studies within this context have not yet been included in the goal-setting meta-analyses of Mento et al. (1987), Kleingeld et al. (2011), and Epton et al. (2017). Supplementary Appendix A Table A.1 offers an overview of all experimental studies examining the effect of goal-setting interventions on academic performance in higher education.

The literature offers three different experimentally tested types of goal-setting in the first year of 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., 2020; Van Lent, 2019; Van Lent & Souverijn, 2020). Van Lent and Souverijn (2020) performed a field experiment with 1,092 Dutch economics students and instructed a random subset of mentors to encourage students to set grade goals. Half of these 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 other goals in a short survey. These students did not perform better than the control group on their exams. Similarly, in their field experiment with 1,967 American microeconomics students, Clark et al. (2020) reported an insignificant increase in the performance of those who set grade goals. The evidence thus far indicates that "grade goal-setting" produces little to no positive effect on academic performance.

The second type of goal-setting intervention targets the specific tasks one wants to complete. Clark et al. (2020) conducted a field experiment with 2,004 American students

enrolled in a microeconomics course. The students randomly allocated to the treatment group were encouraged to set task goals (e.g., number of online practice exams they would complete before their final examination), while those in the control group received no goal-setting encouragement. Students in the treatment group reported significantly higher task completion levels and scored marginally higher on performance.

The third type of goal-setting asks students to write about their personal goals and develop a strategy to achieve these goals (Dobronyi et al., 2019; Latham & Brown, 2006; Morisano et al., 2010; Schippers et al., 2015, 2020; Travers et al., 2015). In a small-scale trial conducted with academically struggling students from a Canadian university ($n = 85$), Morisano et al. (2010) tested a version that combined expressive writing exercises (Pennebaker & Chung, 2011), implementation intentions (Gollwitzer, 1999), and goal-setting theory. The treatment group obtained a significantly higher GPA than the control group. Schippers et al. (2015, 2020) and Dobronyi et al. (2019) used a version that involved similar exercises but included negative scenarios (e.g., what will happen if you do not change your habits?). The appendix of the Morisano et al. (2010) study contains a step-by-step description of the intervention, along with its theoretical rationale. Dobronyi et al. (2019) enclosed a later and shortened version of this intervention in the appendix to their article. The reflective goal-setting method developed by Travers (as described in e.g., Travers et al., 2015), finally, provides a longer goal-setting process and covers different related topics. Although these different versions offer different experiences, they all provide guidance during the goal-setting process and stimulate reflection. For ease of reading this third category of goal-setting will here be referred to as "reflective goal-setting" interventions.

Morisano et al. (2010) reported significant positive effects of the treatments on the GPA, retention, and affect of the students. Schippers et al. (2015, 2020) found that there were significantly more course credits and lower drop-out rates among students in the cohorts that received the goal-setting intervention than students in control cohorts. Dobronyi et al. (2019), however, found no significant effects of the treatment on study grades, or retention, in a large-scale field experiment.

Grade, task, and reflective types of goal-setting interventions in higher education share a common ground in goal-setting theory, but they differ in how directed and extensive they are. Reflective goal-setting interventions seem a promising candidate for replication with variation because results thus far indicate both the largest potential effect as well as contradictory results. As Locke (2015) argued, employing the right moderators or mediators can expand goal-setting theory by improving our understanding of when it works and why. The chosen moderators, which may even be population dependent, may have caused varying effects in previous studies. Furthermore, these studies only included small samples of struggling students and large samples of business or economics students. Schippers et al. (2015) reported a moderating effect for gender and ethnicity: males and students from ethnic minorities benefited more. Therefore, we formulated the following hypotheses:

Hypothesis 1. Students in both business and teacher education, who have 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 (higher effects for males and ethnic minorities) will moderate the intervention's effect on course credits and dropout 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 and additions to the 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 prioritized 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 when students wrote about proximal intermediate goals, this induced an immediate increase in effort regulation, a form of self-regulatory behavior. The increase in effort was sustained through persistence and self-efficacy, and many reported that this had led to an upward spiral of subsequent engagement. This mechanism overlaps with several of Schippers' (2017) propositions. Given that a particular intervention may 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).

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, p. 422) noted that "one commonality across all the theories is that goal-setting triggers self-regulation." In practice, SRL manifests itself in higher levels of academic initiative, such as active class participation, fewer absences, and less misbehavior in class (Hoyle & Sherrill, 2006; Oyserman et al., 2006). These practical implications are why we expect SRL to improve engagement and academic performance (Sitzmann & Ely, 2011).

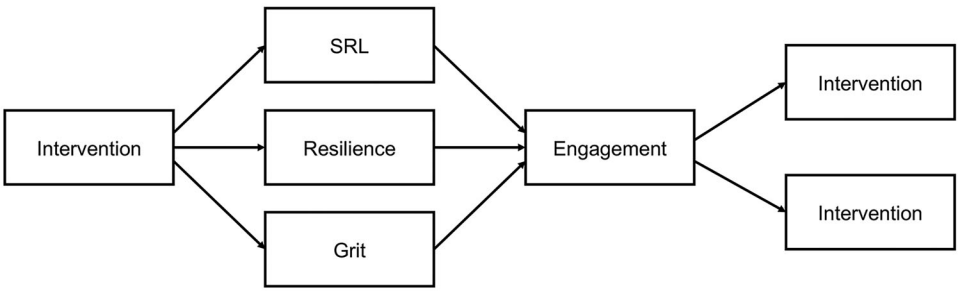

Figure 1. Mediating mechanisms between goal-setting intervention and outcomes. SRL or self-regulated learning is a multidimensional and modular construct (Pintrich et al., 1993). For this study, we used the modules of effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention. Adapted from Schippers (2017).

Setting goals and anticipating how one should act in trying situations, is expected to improve resilience, defined as the capacity to deal with adversity (Connor & Davidson, 2003). Resilience supports both academic performance and well-being (Johnson et al., 2015; Martin et al., 2015), and could 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) defined it as "perseverance and passion for long-term goals" (p. 1087); it can also predict academic performance and engagement (Bowman et al., 2015; Duckworth et al., 2007; Hodge et al., 2018).

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 behavior" (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. Travers et al. (2015) found that students who engaged in a 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., 2015). Accordingly, 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 psychological well-being has become an issue of concern in academia (Auerbach et al., 2018). Policymakers and scientists argue that many measures that aim at improving academic performance, do so, at the cost of students' psychological well-being. However, reflective goal-setting interventions aim to improve both academic performance, and psychological well-being, because they challenge students to set academic, social, and health-related goals (Schippers, 2017; Schippers & Ziegler, 2019). Having the right priorities and strategies should help students engage in activities that allow them to pursue their goals in a healthy way. In a meta-analysis Klug and Maier (2015) synthesized that successful goal pursuit is significantly related to subjective well-being (r = .43). We expect the engagement as a consequence of setting goals and persevered striving (through SRL, resilience, and grit) to lead to an increased general psychological 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 both business and teacher education, who have received a reflective goal-setting intervention at the start of their study, will report higher levels of SRL (effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention), resilience, grit, engagement, and general psychological 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 in both business and teacher education students.

