Increasing Student Involvement to Decrease Underachievement

Increasing Student Involvement to Decrease Underachievement: Experimental Evidence on Gender Differences in Performance

Carla Haelermans
Maastricht University

Maartje van der Eem
Maastricht University and Hermann Wesselink College

This article studies the short-run effect of increased student involvement on academic achievement, motivation, and grade repetition. We use a randomized field experiment among 130 tenth-grade students in a Dutch upper secondary school. Students who are more involved in their own learning process have significantly higher academic performance and a lower chance of grade repetition. Extrinsic motivation was lower for these students, but there was no effect on intrinsic motivation. All effects can be attributed to male students. The main explanation for the findings is that students feel more responsible and in charge of their own learning process.

I. Introduction

The problem of underachievers in education is not new but is definitely still relevant, as it poses threats to students at many different levels (Lee 2016). Underachievement is often interpreted as not performing to one's full potential, which can have drastic consequences. Data show that underachievement is more prevalent for boys than for girls (OECD 2015, among others). Underachievers have a higher risk of grade retention or dropping out (Lee 2016). Therefore, it is of great importance to reduce underachievement, so that students perform at the highest possible level, given

This article has greatly benefited from the feedback of two anonymous reviewers, the associate editor, and the editor in chief at the *Journal of Human Capital*, and from that of Wim Groot, Henriette Maassen van den Brink, Trudie Schils, Eline Sneyers, Inne Vandyck, participants of the Second Lisbon Research Workshop on Economics, Statistics, and Econometrics of Education, participants of the 2015 Leuven Economics of Education Research Workshop on Education Economics, and participants of the EEA (European Economic Association) 2015 in Mannheim.

[*Journal of Human Capital*, 2018, vol. 12, no. 4]
 1932-8575/2018/1204-0004$10.00

their innate abilities, and thereby prevent having to repeat a grade or drop out. Furthermore, the literature shows that higher student performance is associated with higher wages for these individuals (Ashenfelter, Harmon, and Oosterbeek 1999), with fewer health problems (Conti, Heckman, and Urzua 2010), and with a lower chance of committing a crime (Groot and van den Brink 2010). Besides the individual disadvantages of low performance and dropout, there is the actual cost as well. The costs per student per year are rather high (e.g., the Dutch government spends approximately €7,000 per student per year; Teule 2012), and if underachievement can be reduced, (part of) these governmental costs could be prevented.

Many teachers and other actors involved in education believe that educational performance could be improved by increasing student involvement and engagement in their own learning process. This is in line with the literature, which has shown that higher performance and a lower dropout rate are positively associated with a person's locus of control (Coleman and DeLeire 2003), involvement and engagement (Ream and Rumberger 2008), and high self-esteem and better attitude (Waddell 2006).

Therefore, the main aim of this paper it to study the effect of increased student involvement, by means of preparation of a portfolio and implementation of student-led study-progress meetings, on student performance, motivation, and grade repetition in secondary education, using a randomized experiment, with a second focus on studying a possible gender gap in these effects. So far, only few studies have focused on the effect of higher student involvement in general. Positive relations are found with peer social capital and dropout rate (Ream and Rumberger 2008), high school completion and college attendance (Coleman and DeLeire 2003), and labor market outcomes (Waddell 2006). However, none of these studies provides experimental evidence of their claims. To our best knowledge, the only related study providing causal evidence describes three field experiments on programs in Chicago—Becoming a Man and another program in a juvenile detention center—that aim to reduce crime and the dropout rate (Heller et al. 2015). The exercises in these programs teach youth to think more carefully about the situations they are in and the decisions they make regarding the behavior they display in these situations. This study shows that changing the decision-making of economically disadvantaged youth reduces crime and dropout, which the authors attribute to the reflection aspect of the program.

However, none of the previously mentioned studies on increased student involvement focuses on (intrinsic and/or extrinsic) motivation. However, motivation (and the combined notions of intrinsic and extrinsic motivation) and the mechanisms through which this would work are an often-covered topic in relation to student performance (see, e.g., Deci and Ryan 2000; Ryan and Deci 2006). Intrinsic motivation is often defined as a student's own desire to perform a certain task, because of personal interest and engagement and because the students wants to chal-

lenge himself and exercise his capacities, whereas with extrinsic motivation there often is a reward or lack of punishment that explains a certain behavior (Deci 1975; Bénabou and Tirole 2003). Intrinsic motivation is seen as very important in increasing academic performance, through well-being indicators (i.e., self-esteem), which are related to trust (Bénabou and Tirole 2003; Deci and Ryan 2000). However, extrinsic motivation can crowd out intrinsic motivation (Bénabou and Tirole 2003; although this seems to be the case only when rewards are perceived as controlling [James 2005]). On this basis, we would expect that increased student involvement (which entails increased trust in students and at the same time less external "control" by parents and teachers) may increase intrinsic motivation and decrease extrinsic motivation.

With respect to the gender question, it is mentioned above that data show that underachievement is more prevalent for boys than for girls (OECD 2015, among others). At the same time, some of the previous (nonexperimental) studies have studied the gender gap and concluded that girls seem to benefit more from taking the lead in study-progress meetings with the teacher and parents (e.g., Wehmeyer and Lawrence 1995; Vansteenkiste et al. 2009), whereas other studies have found girls to be more intrinsically motivated in the first place (Vansteenkiste et al. 2009), which potentially confounds these nonexperimental findings on girls. On this basis, we might expect different effects for boys and girls, but from the literature it is ex ante unclear who would benefit more from an intervention.

Our study uses a randomized experiment to causally analyze the short-run effect of more student involvement, by means of preparation of a portfolio and implementation of student-led study-progress meetings. In the control situation, the meeting is prepared and led by a mentor (one of the teachers who serves as a coach, i.e., a person of trust, to the student), and the student is merely present, whereas in the treatment situation the student is in control of the preparation and execution of the study-progress meeting. The effect of increased student involvement in their own learning process is studied in a randomized field experiment during one school year in grade 10[1] of a Dutch upper secondary school. We show that increased student involvement leads to higher performance, lower extrinsic motivation, and lower grade repetition. The results also show that all these effects are driven solely by boys. The main explanation for the effects found is the active student involvement, resulting in a combination of metacognitive skills (e.g., reflection), autonomy, and feedback that students actively receive and use in the intervention, through which students may feel more in charge and more responsible for their own learning process.

To our best knowledge, there are no experimental studies on the effects of this type of intervention with increased student involvement in

[1] American grade system used.

upper secondary education. The only experimental study that exists on a slightly related topic is of the previously described randomized controlled trials in Chicago on changing the decision-making of economically disadvantaged youth (Heller et al. 2015). Although that study also focuses on student reflection and metacognitive skills, it is also very different because of the different educational system and the focus on criminal behavior. Furthermore, a few experimental studies have been conducted in special education in the United States, on a group of students who are not comparable with regular students in secondary education and on outcomes that do not include student performance or motivation (Martin et al. 2006). With respect to regular upper secondary schools, most existing studies on student-led meetings are case studies or descriptive studies (see, e.g., Juniewicz 2003; Tuinstra and Hiatt-Michael 2004; Goodman 2008).

Therefore, the contribution of our study to the literature is threefold. First, to our best knowledge, this is the first study to analyze the effect of increased student involvement, focusing on gender differences in student performance, grade repetition, and intrinsic and extrinsic motivation in upper secondary school. Furthermore, given the state of the existing literature on student involvement, the second contribution of our study is the individually randomized experimental design, which allows for causal analysis. By randomizing at the individual level, we construct an appropriate counterfactual and minimize the influence of student characteristics and of a specific teacher or class. Third, by including both performance measures and motivation, we are able to distinguish these two effects.

Section II of this paper focuses on the context, contents, and organization of the experiment. In Section III, we describe the identification strategy and methodology as well as the characteristics of our sample. Section IV contains the results of the intervention. We focus on the full sample as well as on subsamples. Section V contains the robustness analyses and cost-effectiveness analysis. The final section of this paper contains the conclusion and a discussion of the results.