Hypothesis 5. SRL, grit, resilience, and engagement will mediate the intervention's effect on course credits, dropout rates, and general psychological well-being.

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. The current study will explore the implementation fidelity of the intervention. The inconclusive results of previous studies could be a result of the differences in intervention implementation. Following 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 six 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 pollute the potential effects—this is a particular risk for certain elements in goal-setting interventions, considering the theory has been around for decades (Locke & Latham, 2002). *Dosage* refers to "how much" of the intervention was offered. The complete goal-setting intervention (as described in Morisano et al., 2010) is offered once, and students are instructed to reserve around 2.5 h to complete the assignment. *Adherence* refers to whether the treatment's parts were followed in the correct sequence. The goal-setting intervention was offered through an online tool, which did not allow students to change the sequence of their assignments. Students received access to the second part, only after the deadline for finishing part 1. *Quality of delivery* is successful when participants experience the main points as easy to process, true, and emerging naturally. The main points in this study refer to the text-prompts before the assignments. These prompts emphasized the importance of explicating the desired future, of formulating and prioritizing goals, and goal achievement plans. *Student responsiveness* involves students' responses to the adherence and quality of delivery. The right dosage, adherence, and quality of delivery should lead to engaged attention and productivity in the students. This could be estimated with completion rates, time spent on the intervention, or output variables, such as the number of written words. Students are explicitly encouraged in the intervention to "keep on writing" and Schippers et al. (2020) found that the number of written words was related to an increase in academic performance, even when controlling for the number of stages students completed and the quality of their goal achievement plans. *Fidelity of receipt* refers to the degree to which students internalize the main points that the intervention aims to communicate. Did this intervention induce them to think or act differently? Did they (consciously) allocate time differently, or change their behavior? These dimensions require attention, as they provide conditional information expected to influence the results of an experimental study (Durlak, 2015; Durlak & DuPre, 2008).

Table 1. Sample characteristics of the freshmen per faculty and condition.

	Business		Education		Treatment		Control	
	N	%	*N*	%	*N*	%	*N*	%
Participants	302		832		571		563	
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

Method

Participants

The sample consisted of 1,134 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. As part of our selection procedure, we checked whether the curricula of the courses of study contained (parts of) a similar intervention and whether faculty members used similar techniques (reflective writing, mental contrasting, goal setting, implementation intentions) to determine program differentiation. None of the involved 36 faculty members reported already using parts of the intervention or a similar intervention with their students.

Seven percent of the student population in applied sciences universities in The Netherlands followed a preparatory scientific track (students from this track constitute the majority of students in the studies by Schippers et al., 2015, 2020), 43% followed a general academic track in high school, 31% had a vocational education background, and 19% used an admission test or an eligible international degree (The Netherlands Association of Universities of Applied Sciences 'Vereniging Hogescholen', 2020). We controlled for this sample characteristic in our analysis because previous education here is strongly related to central exam scores (similar to SAT scores) (Van der Zanden et al., 2018).

The sample was taken from teacher education and business studies faculties. The engineering and medical faculties were also invited, but they declined to participate. All of the courses of study within the two participating faculties were invited. Within the business faculty, 2 out of the 5 courses of study participated with all of 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.

The internal review board of the researchers' affiliated university approved the experiment before its execution. All the participants signed informed consent forms beforehand. The procedure in the data management plan ensured the use of pseudonyms before datasets were merged, and anonymous and save storage afterward. Directly after

[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 obtain a threshold amount of 42 course credits in the first year and all the required course credits of the first year (60) within two years.

the experiment, all the participants were debriefed and received a book about classroom management (teacher education) or a business journal (business education). The debriefing included information about the design of the experiment and the two assignments. Students who had received the control condition were offered the reflective goal-setting intervention after the experiment.

Research Design

We conducted a double-blind student-level randomized field experiment at the beginning of the 2018–2019 academic year to test hypotheses 1–5. The intervention was a Dutch translation of the goal-setting intervention, described in Morisano et al. (2010), and was tested on Dutch students using the think-aloud method. Minor changes were made to increase understandability. This version was translated back to English and then corrected by the first author of the original version. The students were randomly assigned to a treatment or control group, and were told not to communicate with other students about the assignments. External surveillants treated the interventions similar to an examination, and monitored whether the students did not communicate during the assignments. The control group participants received a control condition that was similar to the one used by Morisano et al. (2010). These web-based tasks included questionnaires about personality and vocational interests, and writing assignments about positive past experiences, which students were asked to answer objectively, without emotional expression. The two parts of the intervention or control assignments were sent to the students by e-mail, who completed them individually in computer rooms at the university. Part 1 was made in the first week of college; part 2 was scheduled three to seven days later. Both parts were made within the introduction weeks, but before the courses began. The intervention was not tied to a specific course. Students had 3 h to complete part 1 and three h to complete part 2. The median time spent on the intervention was 36 min on the first and 51 min on the second. This did not differ significantly from the median time spent on the control condition (37 and 48 min, respectively). To measure the effect of receiving the intervention on SRL, grit, resilience, and engagement we used a baseline survey and two repeated measures after the intervention (T0, T1, T2). We conducted T0 survey at the start of the year and one to three days before the intervention, T1 survey two weeks before the end of the first semester, and T2 survey two weeks before the end of the second. We measured the effects of receiving the intervention on academic performance in accumulated course credits and study status (dropping out of the course of study or not) at T1 (+ 2 weeks) and T2 (+ 2 weeks) with the help of administrative data.

Final Analytical Sample

In total, 1,073 (95%) students started the assignments and 942 (83%) students finished both parts of the treatment or control assignment. The percentage of students who finished both parts in the treatment group (82,3%) and the control condition group (83,8%) did not differ significantly. According to the teaching staff of the participating courses, 83% completion rate for two assignments is similar to normal participation in

course assignments during the first weeks of college. We took the participation rate into account in our assessment of the treatment fidelity.

The final dataset, with 1,134 students, contained no missing data for the main dependent variables: course credits and dropout (enrollment status). Nor did the dataset contain missing demographic data (gender, ethnic background, previous education). However, the survey dataset, gathered for the mediation analyses, did contain missing data. 1,060 students 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 article 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). Assignment to treatment group, gender, or previous education did not significantly correlate with responding to one of the three surveys (T0, T1, T2). Ethnic minority did significantly correlate with non-response, although the difference was relatively small. The strength of the correlation between being an ethnic minority and finishing the survey was $r = (1{,}133)\ .19$, $p < .001$ for survey T0, $r = (1{,}133)\ .10$, $p < .05$ for survey T1, and $r = (1{,}133)\ .08$, $p < .05$ for survey T2. After screening, we treated those who had the same answer to all questions or did not clearly write their identification number in the analog T2 survey (8 cases) as missing. This totaled 104 cases in the T0, 21 cases in the T1, and 23 cases in T2 survey. The final dataset contained 1,134 cases with demographic data, course credits, and dropout status, of whom 956 had T0 survey scores, 483 had T1 scores, and 630 had T2 scores.

We calculated power with the G*Power 3 program Power analyses for testing hypotheses 1–5 can be found in the Supplementary Appendix (Figures B.1 and B.2). We corrected the sample size for multilevel structure (13 clusters with an average n of 87) according to Hox et al. (2018, p. 223) with the following formula:

$$\text{Effective}n = n/[1 + (\text{mean cluster size} - 1) * \text{Intraclass correlation coefficient}]$$

In all instances the effective n was large enough to find at least an effect size of $f^2 = 0.11$ with a power level of 0.90, and $f^2 = 0.08$ with a power level of 0.80.

At the end of the year, we selected 20 students randomly from the treatment group to partake in qualitative focus groups for evaluation purposes; 14 attended. We asked them to evaluate the two parts and describe if they had learned anything and had applied what they had learned beyond the intervention. All courses of study, except preservice 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

Measures and Instrumentation

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 equals 28 study hours). Thus,

we measured academic performance by tracking the participants' obtained ECTS credits and dropout rates supplied by the university administration.

The following standardized scales measured SRL (effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention), resilience, grit, engagement, and general psychological well-being (PGWB). The modular subscales for effort regulation, self-efficacy, intrinsic goal orientation, metacognition, and attention stem from the Motivated Strategies for Learning Questionnaire (MSLQ) (Duncan & McKeachie, 2005; Pintrich et al., 1993). Both subscale selection and the Dutch translation were based on a study that had tested the instruments on Dutch professional higher education students (De Bruijn-Smolders, 2017). We measured resilience with a translated 10-item Connor-Davidson Resilience Scale (Campbell-Sills & Stein, 2007), grit with a translated 10-item GRIT-S scale (Duckworth & Quinn, 2009), and well-being with a translated six-item PGWB-S 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 closely related to having a goal or purpose (Klug & Maier, 2015). This could cloud conceptual clarity and make the correlation between goal pursuit and subjective well-being spurious. The PGWB-S 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.

Six months 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) and made minor language adjustments to replace complicated words and ambiguous formulations.

After we collected all data for the experiment, we tested the validity of the used questionnaire with confirmatory factor analyses in Mplus (Muthén & Muthén, 1998–2006) for each measurement moment (T0, T1, T2). All items loaded significantly on the factor they were supposed to measure, and we also did not find perfect correlations between factors. The overall validity of the instruments was reasonable, the different constructs showed good discriminant validity, and the reliabilities ranged from moderate (Cronbach's alpha = 0.65) to robust (0.86) (Taber, 2018). Tables B.3 and B.4 in the Supplementary Appendix provide a detailed description of the CFA outcomes and reliability measures.

Monitoring Implementation Fidelity

We measured dosage fidelity by tracking the completion rates and the number of words that students wrote. 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 to 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). We recorded and transcribed the two focus group conversations, and followed a protocol to ensure that we evaluated all parts of the intervention, student 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,

Table 2. Baseline balance checks with administrative and survey data.

	Control sample mean (*SD*)	Difference with treatment group (*SE*)	χ^2 or t-value (*df*)	*p* Value	*N*
Male[a]	0.49 (0.50)	0.02 (0.02)	0.58 (1)	.45	1,134
Ethnic min. background[a]	0.30 (0.46)	0.01 (0.02)	0.01 (1)	.92	1,134
Vocational background[a]	0.28 (0.45)	−0.01 (0.05)	0.01 (1)	.94	1,134
Course of study[a]	N.A.	N.A.	12.75 (12)	.39	1,134
GPA high schoolb	6.50 (0.44)	−0.48 (0.24)	−1.56 (700)	.12	701
T0 effort regulation	3.73 (0.52)	0.05 (0.03)	1.47 (959)	.14	960
T0 self-efficacy	3.92 (0.56)	0.01 (0.03)	−0.14 (957)	.89	958
T0 intrinsic g. orient.	4.21 (0.50)	0.05 (0.02)	1.43 (955)	.15	956
T0 metacognition	3.42 (0.62)	0.03 (0.03)	0.67 (947.23)	.50	952
T0 attention	3.46 (0.67)	0.05 (0.03)	1.06 (948)	.29	949
T0 resilience	3.93 (0.48)	0.00 (0.03)	0.01 (948.93)	.99	956
T0 grit	3.65 (0.52)	0.05 (0.03)	1.37 (959)	.17	960
T0 engagement	3.32 (0.66)	0.01 (0.03)	0.34 (955)	.73	956
T0 well-being	4.55 (0.73)	−0.04 (0.03)	−0.75 (955)	.46	956

[a]Tested by means of χ^2 since variable is dichotomous. df = degrees of freedom.
[b]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 needing a GPA score. This is why students with a vocational background are missing on this variable.

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 randomization success. This involved ensuring no significant difference between the control and treatment groups before the intervention (T0) in dependent variables (SRL, grit, resilience, engagement, and well-being), demographics, and high school GPA (previous performance is a strong predictor of future performance). As Levene's test indicated unequal variances for metacognition [F = (949) 4.37, $p = .04$] and resilience [F = (949) 5.86, $p = .02$], we adjusted the degrees of freedom to compensate for the imperfect t-distribution (Table 2). T0 baseline survey scores showed no significant variable differences between the treatment and control groups (Table 2), indicating successful randomization.

Estimating Treatment Effects on Course Credits

The sample consisted of students that are nested in 13 courses of study. The courses are nested in 2 faculties. Students from the same course or faculty tend to be similar to each other because of selection processes. This could violate the assumption of independence of observations on which standard statistical tests rely and lead to incorrect estimations of standard errors. In these cases multilevel models offer a suitable method to correctly estimate standard errors (Hox et al., 2018). On the faculty level we only had two clusters. When a level contains less than 10 clusters, it is preferrable to add the clusters (faculty in our case) as a covariate to the model (Snijders & Bosker, 2012). For each dependent variable, we fitted multilevel models in which we measured significant model fit improvement by testing whether adding variables led to a significant fit improvement in deviance (−2*loglikelihood). The difference in deviance has a χ^2 distribution, with the difference in the number of parameters estimated in both nested models as degrees of freedom. Effect sizes are calculated as the proportions of explained

variance, both for total variance and each of the variances in the random part of the model. Using an RCT as the study design, the condition's (intervention) effects on T1 or T2 reflect the effects of receiving the reflective goal-setting intervention. We analyzed the effects of the intervention, with and without controlling for domain, gender, ethnicity, and previous education. In equation 1.1, Y_{ij} is course credits for student $_i$ in course $_j$ at T1 or T2. The intervention (Tx_{1ij}) is offered to individual students within groups. Previous education (X_{2ij}), ethnicity (X_{3ij}), gender (X_{4ij}), and faculty (Z_{1j}), are included as control variables, with random error at the student (e_{ij}) and course level (μ_{0j}). μ_{0j} is the random effect for cluster and e_{ij} is the student residual. In equation 1.2 we added the interaction effect for the intervention with previous education ($Tx_{1ij}X_{2ij}$). We then replaced interaction between the treatment and previous education with interaction between the treatment and (respectively) ethnicity, gender, and domain, to test if they moderated the treatment effect. For a histogram of the credits in the treatment and control group at T1 and T2 see Supplementary Appendix Figures B.3 and B.4.