II. Context

A. Dutch Educational System and the School under Study

Dutch secondary education consists of three tracks: prevocational education, higher general education, and preacademic education (known by the Dutch acronyms VMBO, HAVO, and VWO, respectively). Prevocational education takes 4 years, higher general education 5 years, and preacademic education 6 years (see fig. 1). Students attend a new school and are placed in a track from grade 7 on, immediately after finishing primary education, when they are about 12 years old. The primary school teacher gives a track advice based on the development of the child and the results of the Cito test. The Cito test is a standardized national test

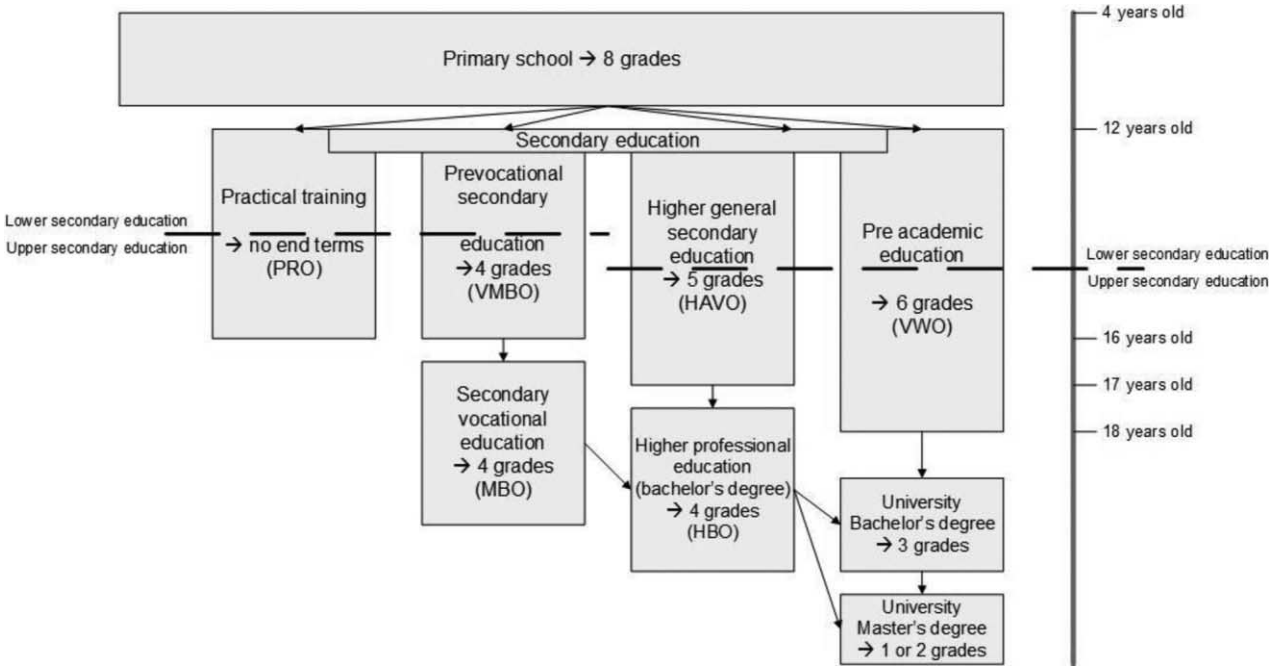

Figure 1.—Dutch educational system (authors' own creation).

almost all Dutch students take in grade 6, the final year of primary education, which focuses on language, mathematics, world orientation, and general study skills. Despite the early tracking in the Dutch system, it is possible to move up a track in secondary education, either immediately after grade 7 or later, for example, after passing the final secondary exam of a lower track in grade 10 or 11. Students can also be placed in a lower track during secondary education when the current track level appears to be too difficult. Note that grade repetition or stepping back a track is common in the Netherlands.

In most schools, in lower secondary higher general school (grades 7–9) each class gets assigned a mentor, who is one of the teachers from lower secondary school who serves as a person of trust on all things except content-related aspects of specific topics and classes. In upper secondary school (grades 10–11), each individual student gets assigned a mentor (as there are no classes in the sense that one student group attends all topics together at the same time, since students have to choose a curriculum and additional courses). In upper secondary school, the mentor is one of the teachers who teaches in upper secondary school. Usually mentors are randomly assigned to classes/students.

Our sample consists of all 130 students in grade 10 (upper secondary school; age of about 15 years) of higher general education, the middle track (HAVO), in one school, although these students are in the upper half of the academic performance distribution, as the lowest track (VMBO) comprises about 50 percent of students. These students have to take their final exams in grade 11. After grade 9, students must choose either a science or an arts curriculum. However, note that all students must take Dutch (mother tongue) and English (first foreign language) as subjects, regardless of the curriculum they choose. Grade 10 is an interesting grade level for our research for several reasons. The first is because the transition from grade 9 to grade 10, from lower to upper secondary school, is difficult for many students. About 46 percent of higher general education students have to repeat a grade (Klomp and Thielen 2010), and a great many of them have to repeat grade 10. The second is because grade 10 of higher general education is considered a melting pot of types of students. There are students who have been in the higher general education track since grade 7. They are accompanied in grade 10 by students who moved up a track after graduating from grade 10 in the prevocational track and by students from the preacademic track (VWO; from both grade 9 and grade 10) for whom the higher academic level was too difficult. Students coming from all these different backgrounds have to find their way in this grade 10 of higher general education. Increasing student involvement can help these students to reflect on their motivation and performance and, hopefully, also increase their performance.

The school we are studying is a regular secondary school, where both lower and upper secondary education are offered. The school is located in the western part of the Netherlands, near Amsterdam. The fact that our

research is conducted at one Dutch secondary school is a potential disadvantage of the study, as this might threaten the external validity. However, we are confident that the results are also applicable to most of the secondary schools in the Netherlands, because this school is very representative of the average secondary school in the Netherlands: all statistics for this school are within half a standard deviation of the average of all variables. Compared to the average Dutch secondary school, this school has about 1,500 students (national average: 1,473 [standard deviation (SD): 1,142]), about 94 full-time equivalent teachers employed (national average: 130 [SD: 101]), a graduation percentage of 88 percent (national average: 90 [SD: 5]), and an average national exam grade of 6.3 (on a scale from 1 to 10; national average: 6.4 [SD: 0.2]).[2]

B. The Experiment

1. The Intervention

Previously, the school under study had organized study-progress meetings between the mentor, the student, and his parents for all students of higher general education in upper secondary school. These meetings were organized twice per year, to discuss the study performance and progress of the student. The mentor prepared, with input from other teachers, and led this meeting. Students were present, but most of the talking was done by the mentor and the parents. This situation is the control situation in our experiment.

The intervention, which started in 2012–13, consists of an active role for the students (increased student involvement) in both the preparation of this study-progress meeting and in the meeting itself. Figure 2 summarizes the steps of the higher-student-involvement intervention for treatment and control groups separately. Figure 2 shows that the first three steps (out of five in total) for both the treatment group and the control group focus on evaluation of study progress and preparation for the meeting using a portfolio. Step 4 is the actual meeting, and step 5 is the evaluation of the meeting and the recording of it in the portfolio.

The intervention consists of three elements that might lead to higher student performance and motivation. The largest element of the intervention is the use and development of metacognitive skills, which have been shown to influence student performance (see, e.g., Masui and De Corte 2005). In their portfolios, students have to describe their strong and weak points and explain why they are satisfied with certain achievements but not so much with others. The students reflect on both how and why they performed the way they did. The reflection by the students is guided by questions: for example, which study skills they would like to

[2] The data are from 2012 and are obtained from the government website containing the Dutch open education data (https://www.duo.nl/open_onderwijsdata/databestanden/vo/; in Dutch).

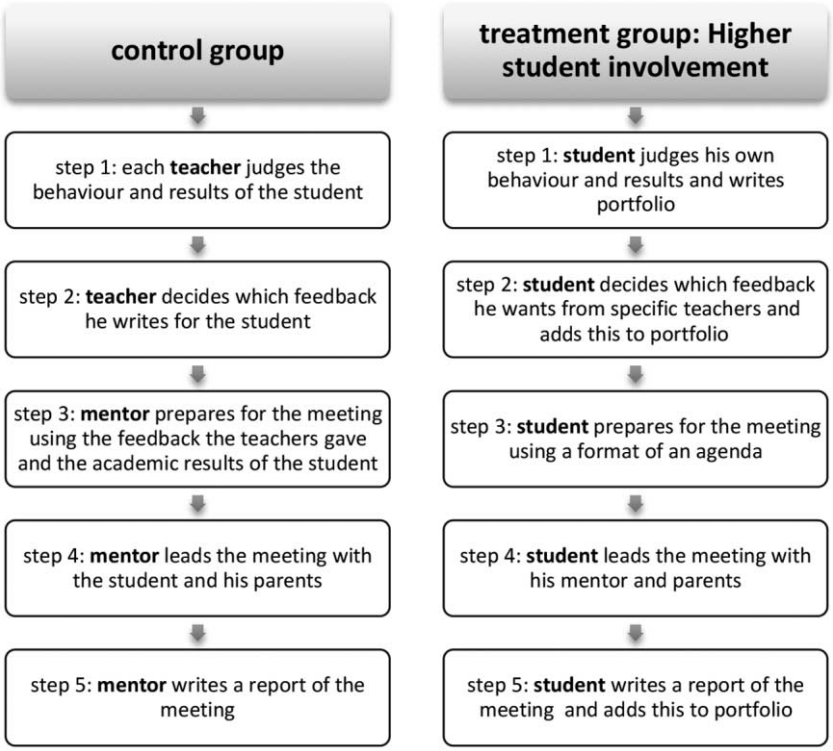

Figure 2.—Content of the experiment.

improve during this year or in what way a teacher can be helpful in their study. Students have to reflect on the outcomes of a survey on study behavior, which they filled out in the beginning of the school year. Metacognitive skills are also developed during the meetings, when students have to take the lead, set goals, and think about how to reach those goals. In the second meeting, students also discuss whether they met the goals they set in the previous meeting. After each meeting, students again have to use metacognitive skills: they have to decide which parts of the meeting are important enough to write down in the report as part of the portfolio.