$$Y_{ij} = \Upsilon_{00} + \Upsilon_{10}Tx_{1ij} + \Upsilon_{20}X_{2ij} + \Upsilon_{30}X_{3ij} + \Upsilon_{40}X_{4ij} + \Upsilon_{01}Z_{1j} + \mu_{0j} + e_{ij} \quad (1.1)$$

$$Y_{ij} = \Upsilon_{00} + \Upsilon_{10}Tx_{1ij} + \Upsilon_{20}X_{2ij} + \Upsilon_{30}X_{3ij} + \Upsilon_{40}X_{4ij} + \Upsilon_{01}Z_{1j} + \Upsilon_{50}Tx_{1ij}X_{2ij} + \mu_{0j} + e_{ij} \quad (1.2)$$

Estimating Treatment Effects on Dropout Rates

As dropping out of a course of study is a binary variable (1 = dropout, 0 = retained), we used multilevel logistic regression analyses with a logit model in MLwiN for this dependent variable (Rasbash et al., 2020). 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). We fitted models with and without controlling for gender, ethnicity, and previous education (hypothesis 2). We used Wald tests to test significance and calculated the relative risk reduction and "number needed to treat" as an indication of the independent variables' effects (Schechtman, 2002). The university where we conducted this study required students to obtain at least 42 course credits, in the first year, to continue studying. Therefore, we also tested whether the treatment group contained more students with at least 42 credits, using the same logistic regression analyses.

Estimating Treatment Effects on Psychological Variables

The intervention's effect on the psychological variables (hypothesis 3) was estimated with multilevel regression analyses in MLwiN. In the analyses, we estimated the treatment's effect on the different psychological variables at T1 and T2. We fitted the same models as we did to measure effects on study credits, with the addition of a covariate for baseline scores on the psychological variable. When no direct effect was found, we could also exclude a mediated effect (Hypothesis 5).

Implementation Fidelity

We assessed implementation fidelity using Horowitz et al.'s (2018) six categories. All the students in the treatment group were offered identical web-based interventions, one time each (dosage). The 87 min that the median of students spent on the intervention

was lower than the minimum of 240 min that students (had to) spent on the intervention in the Morisano et al. (2010) trial. 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 (adherence). We secured a stable quality of delivery, by distributing an identical intervention online, and controlled the conditions in which the students made the assignments in surveilled computer classrooms. 82 percent of the students in the treatment group finished both assignments. On average they wrote 1,134 words ($SD = 671$). The items that measured responsiveness indicate that 69.9% of the participants in the treatment condition, who completed 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 [$r = (941)\ 0.36$, $p < .001$].

In the focus groups, two students reported 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 the intervention had influenced their behavior, and three of them noted its influence in other domains as well. One student said 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, as other researchers reported (Walton & Cohen, 2011). However, some remembered it later during a 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 admitted that 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 help them internalize and utilize the intervention throughout the year. As one student put it: "One's teacher or coach should recall the intervention one period later. […] What about your goals now?" Asked about email reminders, the students reported that they already received too many emails and it would be an extra burden. These results indicate moderate implementation quality and we expect to find a (suboptimal) effect of the intervention.

Results

Treatment Effects on Academic Performance

Students received an average of 17.24 course credits in the first semester (T1). Those in the treatment group, on average, earned 1.08 [95% CI (0.10, 2.06)] course credits more than their peers in the control group during the first semester, which is a significant difference (Table 3, models 2 and 3). At the end of the first year (T2), the students earned an average of 42 course credits. Students assigned to the treatment group earned 2.70 [CI 95% (0.22, 5.19)] credits more ($p < .05$) than their peers in the control group (Table 4, model 2 and 3). The intervention explained 0.41% of the variation in credits at the student level after one semester and 0.42% after one year (both are comparable to a Cohen's d of 0.13).

Table 3. Multilevel regression analyses of treatment effects on course credits after one semester with control variables and moderator analyses.

		Course credits						
Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
				Fixed effects				
Intercept	γ_{00}	16.64 (0.93)	26.07 (1.34)	25.57 (1.37)	25.63 (1.37)	25.69 (1.37)	25.84 (1.37)	25.21 (1.42)
Intervention (= 1)	γ_{10}	1.02 (0.53)		1.08* (0.50)	0.93 (0.59)	0.78 (0.60)	0.47 (0.70)	1.93* (0.97)
Vocational b. (= 1)	γ_{20}		−3.70*** (0.60)	−3.71*** (0.60)	−3.99*** (0.82)	−3.68*** (0.59)	−3.69*** (0.59)	−3.73*** (0.59)
Ethnic min. (= 1)	γ_{30}		−3.58*** (0.59)	−3.60*** (0.58)	−3.59*** (0.58)	−4.09*** (0.81)	−3.61*** (0.58)	−3.62*** (0.58)
Male (= 1)	γ_{40}		−3.31*** (0.54)	−3.30*** (0.54)	−3.29*** (0.54)	−3.30*** (0.54)	−3.93*** (0.74)	−3.31*** (0.54)
Teacher ed. (= 1)	γ_{01}		−6.24*** (1.42)	−6.29*** (1.41)	−6.27*** (1.40)	−6.26*** (1.40)	−6.24*** (1.39)	−5.75*** (1.51)
Intervention*voc.	γ_{50}				0.55 (1.13)			
Intervention*ethnic.	γ_{60}					0.97 (1.09)		
Intervention*male	γ_{70}						1.26 (1.01)	
Intervention*teach.	γ_{11}							−1.16 (1.13)
				Random effects				
Student variance	e_{0ij}	77.80 (3.29)	70.88 (2.99)	70.59 (2.98)	70.58 (2.98)	70.55 (2.98)	70.51 (2.98)	70.53 (2.98)
Course variance	μ_{0j}	8.98 (4.03)	2.52 (1.40)	2.48 (1.39)	2.47 (1.38)	2.45 (1.37)	2.42 (1.36)	2.50 (1.39)
Total variance	$e_{0ij} + \mu_{0j}$	86.78	73.40	73.07	73.05	73.00	72.93	73.03
% expl. var. student level		0.34		0.41				
% expl. var. course level		0.16		1.59				
% expl. var. total		0.32		0.45				
			Goodness of fit					
Deviance		8,184.54	8,066.71	8,062.10	8,061.86	8,061.31	8,060.54	8,061.06
Model of reference		Model 0a	Model 0	Model 2	Model 3	Model 3	Model 3	Model 3
$\Delta\ \chi^2$				$\Delta\ \chi^2_{(1)} = 4.61$	$\Delta\ \chi^2_{(1)} = 0.24$	$\Delta\ \chi^2_{(1)} = 0.79$	$\Delta\ \chi^2_{(1)} = 1.56$	$\Delta\ \chi^2_{(1)} = 1.04$
p Value				$p < .05$	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.