A second, much smaller component, is autonomy. In composing a portfolio and preparing the agenda for the meeting, students have a certain degree of autonomy. They can decide on the content of their portfolios and on the topics they want to discuss with their mentor and parents during the meeting (although the mentor and the parents can bring in additional topics to discuss). According to the self-determination theory of Deci and Ryan (2000; also Ryan and Deci 2006), autonomy is a key element for improving motivation.

Another (also smaller) part of the portfolio is the feedback from teachers. Receiving feedback and dealing with feedback are important learning aspects for students, and asking for feedback also requires more in-

volvement. When students receive feedback from a mentor or a parent during the meeting, the teacher is not present. However, teachers can give more specific feedback than the mentor on a how the student is doing in the specific subject taught by this teacher, and therefore teacher feedback was added to the portfolio. During portfolio classes, students have the option to ask one or more of their subject teachers for specific (digital) feedback, through the digital system in which they write their portfolios. It is important to give students as much autonomy in this as possible, because feedback—although an important factor for improving results—is not effective per se (Hattie and Timperley 2007). According to Shute (2008), feedback is effective only when it is nonevaluative, supporting, and given at the right time. Students in the treatment group are in charge of deciding when to ask for and receive feedback and from whom, by asking a specific question. Students in the control group, on the other hand, receive unsolicited feedback from the mentor during the meeting. Therefore, we expect the feedback to be more effective for treatment students than for control students.

2. Time Line of the Intervention

Figure 3 gives a graphical representation of the time span of the experiment. The week before the start of the school year (2012–13), the students were assigned to one of five classes by the school administrator, usually on the basis of the curriculum the student chose (so not random). Classes have only administrative purposes in upper secondary education, as the composition of classes varies for each subject, depending on whether the student chose that subject. An individual student was randomly assigned to one of 10 mentors (a teacher from upper secondary school). On average, mentors had between 7 and 22 students (with an average of 14). Two students were assigned to a special mentor because of a speaking disorder, for which they might need extra attention, in which these mentors are specialized. In the second week of the school year, the researchers used stratified randomization, based on gender, previous grade level and track, and choice of curriculum, to divide the students between the treatment and control groups. We had a total of 16 strata, with size varying between 5 and 15 students, and within each stratum students were randomized to ensure maximum comparability between treatment and control students, given the fact that we had only 130 students. Students were not informed that they were part of an experiment (which was possible because of the individual nature of the treatment [individual meetings with mentor and parents] and the varying group composition of the tutorial classes, which they were used to), and teachers of grade 10 were not informed which students belonged to which group. Mentors were informed about treatment and controls status of only their own students.

In week 37, the student filled out the pretest motivation and study behavior questionnaire (T0). Treatment students had a first tutorial class in

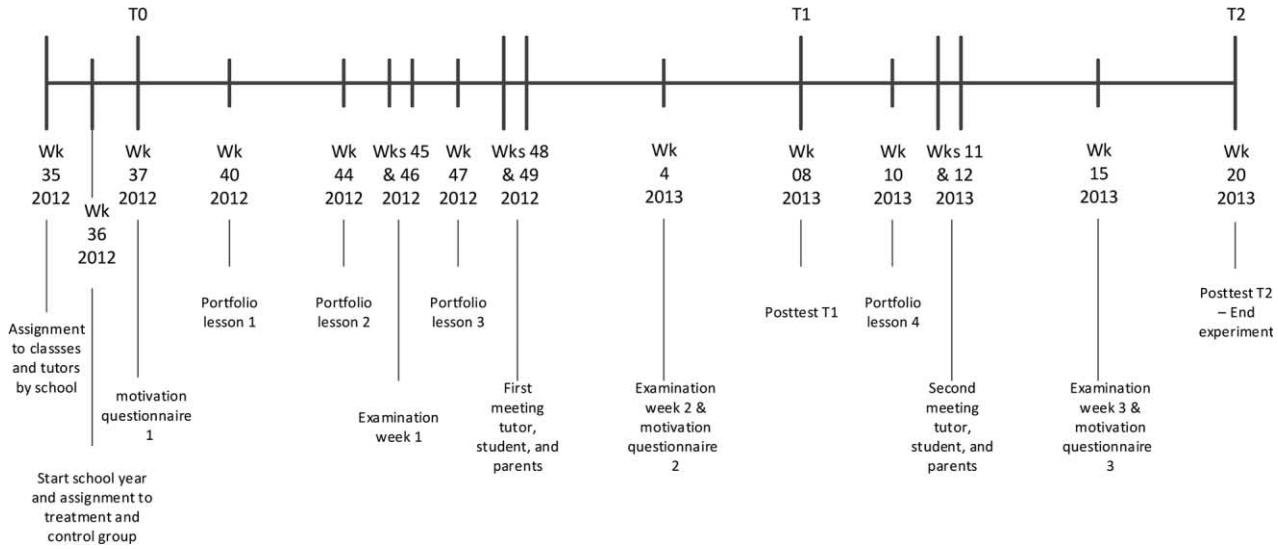

Figure 3.—Overview of the time line of the experiment. Wk: week.

which they set up their electronic portfolio, about 4 weeks after the start of the school year (week 40). At the school under study, tutorial classes of 80 minutes are scheduled on a weekly basis throughout the school year, in which groups of students get remedy classes, comprehensive classes, or time to spend on education activities of their own choice. For treatment students, three of these classes were used as portfolio classes. During these hours, students in the control group have time scheduled to work on educational activities of their own choice. Note that for these tutorial classes, students are divided into six groups of average class size. Since these classes always take place in varying group compositions and various activities are scheduled simultaneously, control students do not know that they are doing something different from treatment students. Four weeks later, the students worked on their portfolios for a second time. Now they had to write about their strengths and weaknesses in each subject. Because the students had been in grade 10 for 2 months at that time, they had some idea about their capacities and whether they would be able to perform well in this grade level. Three weeks later (week 47), the students completed their portfolios, which they then used as a preparation for the first meeting. They had the option to ask some of their teachers for feedback. At that time, students had received most of their grades from the first examination period of grade 10, which took place in weeks 45 and 46. During weeks 48 and 49, all students, from both the treatment and control groups, had the first meeting with their mentors and parents. Four or five school weeks later, in week 4 of 2013, the students had their second examination week and filled out the second motivation questionnaire. To determine whether the first round of the intervention affected student achievement, we measured the average grades of the students after this second examination week (T1).

In week 10, students in the treatment group had their fourth and final portfolio class. One or two weeks later, all students had their second meeting. Three or four weeks after this meeting, students had their third examination week and filled out the third and last motivation questionnaire (week 15). Our last moment of measurement (T2) was 5 weeks after this examination week, when all grades were available.

At T1 and T2, all students took the same number of tests and thereby had the same number of grades, which were weighted in exactly the same way. In examination weeks, all students take exactly the same tests for a subject and are graded the same way, independent of their teacher.

3. Reliability and Validity of the Intervention

During the course of the intervention, we tried to minimize possible threats to the reliability and validity of the intervention. The instructions for students during the portfolio classes were read from paper by the teachers, and the classrooms in which these classes took place (simultaneously) were next to each other, so teachers could discuss issues that would possibly arise during the portfolio classes. By doing this, we tried to ensure

that the treatment students received the same treatment, as much as possible.

To ensure reliability of the instruments, we decided to use an existing, validated questionnaire to measure motivation and study behavior. This is discussed in more detail in Section III.A. The individual randomization strengthens the internal validation of the intervention. However, the number of observations might lead to power problems, which may lead to a type II error.

There are no indications that the behavior of either treatment or control students was influenced during the intervention in any way other than that intended by the intervention (e.g., Hawthorne and/or John Henry effect). Because the students were not explicitly told that they were part of an experiment and the portfolio classes where somewhat "hidden" in the tutorial classes, most students found out only after the first meeting with tutor and parents that there were two version of these progress meetings. However, not telling students that they are part of an experiment may also cause treatment students to believe they are special once they find out about the experiment. Unfortunately, this cannot be ruled out, although additional analyses of the questionnaire with the focus on consistency in answers also do not give any indication to suspect that students started behaving in a different way once they realized that there were different versions of the meetings. There is also no indication at all from observed student behavior or from the questionnaires that there would be a risk of a Hawthorne and/or a John Henry effect.