Note: Standard errors are presented in parentheses.
[a]Model 0 is the empty model without the intervention, it is left out of this table for clarity.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4. Regression analyses of treatment effects on course credits after one year with control variables and moderator analyses.

		Course credits						
Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
				Fixed effects				
Intercept	γ_{00}	40.50 (1.16)	53.31 (2.07)	52.06 (2.18)	52.43 (2.19)	52.30 (2.20)	53.05 (2.25)	51.60 (2.36)
Intervention (= 1)	γ_{10}	2.70* (1.34)		2.70* (1.27)	1.73 (1.50)	2.14 (1.53)	0.64 (1.77)	3.78 (2.47)
Vocational backg. (= 1)	γ_{20}		−10.15*** (2.07)	−10.15*** (1.50)	−11.92*** (2.08)	−10.11*** (1.50)	−10.10*** (1.49)	−10.18*** (1.50)
Ethnic minority b. (= 1)	γ_{30}		−7.11*** (1.47)	−7.13*** (1.47)	−7.06*** (1.47)	−8.06*** (2.04)	−7.18*** (1.47)	−7.16*** (1.47)
Male (= 1)	γ_{40}		−8.15*** (1.36)	−8.12*** (1.36)	−8.09*** (1.36)	−8.13*** (1.36)	−10.27*** (1.86)	−8.14*** (1.36)
Teacher education (= 1)	γ_{01}		−3.21 (2.01)	−3.35 (2.04)	−3.23 (2.02)	−3.29 (2.02)	−3.27 (2.02)	−2.65 (2.45)
Intervention*voc.	γ_{50}				3.50 (2.85)			
Intervention*ethn.	γ_{60}					1.84 (2.77)		
Intervention*male	γ_{70}						4.30 (2.55)	
Intervention*teach.	γ_{11}							−1.47 (2.89)
					Random effects			
Student variance	e_{0ij}	502.44 (21.21)	459.89 (19.40)	457.95 (19.32)	457.42 (19.30)	457.83 (19.32)	456.85 (19.28)	457.85 (19.32)
Course variance	μ_{0j}	4.71 (4.30)	2.31 (2.96)	2.50 (3.05)	2.38 (2.99)	2.40 (3.00)	2.42 (3.01)	2.49 (3.05)
Total variance	$e_{0ij} + \mu_{0j}$	507.15	462.20	460.45	459.80	460.23	459.27	460.34
% expl. var. student level		0.35		0.42				
% expl. var. course level		1.73		–				
% expl. var. total level		0.37		0.38				
			Goodness of fit					
Deviance		10,278.21	10,175.13	10,170.64	10,169.14	10,170.20	10,167.80	10,170.38
Model of reference		Model 0a	Model 0	Model 2	Model 3	Model 3	Model 3	Model 3
$\Delta\ \chi^2_{(df)}$				$\Delta\ \chi^2_{(1)} = 4.49$	$\Delta\ \chi^2_{(1)} = 1.50$	$\Delta\ \chi^2_{(1)} = 0.44$	$\chi^2_{(1)} = 2.84$	$\chi^2_{(1)} = 0.26$
p Value				$p < .05$	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.

Note. Standard errors are presented in parentheses.
[a]Model 0 is the empty model without the intervention, it is left out of this table for clarity. Student $n = 1{,}134$; course of study $n = 13$.
*$p < .05$. **$p < .01$. ***$p < .001$.

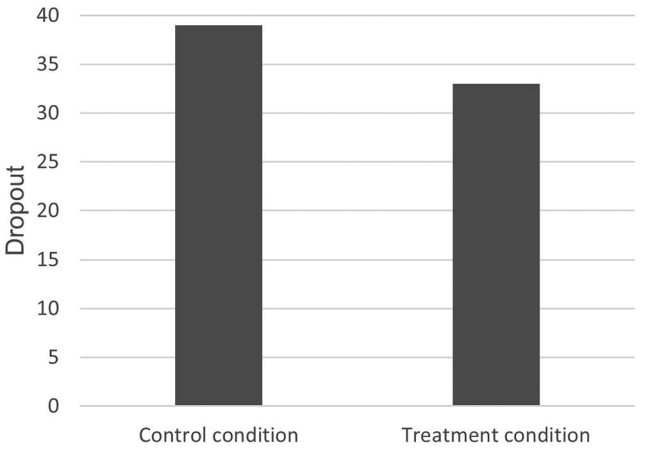

Figure 2. Treatment effects on dropout.

The intervention, on average, cost students less than 2 h, while 2.70 study credits correspond to 75.6 study hours. Kraft (2020) proposed taking scalability and costs into account when interpreting effect sizes from experimental studies as small, medium, or large. Given that the intervention can be sent to any number of students and requires little time of the teaching staff or university funding, it can be considered low-cost and scalable. According to Kraft (2020), an effect of .13 standard deviation "should be considered large and impressive when they arise from large-scale field experiments that are pre-registered and examine broad achievement measures" (p. 248).

With respect to dropout rates, the results were similar: 39% of all students in the control group dropped out of their course of study during the first year, compared to 33% in the treatment group (Figure 2). The logistic regression shows that the intervention significantly predicts dropout rates before and after controlling for previous education, gender, ethnicity, and domain (Table 5, models 1–3). The relative risk reduction of the intervention is 15.17% [95% CI (1.00, 27.31)], which stands for the reduced risk of bad outcomes relative to the control group. Therefore, Hypothesis 1 is accepted.

To find out whether these effects might be due to less students obtaining zero credits and/or more students obtaining more than 42 credits (the threshold for being allowed to stay enrolled) we ran additional analyses. The percentage of students who obtained more than 42 course credits in the treatment group was 68.3%, compared to 62% in the control group ($p = .03$, $r^2 = 0.01$) (Table B.5.1, Supplementary Appendix). Only 4.9% of the students in the treatment group obtained zero credits, compared to 8.9% in the control group ($p = .02$, $r^2 = 0.01$) (Table B.5.2, Supplementary Appendix). It, therefore, seems that the intervention both decreased the percentage of students who obtained no credits, and increased the percentage of students who obtained more than the required 42 credits.

Subgroups

To test whether the intervention works better for subgroups, as determined by Schippers et al. (2015), we added the interaction effects between the intervention and

Table 5. Multilevel logistic regression – effects on dropping out.