Another factor in this experiment was the mentor. In the control condition, mentors had to analyze the study progress of each student, on the basis of input from all teachers of that student, and lead the study-progress meeting. However, in the treatment condition the mentor prepared behind the stage (to be able to interrupt when necessary, although that hardly ever happened during the intervention) and did not take the lead during the meeting itself, which enforced a different type of action for the mentor. Martin et al. (2006) and Tholander (2011) pointed out that the mentor can play a role that is not desired for this setup during the student-led meeting, as mentors can unintentionally encourage students in the control group to behave like the students in the treatment group and take the lead in the meeting, whereas the mentor could unintentionally take over the lead from the student in the treatment group during the meeting. Therefore, mentors were informed and reminded about their role in both situations several times during the experiment. Additionally, to get an idea whether a mentor played the "wrong" role during the meeting, we recorded two meetings per mentor, with random students. These recordings show that the mentors played their role quite well in both the student-led and the mentor-led meetings. The applied analysis and the results of these recordings are discussed in Sections III.B and IV.A.

Furthermore, there is also no indication that the teachers may have played an unexpected role in the intervention. Although teachers of course

notice when they get specific questions from students (which is done digitally), it is unlikely that this influenced the teachers' behavior toward these students in class, since for each teacher that would be only a few students; and since each student asked only few questions (to only a few teachers), it is unlikely that the average grade (over many subjects) of these few students was influenced by a possibly biased teacher in such a way that it influenced the overall result of the experiment.

Unfortunately, things did not always go as planned in the intervention. All students were supposed to have four portfolio meetings. All the treatment students had the first two preparation classes, but the third class was scheduled close to the first study-progress meeting, and it appeared that for a group of 18 students there was another class scheduled simultaneously, where they were obliged to be present. These students were asked to finish their portfolios at home, and 16 of them did. Only 2 students had a partially complete portfolio as a preparation of the first meeting. Furthermore, all students were supposed to have two meetings during the school year. However, 7 students (or their parents) opted to not have the second meeting, and, against all rules, the mentors of these students agreed to that, so not all students actually had two meetings. It is unclear whether and to what extent this may have influenced the result of this study. On the one hand, these students still wrote their portfolios and used their autonomy to not have the second meeting; on the other hand, these students did not participate in the full intervention and thereby missed opportunities to use and further develop their metacognitive skills.

III. Data and Method

A. Motivation Questionnaire

Before the start of the experiment, students filled out a survey on study behavior and motivation, for which we used a validated LEMO (LEarning and MOtivation) questionnaire constructed by Donche et al. (2010). In the analysis, we use the scales on intrinsic motivation (called "autonomous motivation" in the questionnaire) and extrinsic motivation (called "controlled motivation" in the questionnaire), because these types of motivation are said to have an influence on academic results (Deci and Ryan 2000; Ryan and Deci 2006). The two motivation factors both consisted of 6 questions, the students' answer to which (on a 5-point Likert scale) were added to get the overall motivation scale measure, with a minimum of 6 and an absolute maximum of 30. Donche et al. (2010) argue that all subscales in their survey, among which the ones on motivation, are reliable, with a Cronbach's $\alpha \geq 0.69$.[3] For additional validation, we calculated Cronbach's α for the answers the students in our sample gave: intrinsic

[3] Cronbach's α represents the reliability of a scale, where a higher number indicates higher internal consistency between the items of that scale and thereby a higher reliability. An α of 0.70 and up is considered to represent a reliable scale (Field 2013).

motivation had a Cronbach's α of 0.70 and extrinsic motivation one of 0.84 (similar to those that Donche et al. found). The questions that were used for the two motivation subscales (translated to English by the authors) can be found in appendix A.

B. Sample and Data

At the start of our experiment, 133 students started in grade 10 at the higher general education level at the school under study. Three of these students were excluded from the sample because they attended very few classes, did not participate in the intervention (or in control activities), did not fill out the questionnaires, and received (almost) no grades, as a result of physical or psychological problems. Therefore, our starting sample for analysis consisted of 130 students. The characteristics of the sample are summarized in table 1.

Table 1 shows that girls were the majority in this group: about two-thirds of the students were female. The students were, on average, 15 years of age at the start of the intervention. Almost 60 percent chose an arts curriculum; the others chose science. About 18 percent of the students had been diagnosed with dyslexia or attention deficit/attention deficit hyperactivity disorder (AD(H)D). Almost all students had Dutch as a first language. Two-thirds of the students moved from grade 9 of higher general education to grade 10 of higher general education. Sixteen students in

TABLE 1
Descriptive Statistics of the Sample

	N	Mean	Standard Deviation	Minimum	Maximum
Girl	130	.63	.48	0	1
Age (years)	130	15.38	.56	15	17
Dyslexia/AD(H)D	130	.18	.39	0	1
Dutch as first language	130	.92	.27	0	1
Science curriculum	130	.42	.5	0	1
Grade and track in 2011–12:					
Grade 9 HAVO	86	.66	.48	0	1
Grade 10 HAVO	16	.12	.33	0	1
Grade 9/10 VWO	13	.1	.3	0	1
Grade 10 VMBO	15	.12	.32	0	1
Cito test score	122	538.55	5.3	520	549
Average grade in 2011–12 (1–10 scale):					
Overall grade	129	6.44	.53	5	8.2
Dutch	129	6.17	.55	4.9	7.9
English	129	6.23	.83	4.1	8.8
Mathematics	118	6.11	.98	3.4	8.7
Extrinsic motivation at T0	125	16.67	4.63	6	28
Intrinsic motivation at T0	125	18.02	4.46	7	30

Note.—We combined the students coming from grades 9 and 10 of the preacademic track. The groups were too small to include separately, and they both have had the same experience anyway: being placed in a lower track because of bad performance. HAVO: higher general education track; VWO: preacademic track; VMBO: prevocational track.

the sample had to repeat grade 10. Exactly 10 percent of the students were placed in a lower track, compared with the previous year (i.e., came from grade 9 or 10 of preacademic education), while almost 12 percent of the students moved to a higher track (i.e., came from grade 10 of prevocational education). These numbers also show why grade 10 of higher general education is called a melting pot.

Table 1 shows that the pretest extrinsic motivation had an average of 16.7 and intrinsic motivation an average of 18. We also see in table 1 that five students did not fill out the motivation questionnaire, because of illness in the week in which the questionnaire was filled out as well as in the week after, when a second opportunity to fill it out was scheduled. Two of these students were in the control group and three in the treatment group. On the basis of observable characteristics, there is no reason to assume that this attrition was selective. Attrition with respect to the questionnaires throughout the experiment is further discussed in the next section.

The scores on the standardized Cito test, which is a primary school exit test, can vary from 501 to 550, the latter being the highest score. The national guideline for admittance to the higher general education track is a score between 538 and 541. Although the average score in this sample lies within the higher general education track range, the large standard deviation shows the wide variety in these scores, because of students moving up or down a track during junior high or high school.

Finally, we look at the average final grades from the school year before the experiment. For each student, the seven subjects he follows in his curriculum in grade 10 have been used to calculate his overall grade average. These are also the seven subjects a student has to take at the national exit exam at the end of grade 11 of higher general education. Two of these subjects, Dutch and English, are compulsory for all students; participation in the other subjects varies across students.

Grades are measured on a scale from 1 to 10, using one decimal place, 10 being the highest score. A 5.5 is considered sufficient to pass the subject. Students are allowed to have a few average grades below 5.5; however, an average grade below 3.5 for one of the seven subjects makes it impossible to continue to the next grade. Next to the overall grade we focus on two separate grades: Dutch and mathematics, mainly because these subjects are key subjects in the national exit exam but also because these are interesting grades to compare internationally, because many researchers focus on the grades for language and mathematics (see, e.g., Perry, Albeg, and Tung 2012).

Note that we have average final grades from the year before for only 129 students for Dutch and English, as one student was transferred from another school and we do not have prior data from this student, and that we have only 118 observations for mathematics, because 10 students did not have mathematics (as it is not a compulsory subject) and for the one student we do not have the data.

Furthermore, additional data were collected on how active the students were in meeting preparation and in the recorded meetings. All 65 students in the treatment group composed a portfolio for the first meeting; 63 of them added new information to their portfolio for the second meeting. For the first meeting, treatment group students asked their teachers an average of 2.26 feedback questions (range: 0–5); for the second meeting, there were 2.32 questions per student (range: 0–8). All feedback questions were provided with an answer. Teachers also had the possibility of adding a comment about the student's performance in their class to that student's portfolio without being asked a specific question. For the first meeting, there were an average of 0.85 comments per student (range: 0–5); for the second meeting, there were 1.02 comments per student (range 0–4). Comments could be both appraisal and constructive feedback.

Additionally, the recordings of the meetings gave us information about how many minutes each of the three actors (mentor, student, and parents) talked during the meetings. This is discussed in more detail in Section IV.A.

C. Outcome Indicators and Attrition

In this study, we have three outcomes: student performance measured as the average grade, student performance measured in grade repetition, and student motivation (both intrinsic and extrinsic). For the first measure of student performance, we have average grades over all subjects and grades for Dutch and mathematics separately. As mentioned above, not all students had mathematics, and therefore we have a subsample for that. Student motivation is based on three questionnaires. Here we also have subsamples in our analysis, as not all students filled out all questionnaires.

A total of 105 students filled out all three questionnaires (125 students filled out the first questionnaire; of those, 114 filled out the second one, and of those, 105 also filled out the third one). Some students filled out only one or two of the three questionnaires. Table 2 shows the attrition for the mathematics subsample as well as for the motivation questionnaire subsample (based on who filled out all three of them), including a statistical comparison. Table 2 shows that the questionnaire sample is very comparable to the group of students who did not fill out all three questionnaires, except for the prior mathematics grade, which is significantly higher for the group that did fill out all three questionnaires. If we look at the mathematics subsample, we see that this subgroup differs significantly with respect to science curriculum and all three prior grades. These significant differences make sense, as mathematics is a compulsory subject if a student has a science curriculum. Furthermore, mathematics is considered a difficult topic by most students, and grades are not very high, which weighs down the average grade and explains why the overall grade was significantly higher for the group who did not have mathematics. Also, students who do not choose mathematics are usually a lot better

TABLE 2
Representativeness of Subsamples: *T*-Statistics on Observable Characteristics of Students in Subsamples versus Other Students

	Questionnaire Sample					Mathematics Sample				
Variable	*N* Yes	Yes	*N* No	No	*p*-Value of Difference	*N* Yes	Yes	*N* No	No	*p*-Value of Difference
Girl	105	.62	25	.68	.57	120	.61	10	.90	.07*
Age	105	15.37	25	15.44	.59	120	15.38	10	15.40	.93
Dyslexia/AD(H)D	105	.21	25	.24	.74	120	.23	10	.00	.09*
Dutch as first language	105	.92	25	.92	.95	120	.92	10	1.00	.35
Science curriculum	105	.45	25	.28	.13	120	.45	10	.00	.01***
Grade and track in 2011–12:										
Grade 9 HAVO	105	.69	25	.56	.24	120	.67	10	.60	.67
Grade 10 HAVO	105	.10	25	.20	.20	120	.13	10	.10	.82
Grade 9/10 VWO	105	.09	25	.16	.27	120	.10	10	.10	1.00
Grade 10 VMBO	105	.12	25	.08	.54	120	.11	10	.20	.39
Cito score	99	538.51	23	538.74	.85	113	538.78	9	535.67	.09*
Average grade in 2011–12 (1–10 scale):										
Overall grade	104	6.48	25	6.29	.12	120	6.41	9	6.91	.01***
Dutch	104	6.20	25	6.05	.21	120	6.13	9	6.76	.00***
English	104	6.22	25	6.24	.91	120	6.15	9	7.28	.00***
Mathematics	96	6.22	24	5.70	.02***					
Extrinsic motivation at T0	105	16.51	20	17.50	.38	115	16.76	10	15.70	.49
Intrinsic motivation at T0	105	18.08	20	17.75	.77	115	17.93	10	19.10	.43

Note.—Comparisons are based on observable characteristics of students: those who did fill out motivation questionnaire versus those who did not, and those who did have mathematics as a subject versus those who did not. HAVO: higher general education track; VWO: preacademic track; VMBO: prevocational track.
* $p < .10$.
*** $p < .01$.

in languages, which is reflected in the higher grades for Dutch and English. However, although the subsamples for mathematics and motivation were selected groups, the treatment and control students within these subsamples were not significantly different from each other (see app. B, available online).

D. Identification

Because the experiment itself did not take place within the context of a classroom, we were able to randomize the 130 students in our sample at the individual level. As explained above, we used stratified randomization, based on gender, previous grade and track, and choice of curriculum. Because of the individual random assignment of the students in

the sample, all mentors and all teachers in grade 10 had students in both the control and treatment groups. This implies that the results are independent of an individual mentor, teacher, or class. Note that, if we were to find a small effect of 0.2 of a standard deviation, the chance of a type II error would be quite high with a sample of 130 students, as the power would be only 0.20. However, if we were to find a medium effect of 0.5 of a standard deviation, we do have enough observations for a power of 0.80.

Table 3 presents a comparison of the treatment and control groups on observable characteristics. As expected, there were no significant differences between the two groups. These results suggest that we successfully randomized the students and that it is likely that unobservable characteristics were randomly distributed as well.

E. Methodology

To determine whether student-led study-progress meetings have an effect, we estimate the average treatment effect (ATE), using the notation described by Rosenbaum and Rubin (1983). Because the randomization ensures independence between the treatment and the potential outcome, the ATE can be formulated as follows:

TABLE 3
Independent *T*-Tests on Observable Characteristics

	Treatment Group (n = 65)			Control Group (n = 65)			
Variable	*N*	Mean	Standard Deviation	*N*	Mean	Standard Deviation	Difference
Girl	65	.63	.49	65	.63	.49	1.00
Age	65	15.32	.56	65	15.45	.56	.21
Dyslexia/AD(H)D	65	.20	.40	65	.23	.42	.67
Dutch as first language	65	.89	.31	65	.95	.21	.19
Science curriculum	65	.40	.49	65	.43	.50	.72
Grade and track in 2011–12:							
Grade 9 HAVO	65	.65	.48	65	.68	.47	.71
Grade 10 HAVO	65	.12	.33	65	.12	.33	1.00
Grade 9/10 VWO	65	.11	.31	65	.09	.29	.77
Grade 10 VMBO	65	.12	.33	65	.11	.31	.79
Cito score	60	538.57	5.22	62	538.53	5.41	.97
Average grade in 2011–12 (1–10 scale):							
Overall grade	65	6.44	.57	64	6.44	.49	.98
Dutch	65	6.19	.62	64	6.15	.47	.67
English	65	6.20	.88	64	6.26	.77	.70
Mathematics	61	6.13	.86	59	6.11	1.07	.91
Extrinsic motivation at T0	62	16.24	4.93	63	17.10	4.30	.30
Intrinsic motivation at T0	62	18.03	4.70	63	18.02	4.24	.98

Note.—HAVO: higher general education track; VWO: preacademic track; VMBO: prevocational track.

$$\text{ATE} = E(Y_i(1) - Y_i(0)). \quad (1)$$

Here, E is the expected value, $Y_i(1)$ is the outcome for the student i who is assigned to the treatment group and $Y_i(0)$ is the outcome for the student i who is assigned to the control group. We can use independent T-tests and regression analyses to determine the ATE. First, we use single linear regression analysis (model 1), where the treatment is the only explanatory variable in the regression:

$$Y_i = \alpha_0 + \alpha_1 T_i + \varepsilon_i. \quad (2)$$

In equation (2), Y_i is the dependent variable for student i; T_i is the treatment indicator: $T_i = 1$ represents the students assigned to the treatment group, whereas $T_i = 0$ are the students assigned to the control group. The last element of this equation is ε_i, the normally distributed error term.

We then use multivariate analysis (model 2), where we take into account covariates to estimate the model more precisely and reduce standard errors:

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 X_i + \ldots + \varepsilon_i. \quad (3)$$

In equation (3), we add n covariates (X_i). All covariates added to the regressions were measured at T0. We add gender, curriculum (arts or science), Cito score, and grade/track in the year before the experiment as controls. We also included dyslexia/AD(H)D (all but one of the affected students have both) and Dutch as a first language in the analysis, because these factors can influence student performance as well, especially language performance. Furthermore, we add motivation (also measured at T0) and mentor fixed effects to the regression. Note that all covariates and fixed effects add solely to the precision of the estimation of the model. Given the randomized nature of the experiment, the results of the effect of the treatment should be consistent with and without these covariates.

We run these regression models separately for our different types of outcome variables. The first type is the performance outcome variables, which are the overall grade average, based on the seven subjects each student follows in his curriculum, and the individual subject grades for Dutch and mathematics.

The second type of outcome is motivation. We use the two motivation scales, with intrinsic motivation and extrinsic motivation as outcomes. The models we run for this are slightly different, as we add the pretest of that particular type of motivation as a control variable.

Finally, we use grade repetition as the third type of outcome. This is a combined indicator, as it includes students who had to repeat grade 10 after the year of the experiment, students who were initially promoted to grade 11 but were placed back in grade 10 within the first 2 months of the next school year, and students who dropped out completely in the next school year (this happened to only 2 students).

Because the experiment took place at the individual level and was randomized at the individual level, whereas students were also randomized into classes and to mentors, as explained above, the models presented in Section IV have no clustered standard errors, as is discussed in Weiss, Lockwood, and McCaffrey (2014). However, if we cluster standard errors at the mentor level (which is the only level where we potentially expect a difference, as the mentor has a large influence on the actual performance meeting), we do not see notable differences, as can be seen from the results in appendix B.