Effect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Constant	−0.44 (0.09)	−1.23 (0.14)	−1.24 (0.19)	−1.23 (0.19)	−1.31 (0.15)	−1.24 (0.20)
Intervention	−0.25* (0.12)	−0.26* (0.13)	−0.19 (0.16)	−0.22 (0.16)	−0.16 (0.19)	−0.19 (0.29)
Vocational background		0.83*** (0.15)	0.93*** (0.20)	0.82*** (0.15)	0.82*** (0.15)	0.82*** (0.15)
Ethnic minority backg.		0.64*** (0.14)	0.64*** (0.15)	0.70*** (0.20)	0.65*** (0.15)	0.64*** (0.15)
Male		0.75*** (0.14)	0.74*** (0.14)	0.75*** (0.14)	0.84*** (0.19)	0.74*** (0.14)
Teacher education		−0.05 (0.15)	−0.06 (0.15)	−0.05 (0.15)	−0.05 (0.15)	−0.00 (0.20)
Intervention*vocational			−0.22 (0.28)			
Intervention*ethnic min.				−0.11 (0.27)		
Intervention*male					−0.19 (0.26)	
Intervention*teacher ed.						−0.10 (0.29)
Student variancea	3.29	3.29	3.29	3.29	3.29	3.29
Course variance	0.01 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Goodness of fit						
Deviance	1,482.96	1,395.82	1,395.20	1,395.64	1,395.26	1,395.69
Model of reference	Model 0b	Model 0	Model 2	Model 2	Model 2	Model 2
$\Delta\ \chi^2_{(df)}$	$\chi^2_{(1)} = 4.38$	$\chi^2_{(5)} = 91.53$	$\chi^2_{(1)} = 0.62$	$\chi^2_{(1)} = 0.18$	$\chi^2_{(1)} = 0.56$	$\chi^2_{(1)} = 0.13$
p Value	$p < .05$	$p < .001$	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.	$p =$ n.s.

Note: Dependent variable is "Dropout": higher regression coefficients indicate higher chances of dropping out.
[a]Variance at the lowest level is by default 3.29 with multilevel logistic regression.
[b]Model 0 is the empty model without the intervention, it is left out of this table for clarity. Student $n = 1{,}134$.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6. Treatment effects of intervention on variables after one (T1) and two semesters (T2).

Variable	B (SE)	95% CI	$\Delta\ \chi^2$	n	*p*	r^2
Attention T1	0.07 (0.05)	−0.03, 0.17	1.46	442	.23	0.003
Attention T2	0.00 (0.05)	−0.10, 0.10	0.00	580	1.00	0.000
Effort T1	0.06 (0.05)	−0.04, 0.16	1.13	450	.29	0.003
Effort T2	−0.02 (0.05)	−0.12, 0.10	0.22	585	.64	0.000
Intrinsic or. T1	−0.04 (0.05)	−0.14, 0.06	0.58	444	.45	0.000
Intrinsic or. T2	−0.03 (0.04)	−0.11, 0.05	0.72	583	.40	0.000
Metacogn. T1	0.01 (0.05)	−0.11, 0.09	0.06	444	.81	0.000
Metacogn. T2	−0.05 (0.05)	−0.15, 0.05	1.01	583	.35	0.003
Self-effic. T1	0.03 (0.06)	−0.88, 0.15	0.33	448	.57	0.000
Self-effic. T2	−0.04 (0.04)	−0.12, 0.04	0.74	583	.39	0.000
Engagement T1	−0.01 (0.06)	−0.13, 0.11	0.03	438	.86	0.000
Engagement T2	−0.02 (0.05)	−0.12, 0.10	0.27	585	.60	0.000
Grit T1	0.02 (0.04)	−0.58, 0.10	0.35	454	.55	0.000
Grit T2	−0.00 (0.04)	−0.08, 0.08	0.01	585	.92	0.000
Resilience T1	0.06 (0.05)	−0.04, 0.16	0.39	442	.53	0.000
Resilience T2	0.00 (0.05)	−0.10, 0.10	2.59	580	.11	0.005
Well-being T1	0.04 (0.07)	−0.10, 0.18	0.68	433	.41	0.000
Well-being T2	−0.06 (0.07)	−0.20, 0.08	0.93	580	.33	0.000

vocational background, ethnic minority, male, and domain, respectively, to a model. None of these moderator effects proved a significant improvement (Tables 3 and 4, models 4–7). The confidence intervals of the interaction coefficient for teacher education [CI 95% (−7.13, 5.66)], vocational background [CI 95% (−5.59, 9.09)], ethnic minority background [CI 95% (−5.43, 7.27)], and being male [CI 95% (−5.00, 9.30)] at T2 show that both negative, positive, and no differences in credits, are all consistent with the data. These results are, therefore, still inconclusive. We did not find sufficient evidence to accept or reject hypothesis 2.

Exploring the Mechanism

Hypothesis 3 predicted that students in the treatment group score significantly higher on SRL, resilience, grit, engagement, and well-being. Contrary our predictions, the treatment group did not score significantly higher on any of these variables after one or two semesters. Table 6 summarizes the results from these analyses. We tested whether using repeated measure analyses with slightly higher power led to different results. The findings from these analyses were consistent.

Both students in the treatment and control group showed a significant decline in well-being, engagement, and SRL at the end of the two terms of the first year. This decline is typical for the first year and end of term (Corpus et al., 2020; Hudig et al., 2022; Wang et al., 2014).

Although a moderator effect without a direct effect is unlikely, it is still possible. We continued to test whether significant treatment effects could be found with gender, domain, previous education, and ethnicity as moderators (hypothesis 4). After two semesters, students from an ethnic minority in the treatment group showed a significantly lower drop in well-being ($p < .01$, $r^2 = 0.014$) and higher metacognition ($p < .05$, $r^2 = 0.007$) than those in the control group (B.14.2 and B.12.2, Supplementary Appendix). Both teacher education students ($p < .05$, $r^2 = 0.031$) and students with a vocational education background students ($p < .01$, $r^2 = 0.013$) in the treatment group reported significantly higher effort regulation after one semester. Students with a

vocational education background scored higher on self-efficacy ($p < .05$, $r^2 = 0.009$) and engagement ($p < .05$, $r^2 = 0.013$) at T1 than their peers in the control group (Table B.6.1, Supplementary Appendix). Teacher education students in the treatment group reported higher grit at T2 ($p < .05$, $r^2 = 0.006$) (Table B.8.2., Supplementary Appendix). However, we performed 72 of these moderator tests, which results in a high family-wise error rate. After using the Bonferroni correction (which led to an alpha threshold of 0.0007) none of the moderator analyses remained significant. Therefore, we did not find sufficient evidence to support 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, we did not find a direct effect of the intervention on well-being, hence no mediation could occur, rejecting Hypothesis 5. Additionally, we found that grit, attention, self-efficacy, metacognition, intrinsic goal orientation, and effort regulation at the start of the year (T0), did not significantly predict academic performance in course credits (Table B.15, Supplementary Appendix) or dropout rates (Table B.1, Supplementary Appendix), rendering mediation even less plausible.