IV. Results

A. Student Activity during the Intervention

As explained above, the crucial part of the intervention is for the student to be active and involved, via increased autonomy, focus on metacognitive skills, and solicited feedback. This holds for creating the portfolio but also for the participation in the student-led meeting. As described in the last part of Section III.B, students in the treatment group behaved as was expected of them. They wrote their portfolio, and they asked their teachers questions. Note that there was a slight increase in the average number of feedback questions and comments for the second meeting and a decrease in the range. This points toward a more equal distribution of the number of questions and comments. A possible explanation for that is that both students and teachers got more used to how the system worked, in comparison to the first round of meetings.

During the meeting, the students in the treatment group also were more active than the students in the control group. We recorded meetings of both types and counted the seconds each of the participants in the meeting was speaking. The results are reported in table 4. Treatment group students spoke significantly more in the meetings than the students in the control group, whereas their mentors spoke significantly less

TABLE 4
Time Speaking during the Meeting (%)

	Treatment Group ($n = 7$)		Control Group ($n = 7$)			
	Mean	Standard Deviation	Mean	Standard Deviation	Percentage Point Difference	*T*-Statistic
Student	42.57	7.48	19.00	7.19	23.57	−6.01***
Mentor	42.14	12.21	66.71	13.15	−24.57	3.62***
Parents	15.14	9.53	14.71	11.27	.43	−.08

Note.—Seven out of nine mentors recorded a meeting with a student from the control group and a meeting with a student from the intervention group. Two mentors thought they had recorded the meetings, but the recording device did not work, so the meeting recordings of these mentors could not be included in this table.
*** $p < .01$.

than those for control group students. There was no significant difference in the parents' activity during the meetings. This is exactly as we would expect. Note that we did not see differences in gender (not presented in the table).

B. Effect on Performance, Motivation, and Grade Repetition: Basic Results

To estimate the ATE, we start with independent *T*-tests, where the outcome variables are standardized with a mean of 0 and a standard deviation of 1. All results from these tests are presented in standardized measures. Table 5 presents the average standardized grades and motivation scores of the students at T0 (pretest), T1 (after the first meeting), and T2 (after the second meeting). It also shows the grade retention rates at T2. Table 5 shows that treatment group students performed better than control group students on all three performance outcome variables at both T1 and T2. The overall average grade and the average mathematics grade of the treatment group were both significantly higher at T2, at the 5 percent significance level. Although the students in the treatment

TABLE 5
INDEPENDENT *T*-TESTS ON GRADES, MOTIVATION, AND GRADE RETENTION

	Treatment Group			Control Group				
	N	Mean	Standard Deviation	*N*	Mean	Standard Deviation	Difference	*p*-Value
Overall average grade:								
T0	65	.00	1.08	64	.00	.92	.00	.98
T1	65	.15	.93	65	−.15	1.05	.31	.08*
T2	65	.18	.90	65	−.18	1.07	.36	.04**
Dutch:								
T0	65	.04	1.13	64	−.04	.86	.08	.67
T1	65	.16	.90	65	−.16	1.07	.32	.07*
T2	65	.16	.92	65	−.16	1.05	.31	.07*
English:								
T0	65	−.03	1.06	64	.03	.94	.07	.70
T1	65	.11	.91	65	−.11	1.08	.21	.22
T2	65	.12	.92	65	−.12	1.06	.23	.18
Mathematics:								
T0	61	.01	.89	59	−.01	1.11	.02	.91
T1	61	.14	.95	60	−.14	1.04	.27	.13
T2	61	.19	.91	60	−.19	1.06	.39	.03**
Extrinsic motivation:								
T0	51	−.14	.99	54	.07	.94	.21	.27
T1	51	−.10	.92	54	.17	.92	.27	.13
T2	51	−.29	.90	54	.29	.97	.58	.00***
Intrinsic motivation:								
T0	51	.01	1.10	54	.01	.93	.00	1.00
T1	51	.11	1.16	54	−.14	.82	.25	.20
T2	51	.04	1.18	54	.01	.75	.03	.87
Grade retention at T2	65	−.17	.76	65	.17	1.18	.34	.06*

* $p < .10$.
** $p < .05$.
*** $p < .01$.

group performed better in Dutch and English as well, these differences are not significant at the 5 percent level.

For the two motivation outcomes, we see that at T1 and T2, extrinsic motivation was always lower for treatment students, whereas intrinsic motivation was always higher, which is in the expected direction. However, this was significant (at the 1 percent level) only for extrinsic motivation at T2.

Furthermore, table 5 shows that significantly more students had to repeat a grade in the control group than in the treatment group. This is significant at the 10 percent level ($p = .06$).

C. *Effect on Performance, Motivation, and Grade Repetition: Regression Results*

In tables 6, 7, and 8, the standardized results of the regression analyses are presented for the six outcome measures separately, for both the model without covariates (model 1 [M1]) and the model with all the covariates (model 2 [M2]) to estimate more precisely. Note that all regression estimations (both M1 and M2) are based on the smaller sample of students for whom we have all covariate information available for that particular outcome, leading to slightly different results for M1 than we presented in table 5.

The coefficients for overall grade are positive in table 6, implying that the treatment group students, on average, scored higher grades throughout the year than the students in the control group. The grade repetition variable is a reversed variable: the lower the better. Therefore, we also expected a negative sign here, as is indeed the case. For the Dutch and mathematics grades in table 7, we see only positive coefficients, as expected. In table 8, we see all negative coefficients for extrinsic motivation but positive coefficients (except for M2 at T2) for intrinsic motivation. The negative coefficients for extrinsic motivation are also as expected, because extrinsic motivation has a reversed interpretation: one would expect that

TABLE 6
Regression Analyses on Overall Average Grade and Grade Repetition

	Overall Grade				Grade Repetition	
	T1		T2		T2	
	M1	M2	M1	M2	M1	M2
Treatment	.286	.327*	.344*	.397**	−.353*	−.429**
	(.182)	(.191)	(.183)	(.193)	(.192)	(.207)
Covariates	No	Yes	No	Yes	No	Yes
Mentor fixed effects	No	Yes	No	Yes	No	Yes
Observations	117	117	117	117	117	117

Note.—Standard errors are in parentheses. Covariates (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, intrinsic motivation, and extrinsic motivation.
* $p < .10$.
** $p < .05$.

TABLE 7
Regression Analyses on Dutch and Mathematics Grades

	Dutch Grade				Mathematics Grade			
	T1		T2		T1		T2	
	M1	M2	M1	M2	M1	M2	M1	M2
Treatment	.259	.251	.247	.244	.309	.433**	.421**	.548**
	(.185)	(.189)	(.184)	(.193)	(.193)	(.212)	(.192)	(.210)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
Mentor fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	117	117	117	117	109	109	109	109

Note.—Standard errors are in parentheses. Covariates (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, intrinsic motivation, and extrinsic motivation.
** $p < .05$.

treated students felt less extrinsically motivated than control students, which is indeed what we see in table 8.

For the overall average grade in table 6, we see significant results at the 5 percent level for M2 at T2 and at the 10 percent level for M2 at T1 and M1 at T2. The last column of table 6 shows the results for grade repetition. As discussed above, the negative sign is as expected, since this is a reversed outcome. Table 6 shows that students in the treatment group had a lower likelihood of having to repeat a grade or dropping out, which is significant at the 5 percent level in M2.

The results of the regression analyses on the two separate subjects Dutch and math, in table 7, show that the positive significant results for the overall grade at T2 seem mostly driven by the improved math performance of the treated students. The results for mathematics show significant effects, at the 5 percent level, at both T1 and T2 (M2). Note that these math results also hold if we apply a Bonferroni correction for having three outcome

TABLE 8
Regression Analyses on Extrinsic and Intrinsic Motivation

	Extrinsic Motivation				Intrinsic Motivation			
	T1		T2		T1		T2	
	M1	M2	M1	M2	M1	M2	M1	M2
Treatment	−.311	−.239	−.601***	−.543***	.288	.409**	−.000542	.105
	(.189)	(.188)	(.193)	(.185)	(.202)	(.163)	(.201)	(.151)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
Mentor fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	99	99	99	99	99	99	99	99

Note.—Standard errors are in parentheses. Covariates for extrinsic motivation (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, and extrinsic motivation at T0. Covariates for intrinsic motivation (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, and intrinsic motivation at T0.
** $p < .05$.
*** $p < .01$.

measures (grades/grade repetition and two types of motivation), as the maximum *p*-value for significant results is then reduced to .02 (instead of .05). Although the coefficient for the difference in performance between the students who led their own meeting and the students who did not is positive in all models for Dutch as well, these differences are not significantly different from zero.

Table 8 show the results with respect to the two motivation outcome measures. From the first column of table 8, with extrinsic motivation as the outcome measure, a few things emerge. First, we see that the treatment has no significance in T1 but does have a significant effect in T2, with a standardized coefficient of 0.54 of a standard deviation, which is a medium effect. The former result implies that it might take some time to change the perceived extrinsic motivation. Second, we see that the coefficient more than doubles between periods T1 and T2. This result also holds after the Bonferroni correction, as the effect is significant at the 1 percent level, whereas the Bonferroni threshold is the 2 percent level.

In the second part of table 8, with intrinsic motivation as the outcome measure, we see that the coefficient becomes considerably smaller between T1 and T2 and that only the coefficient for M2 at T1 is significantly different from zero. Interestingly, there seems to be an effect on intrinsic motivation only in the short run, if we include all covariates.

A quantile regression on the full sample for mathematics shows that the effect is similar in magnitude for all the quantiles but significant only for the middle two quantiles.[4] For the extrinsic motivation, quantile regression shows that the effect is significant for three out of four quantiles, although the magnitude of the effect decreases slightly. Remember that extrinsic motivation is a reversed variable, where a higher score means feeling more pressure from others.

D. Gender Analysis

Table 9 shows the results of the analysis of potential differential effects by gender for the most extensive model (M2).[5] We first divided the sample by gender. When we compare the intervention group with the treatment group for only boys or only girls, there are no significant differences between the groups on observable characteristics (underlying tables can be found in app. B). The results in table 9 confirm the positive results we presented in tables 6–8. On all outcome variables, both the boys and the girls in the treatment group scored higher grades than their counterparts in the control group, and for motivation the signs are also similar to those presented above.

The results show that the positive significant effect of student-led study-progress meetings on performance that we presented above is primarily driven by the performance of boys (even though we had considerably fewer

[4] Results can be found in appendix B.
[5] The results for M1 and M2 can be found in appendix B.

TABLE 9
Regression Analyses on Different Groups by Gender

	Boys		Girls	
	T1	T2	T1	T2
A. Overall grade average	.823**	.813**	.246	.337
	(.312)	(.335)	(.256)	(.255)
B. Dutch	.309	.400	.148	.0734
	(.341)	(.327)	(.256)	(.264)
C. Mathematics	.877**	.968***	.276	.370
	(.323)	(.332)	(.324)	(.302)
D. Extrinsic motivation	−.105	−.909***	−.119	−.290
	(.275)	(.247)	(.229)	(.233)
E. Intrinsic motivation	.494	.269	.362*	.104
	(.366)	(.267)	(.195)	(.183)
F. Grade repetition		−1.035**		−.398*
		(.428)		(.237)
Observations (A, B, and F)	45	45	72	72
Observations (C)	45	45	64	64
Observations (D and E)	41	43	65	64

Note.—Standard errors are in parentheses. Regression model 2 (M2) includes covariates (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, intrinsic motivation, extrinsic motivation, and mentor fixed effects.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

boys in our sample than girls, which gave us a potential power problem). The positive effect for boys on overall average grade seems to be driven by mathematics, similar to when we considered the full sample. The same holds for the effect on grade repetition/dropout, although the results for girls are significant at the 10 percent level as well. However, all coefficients for girls are much smaller than the coefficients for boys, so the effects are much larger for boys.

For the motivation outcomes, we find a positive and significant effect of intrinsic motivation at T1 for girls, although only at the 10 percent level. However, this effect does not last, as it disappears at T2, similar to what we saw in the overall analysis in table 8. Furthermore, the coefficients for intrinsic motivation are not very different between boys and girls. On the other hand, the effects for extrinsic motivation again show a different story. We clearly see a significant large negative effect for boys only and a much larger coefficient, which appears to be the reason that we found a significant effect in the total sample.

Note that all these gender effects of boys also hold if we apply the Bonferroni correction, again with a reference significance level of 2 percent.

V. Robustness Analyses and Cost-Effectiveness

A. Robustness Analyses

To check the robustness of our results, we perform several additional analyses, shown in table 10. Table 10, first of all, presents the effect on

TABLE 10
Robustness Analyses Based on Subsamples

	Questionnaire Subsample			Mathematics Subsample			
	Overall Grade T2	Dutch T2	Math T2	Overall Grade T2	Dutch T2	Extrinsic Motivation T2	Intrinsic Motivation T2
Treatment	.446**	.154	.566**	.426**	.326	−.619***	.0316
	(.213)	(.212)	(.242)	(.201)	(.200)	(.184)	(.161)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mentor fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	99	99	92	109	109	100	100

Note.—Standard errors are in parentheses. Covariates (measured at T0): gender, curriculum, Cito score, grade/track year before, dyslexia, age, Dutch as first language, intrinsic motivation, and extrinsic motivation.
** $p < .05$.
*** $p < .01$.

the average grade, Dutch grade, and motivation outcomes for the subsample of students who followed math. Furthermore, table 10 presents the results for the outcomes overall grade, Dutch grade, and math grade for the subsample who filled out the motivation questionnaire. This is done to check that the effect we find is not due to the fact that we have a slightly different sample for Dutch than for math or to the fact that only a select group of students filled out the motivation questionnaires, for whom the intervention would have more effect. In both cases, we see that the results are very similar to the results in tables 6–8.

All in all, the robustness checks in table 10 show effects similar to those presented above.

B. Cost-Effectiveness

The costs for implementing student-led study-progress meetings are relatively low. At the school under study, the students worked at their portfolio during regular scheduled classes, so no extra lessons were added to the curriculum. As most schools have scheduled special mentor lessons in their timetable, there were no extra costs there either. The existing electronic learning environment was used for the students' portfolios. However, there were some extra costs: for mentor training (one-time-only cost) and for an overall coordinator (yearly costs). When student-led study-progress meetings were implemented for all students in grade 10 and 11 of the higher general education level (a total of 261 students) at the school under study, a training was organized for mentors to help them prepare for their role in the student-led study-progress meetings. The cost of this 4-hour training was €800 in total for 19 mentors. These mentors counseled a total of 261 students, so the cost of this training was €3.07 per student. Yearly costs must be paid to provide time for an overall coordinator in the current setup, although these are not only for the treatment but also for the total setup of study-progress meetings with this design. These

costs are relevant for schools that come from a different system of teacher-parent meetings, but they do not directly contribute to the cost of the intervention in the school under study. This person is responsible for the overall planning and preparation of the student-led study-progress meetings during the year and is the point of contact for the mentors. To perform that task well, the coordinator would need 2 hours per class per year. At the school under study, the 261 students were distributed across nine classes, which makes the total costs for the overall coordinator €900 (on average, a working hour of a teacher is estimated by the school administration at €50), which is €3.45 per student per year.

As discussed above, analysis of the students who had to repeat a grade or dropped out also shows significant effects. Given that the Dutch government spends approximately €7,000 per student per year (Teule 2012), this is an important finding, as a student who repeats a grade costs an additional €7,000 over his school career. If students can improve their academic achievement by leading their own study-progress meeting, it can be a relatively cheap way to increase academic performance and prevent students from having to repeat a grade, which in turn prevents these governmental costs.

Given the medium-sized standardized effect and the average cost of only €6.50 per student, the cost-effectiveness of this intervention is rather large.

VI. Conclusions and Discussion

A. Conclusions

In this paper, we study the short-term ATE of an intervention of increased student involvement, consisting of writing a portfolio on their own learning process and implementing student-led study-progress meetings, on academic results, motivation, and grade repetition of grade 10 students in upper secondary school. A randomized experiment was carried out during one school year with 130 students. Students in the treatment group had to prepare a portfolio and lead two study-progress meetings during the school year. Control group students did not write a portfolio and attended two mentor-led meetings. The results show that there was a medium-sized significant effect of increased student involvement on the overall average grade and the grade in mathematics. Although treatment group students also had higher grades for Dutch, no significant differences were found for that grade. The medium-sized effect is similar to what is found in the literature for the average of studies on metacognitive skills as well as on autonomy and feedback (for an overview chapter in which all these aspects are discussed, see Haelermans and Ghysels 2017).

Extrinsic motivation was significantly lower for students who had increased student involvement (also with a medium effect size), and in the short run intrinsic motivation was significantly higher. Furthermore, grade repetition was significantly lower for students in the treatment group. All

results, except for intrinsic motivation, can be completely attributed to the male students. Interestingly, the result on intrinsic motivation is fully due to the female students. Robustness analyses confirm the results. Quantile regressions show that the results are not limited to high or low achievers only.

B. Discussion

The existing literature on increased student involvement with portfolios and student-led study-progress meetings suggests that these meetings have positive effects on student performance, behavior, and motivation. Unfortunately, most of these studies lack statistical evidence for these claims. Our contribution to the literature is the use of a randomized experimental design, which allows for causal analysis. Randomization at the individual level increased the internal validity of this study. A limitation is the fact that the analysis is based on only one school in the Netherlands. However, we are confident that the results are also applicable to most of the secondary schools in the Netherlands, because this school is representative of the average secondary school in the Netherlands.

Composing a portfolio and having a meeting are done individually, which reduces the chance that there are crossovers or that control students' behavior is influenced by treatment students. However, spillover effects are possible, as students talk to each other over lunch breaks.