Discussion

As universities are looking for scalable and low-cost interventions that could aid a broad population, a reflective goal-setting intervention could provide a solution. However, the evidence about its effectiveness is divided, mechanisms that could explain why and when it works are still underexplored, and the domains in which it is tested are relatively limited. Offering the reflective goal-setting intervention in this study yielded a significant positive effect on course credits and dropout. The standardized effect size ($d = 0.13$) is relatively large when benchmarked against studies with a similar design and outcome measure, especially when the low costs per student and its scalability are also taken into account (Kraft, 2020). The intervention, on average, cost students less than 2 h, while its gains equaled 75.6 study hours and an absolute dropout risk reduction of 5.98%. In contrast to earlier results (Schippers et al., 2015), we did not find evidence that the treatment effect was significantly different across domain, gender, ethnicity, or educational background. Additionally, contrary to expectations, the treatment group did not differ significantly in SRL, resilience, grit, and engagement.

Our findings expand the literature on reflective goal-setting and life-crafting's effects on academic performance in several ways.

First, this study found significant positive treatment effects on course credits both after a semester and at the year-end. The effect on dropout was only significant after a year and not after one semester. This result most likely means that the treatment improved course credits, which then allowed the students to continue their enrollment. As the treatment effect on obtained course credits and retention grew proportionately, the intervention had a durable benefit that improved 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. Indeed, students in the focus group

had mentioned that participating in the intervention had helped them organize their studies, finances, and social lives.

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

Third, this study showed how an implementation fidelity framework (adapted from Carroll et al., 2007) can be applied to monitor the fidelity of this type of intervention. Some of these aspects were not yet reported by previous studies, such as program differentiation, or the degree to which (elements of) the intervention was new to the context. Other aspects, such as dosage fidelity could be compared with prior research. Students in our study wrote an average of 1,134 ($SD = 671$) words, or around one third of the average of around 3,000 words in Morisano et al. (2010) and Schippers et al. (2015, 2020), and spent only half the amount of time on the assignment compared to the study from Morisano et al. (2010).[2] Writing more can be an indicator of more extensive reflections and more specific goal achievement plans. Thus, the effect of the intervention might be increased even further through higher dosage fidelity. Future studies can build on this approach to ensure that implementation fidelity is closely monitored, and taken into account, through a meta-analysis. Practitioners could monitor this variable as a potential condition for optimal success.

Fourth, we found no sufficient evidence that the intervention improved the self-regulated learning (SRL) modules, grit, resilience, engagement, or general psychological well-being. There was no evidence that these constructs mediate the treatment effect, contrary to Schippers' (2017) expectations, nor did the intervention lead to significant benefits to well-being, as suggested by Schippers and Ziegler (2019). Our inconclusive results could be attributed to different causes. It could be that we failed to correctly measure the intended constructs, although the validity and reliability of the used measures were relatively good. If the constructs were measured correctly, the inconclusive results might be due to power issues. However, our power analyses and outcomes show that we were able to find small effects. Increasing power would lead to more significant results but not to larger effect sizes. It is also possible that we chose the wrong constructs, the intervention might be mediated by different internal and behavioral processes. Yet another explanation could be that the chosen constructs did not predict academic performance in this context. We indeed found that SRL and grit, at the start of college, did not predict obtaining course credits or dropping out at the end of the first year, which would have been expected, based on previous findings in the field (e.g., Bowman et al., 2015; De Bruijn-Smolders et al., 2016; Duckworth et al., 2007; Hodge et al., 2018; Sitzmann & Ely, 2011). Resilience and engagement did correlate significantly with outcome measures, but they predicted less than 1 percent of performance (Table B.15, Supplementary Appendix). This, in turn, could be due to different reasons. It could be explained by the time of measurement, the chosen measures for performance, or publication bias. While scores at T0 were not strong predictors of academic success, the scores at T1 did correlate significantly and/or more strongly with academic performance at T2. Self-efficacy, for example, showed no significant correlation with course

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

credits at T0, but for T1 it did: $d = 0.22$ $p < .01$. For T2 this was: $d = 0.29$, $p < .01$. This could mean that performance might influence self-efficacy more than self-efficacy impacts performance, and/or it could mean that self-efficacy becomes a reliable predictor only after receiving sufficient performance feedback. Both explanations can also account for the differences in the correlations with these other constructs. It is feasible that students over- or underrate their SRL, grit, and engagement within the (new) context of higher education when they just start. The studies about SRL, grit, engagement, and resilience, on which we based our expectations of the potential mediating effect did not measure those constructs right at the start of college, or did not report exactly when the measure was taken (Bakker et al., 2015; Bowman et al., 2015; de Bruijn-Smolders et al, 2016; Duckworth et al., 2007; Hodge et al., 2018; Johnson et al., 2015; Martin et al., 2015; Sitzmann & Ely, 2011). We used the accumulation of standardized test scores and dropping out of college over long periods of time with a large sample, these are likely to show smaller effects than lab and small-scale studies, researcher-made tests, or self-report measures of performance taken shortly after the intervention (Kraft, 2020; Lortie-Forgues & Inglis, 2019). The studies reviewed by Sitzmann and Ely (2011) and De Bruijn-Smolders et al (2016), such as Ford et al. (1998), use tests that are taken shortly after the intervention and are targeted at the specific training or course during which the SRL measure was taken. This is also the case for grit, engagement, and resilience. There are studies that similarly report no predictive value of the constructs that we studied, and based on their meta-analysis, Sitzmann and Ely (2011) state that there is "evidence of publication bias in self-regulated learning research" (p. 433). This could also be the case for resilience, engagement, and grit.

Limitations and Future Directions

On account of the rigorous double-blind controlled experimental design, the students and teachers received limited information about the intervention and none about its expected benefits. This situation might have lowered participation rates: 81% of all enrolled students finished both parts of the intervention or control assignment. Analyzing only these students who finished both parts would probably lead to a larger effect size and more precise estimation of the intervention's effect, but measuring the effects of *offering* the intervention, instead of *participating* in it, offers a more realistic estimation of effectiveness in a field setting. In the focus-group interviews, students mentioned that the limited information and experimental status had made them skeptical. They reported that integrating the intervention in the regular curriculum and having a mentor follow-up during the regular coaching sessions would increase the positive effect. Some students remarked experiencing too little follow-up, except for the emails that they perceived bothersome. Future studies could look into personalized ways of organizing follow-ups, such as using a chatbot-coach (Dekker et al., 2020), for such interventions to yield a larger effect.

Although this study monitored the degree of assignment completion and the number of words written by students, it did not assess the degree to which they are setting goals in a way that is consistent with goal-orientated behavior. Future studies could further explore ways in which this could easily be monitored, or improved, through feedback.

In line with the principles of replication with variation (Locke, 2015), this study explored the role of grit, engagement, resilience, and several modules of SRL, as mediators for the intervention's effect, to expand the related literature's generality. We found no evidence to suggest that these constructs were part of the core mechanisms in this context. Potentially, the effects were too small to find with the statistical power of this study, or temporary effects that faded within one semester might have been found if the impact on the psychological variables was measured at an earlier time point after the intervention. Future studies could explore the mediating or moderating effects of other potential constructs or other variables that do not require self-reported measures and test whether there might be an effect directly after participating in the intervention. Further information on mediating constructs can aid the effective directed implementation in the right conditions and contexts.

Finally, we found that gender, previous education, and ethnicity were strong predictors of academic performance and dropping out in the first year of college. Studying interventions that could potentially mitigate these negative effects, both in the first year and later years, remains a relevant topic.

Conclusion

There were significantly more course credits and lower dropout rates among the teacher and business education students who received a reflective goal-setting intervention at the beginning of their study than those who received a control assignment. This study found no evidence to suggest that the treatment effects were different for males, ethnic minorities, or students with a vocational education background. Nor did we find sufficient evidence of a treatment effect on grit, resilience, engagement, SRL, or general psychological well-being, The implementation fidelity of this study was moderate, suggesting that the treatment's effects might potentially increase further with higher treatment fidelity. These findings indicate that reflective goal-setting has a significant effect on academic performance. As the intervention took students less than 2 h to complete and their gains equaled 75.60 study hours (2.70 course credits) and an absolute risk reduction of 5.98% of dropping out (relative risk reduction = 15.17%), this is good news for educators seeking to improve academic performance. Carefully implementing this low-cost and scalable intervention can ensure that more students benefit from the intervention's positive effects.

Open Research Statements

Study and Analysis Plan Registration

There is no study and analysis plan registration associated with this manuscript.

Data, Code, and Materials Transparency

The data, code, and materials underlying the results reported in this manuscript are not publicly available.

Design and Analysis Reporting Guidelines

There is not a completed reporting guideline checklist for randomized trials included as a supplementary file for this manuscript.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports a replication of the following study: 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.

Acknowledgements

We wish to thank the two anonymous reviewers, Monique de Bruijn-Smolders, Bettina de Jong, Clementine Degener, Laurie Delnoij, Claudia Gomes, Conrad Heilmann, Ellen Klatter, Dominique Morisano, Philip Oreopoulos, Anders Schinkel, and Monique Volman, for their helpful feedback. We thank Nina Adriaanse, Annet Brugel, Els de Bock, Emiel de Ruiter, Anneke Delhaas, Fleur Doolaard, Henriett Duijs, Lex Freund, Ishany Gaffar, Jeanine Hoogerland, Juliette Hoogeveen, Marika Kouvelis, Anjali Raghunath, Mehmet Sahan, Aad Schippers, Justin Slabbekoorn, Janine Slootmaekers, Dian van Arkel, Erwin van Braam, Annemarie van Efferink, Bert van Wijk, and Chanté Zuidgeest for their aid in the execution of the experiment.

Author Contributions

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.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

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.

ORCID

Izaak Dekker http://orcid.org/0000-0002-6858-4001
Michaéla Schippers http://orcid.org/0000-0002-0795-5454
Erik Van Schooten http://orcid.org/0000-0002-2401-9115

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.org/10.1037/abn0000362

Bakker, A. B., Sanz Vergel, A. I., & Kuntze, J. (2015). Student engagement and performance: A weekly diary study on the role of openness. *Motivation and Emotion*, *39*(1), 49–62. https://doi.org/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

Bowman, N. A. H., Patrick, L., Denson, N., & 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*(6), 1019–1028. https://doi.org/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

Clark, D., Gill, D., Prowse, V., & Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments. *The Review of Economics and Statistics*, *102*(4), 648–663. 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.org/10.1002/da.10113

Corpus, J. H., Robinson, K. A., & Wormington, S. V. (2020). Trajectories of motivation and their academic correlates over the first year of college. *Contemporary Educational Psychology*, *63*, 101907. https://doi.org/10.1016/j.cedpsych.2020.101907

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/

De Bruijn-Smolders, M., Timmers, C. F., Gawke, J. C., Schoonman, W., & Born, M. P. (2016). Effective self-regulatory processes in higher education: Research findings and future directions. A systematic review. *Studies in Higher Education*, *41*(1), 139–158. https://doi.org/10.1080/03075079.2014.915302

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

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

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.org/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), 1087–1101. 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*(2), 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: The Official Journal of the Society for Prevention Research*, *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), 1182–1198. https://doi.org/10.1037/ccp0000260

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, andpractice strategies with learning outcomes and transfer. *Journal of Applied Psychology*, *83*(2), 218–233. https://doi.org/10.1037/0021-9010.83.2.218

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*(1), 40–67. https://doi.org/10.1080/19345747.2016.1246632

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.org/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

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

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

Hudig, J., Scheepers, A. W., Schippers, M. C., & Smeets, G. (2022). Motivational mindsets, mindset churn and academic performance: The role of a goal-setting intervention and purpose in life. *Current Psychology*, 1–20. https://doi.org/10.1007/s12144-022-03462-8

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. *The 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

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

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). Jossey-Bass.

Latham, G. P., & Brown, T. C. (2006). The effect of learning vs. outcome goals on self-efficacy, satisfaction and performance in an MBA program. *Applied Psychology*, *55*(4), 606–623. https://doi.org/10.1111/j.1464-0597.2006.00246.x

Locke, E. A. (2015). Theory building, replication, and behavioral priming: Where do we need to go from here? *Perspectives on Psychological Science: A Journal of the Association for 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. *The American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

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

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. (2013). Goal setting in the academic arena. In E. A. Locke & G. Latham (Eds.), *New developments in goal setting and task performance* (pp. 495–506). Routledge.

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. *The 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.). Muthén & Muthén.

Organisation for Economic Co-operation and Development. (2019). *Education at a glance 2019: OECD indicators*. 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), 188–204. 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* (pp. 417–437). Oxford University Press.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2020). *A user's guide to MLwiN version 3.05*. 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*(3), 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 dropout and university grades: A longitudinal approach. *Journal of Educational Psychology*, *112*(5), 987–1002. https://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

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.org/10.1002/job.248

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement. *Educational and Psychological Measurement*, *66*(4), 701–716. https://doi.org/10.1177/0013164405282471

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—Which of these should we use? *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *5*(5), 431–436. https://doi.org/10.1046/J.1524-4733.2002.55150.x

Schippers, M. C. (2017). *Ikigai: Reflection on life goals optimizes performance and happiness*. 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.org/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

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage Publications Ltd.

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*(6), 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. *The 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. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*(1), 73–82. https://doi.org/10.1177/0963721413512856

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*(6023), 1447–1451. https://doi.org/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), 468–485. https://doi.org/10.1037/a0037461

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3–14). https://doi.org/10.1145/2632048.2632054

Willcoxson, L. (2010). Factors affecting intention to leave in the first, second and third year of university studies: A semester-by-semester investigation. *Higher Education Research & Development*, *29*(6), 623–639. https://doi.org/10.1080/07294360.2010.501071