The most likely explanation for the significant and considerable effects we found in this study is the student involvement and the required metacognitive skills of the students, which were present in both the portfolio and meeting phases. The intervention forced the students to be active participants in their own learning process and thus to feel more responsible and be much more involved: in preparing the portfolio, during the meeting, and in updating the portfolio. Furthermore, the required increased autonomy and asking for and dealing with feedback are also aspects that most likely played a role in the positive effect. However, it is unclear whether the chosen division of attention to these three aspects is the most efficient. We would suggest further research to focus on separate aspects, or on a different share of time spent on each aspect, to determine which one of those is most important for the significant positive results.

Furthermore, unlike the expectations, we did not find positive effects on the intrinsic study motivation of students in the longer run, although we did find effects on extrinsic motivation, which implies that students felt less externally controlled. Although we would have expected that the decreased extrinsic motivation would lead to increased intrinsic motivation (following, e.g., Vansteenkiste et al. 2004) and also that the increased use of metacognitive skills would lead to higher intrinsic motivation, we did not observe this in our results. It is likely that this transition takes longer than the time span of our study. Students felt less externally controlled and pressured as a result of the intervention, but that does not

immediately mean that their intrinsic motivation increased in the longer run. The significant effect on intrinsic motivation in the short run implies that this might be the effect of doing something different, which might fade out after the first few months.

As results show, the significant effect on the overall average grade seems mainly driven by mathematics, as there was no significant effect on the grades for Dutch. It is possible that language skills are developed at a much younger age than math skills and that therefore language skills are much less influenceable at later age than math skills. Another possible explanation for this is that in their portfolio and during the meeting, many treatment students formulated so called "action" goals, such as doing more homework or paying more attention during class. It is likely that these goals would affect grades in mathematics more than the language grades, because of the practice repetition that is often more beneficial in math.

The results show that boys benefit more from this intervention than girls. The most plausible explanation is that the intervention is mostly based on using metacognitive skills, which girls might already possess to a larger degree. In line with this is the possible explanation that girls were already much more concerned with their own learning process, even when it was not institutionalized. This implies that there is much more to gain for boys from this experiment than there is for girls. It is also possible that the peer effects among girls are higher, such that there is more spillover in the acquired skills among girls between treated and control students. On the basis of the literature discussed in Section I, one would think that girls have higher intrinsic motivation. However, we did not see differences in intrinsic motivation scores between boys and girls, and additional analyses of the effect of the intervention at the bottom of the distribution of intrinsic motivation also do not show a difference, which is why we do not believe this a good explanation. Another potential explanation for the gender effects is the gender of the mentor. Previous literature points toward the existence of a same-gender effect, in which students perform better if they have a teacher of the same gender (Dee 2007). It is possible that this is the case here, as most mentors were male (6 out of 9 mentors), and only 23 students had a female mentor, of whom 16 were female as well. Unfortunately, these numbers in our data set are too small to statistically test this hypothesis. Finally, given that most mentors are male, one possible explanation for the strong results for boys or math is that these male mentors all teach science courses, such as math or physics. However, a closer look at the data shows that this was not the case. Four out of the six male mentors did not teach sciences.

As a policy implication, the results indicate that implementing this intervention in upper secondary school in only one grade level already gives positive and significant results. The costs for implementing this intervention are also quite low. It could therefore be worth considering implementing this intervention only in upper secondary school and still getting the benefits. However, it is possible that expanding this intervention

to multiple years in secondary school would generate an even greater effect. This should be studied in future research on this topic.

Furthermore, implementing this intervention does not require major adjustments in, for example, the curriculum or the organization of most schools in developed countries. Quite a number of schools already have once- or twice-yearly meetings where, in almost all cases, the mentor, the parent(s), and the student are present, especially in high school, and in upper secondary school, all students have idle hours in their schedule in which the portfolio writing could be scheduled without increasing school hours. To implement this, one needs to change only the roles of mentor and student with respect to preparing for and leading the meeting, implying higher student involvement.

Appendix A

Questions for the Two Motivation Subscales

A1. Intrinsic Motivation

I am motivated to study because:

36. I want to learn new things
39. I am interested in studying
42. It is an important goal in life for me
45. I like studying
47. I personally find studying very valuable
49. Studying is fun

A2. Extrinsic Motivation

I am motivated to study because:

35. I am supposed to study
38. I feel guilty towards others if I would not study
41. Others (parents, friends, teachers) oblige me to study
44. I would feel bad about myself towards others if I would not study
46. Others force me to study
48. I would be very disappointed in myself if I would not study

References

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Econ.* 6 (4): 453–70.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Rev. Econ. Studies* 70 (3): 489–520.

Coleman, Margo, and Thomas DeLeire. 2003. "An Economic Model of Locus of Control and the Human Capital Investment Decision." *J. Human Resources* 38 (3): 701–21.
Conti, Gabriella, James Heckman, and Sergio Urzua. 2010. "The Education-Health Gradient." *A.E.R.* 100 (2): 234–46.
Deci, Edward L. 1975. *Intrinsic Motivation*. New York: Plenum.
Deci, Edward L., and Richard M. Ryan. 2000. "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior." *Psychological Inquiry* 11 (4): 227–68.
Dee, Thomas S. 2007. "Teachers and the Gender Gaps in Student Achievement." *J. Human Resources* 42 (3): 528–54.
Donche, Vincent, Peter Van Petegem, Herman Van de Mosselaer, and Jan Vermunt. 2010. *LEMO: een instrument voor feedback over leren en motivatie* [LEMO: An Instrument for Feedback about Learning and Motivation]. Mechelen, Belgium: Plantijn.
Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics*. 4th ed. London, Sage.
Goodman, Amy. 2008. "Student-Led, Teacher-Supported Conferences: Improving Communication across an Urban District." *Middle School J.* 39 (3): 48–54.
Groot, Wim, and Henriëtte Maassen van den Brink. 2010. "The Effects of Education on Crime." *Appl. Econ.* 42 (3): 279–89.
Haelermans, Carla, and Joris Ghysels. 2017. "Evaluating Didactical Interventions in Education: Where Do We Stand?" In *Handbook of Contemporary Education Economics*, edited by Geraint Johnes, Jill Johnes, Tommaso Agasisti, and Laura López-Torres, 141–61. Cheltenham: Edward Elgar.
Hattie, John, and Helen Timperley. 2007. "The Power of Feedback." *Rev. Educ. Res.* 77 (1): 81–112.
Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2015. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." Working Paper no. 21178 (May), NBER, Cambridge, MA.
James, Harvey S., Jr. 2005. "Why Did You Do That? An Economic Examination of the Effect of Extrinsic Compensation on Intrinsic Motivation and Performance." *J. Econ. Psychology* 26 (4): 539–66.
Juniewicz, Kit. 2003. "Student Portfolios with a Purpose." *The Clearing House* 77 (2): 73–77.
Klomp, J., and S. Thielen. 2010. *Bovenbouw havo problematiek* [High School Middle Level Problems]. Heerlen, Netherlands: Ruud de Moor Centrum/OUNL i.s.m. Orion.
Lee, Jaekyung. 2016. *The Anatomy of Achievement Gaps: Why and How American Education Is Losing (but Can Still Win) the War on Underachievement*. New York: Oxford Univ. Press.
Martin, James E., Jamie L. Van Dycke, W. Robert Christensen, Barbara A. Greene, J. Emmett Gardner, and David L. Lovett. 2006. "Increasing Student Participation in IEP Meetings: Establishing the Self-Directed IEP as an Evidence-Based Practice." *Exceptional Children* 72 (3): 299–316.
Masui, Chris, and Erik De Corte. 2005. "Learning to Reflect and to Attribute Constructively as Basic Components of Self-Regulated Learning." *British J. Educ. Psychology* 75 (1): 351–72.
OECD (Organisation for Economic Cooperation and Development). 2015. "What Lies Behind Gender Inequality in Education?" *PISA in Focus*, no. 49.
Perry, Valerie, Loren Albeg, and Catherine Tung. 2012. "Meta-Analysis of Single-Case Design Research on Self-Regulatory Interventions for Academic Performance." *J. Behavioral Educ.* 21 (3): 217–29.

Ream, Robert K., and Russell W. Rumberger. 2008. "Student Engagement, Peer Social Capital, and School Dropout among Mexican American and Non-Latino White Students." *Sociology Educ.* 81 (2): 109–39.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Ryan, Richard M., and Edward L. Deci. 2006. "Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will?" *J. Personality* 74 (6): 1557–86.

Shute, Valerie J. 2008. "Focus on Formative Feedback." *Rev. Educ. Res.* 78 (1): 153–89.

Teule, Paul. 2012. "Wat kost zittenblijven nou echt?" Published July 25, 2012. http://sargasso.nl/wat-kost-zittenblijven-nou-echt/.

Tholander, Michael. 2011. "Student-Led Conferencing as Democratic Practice." *Children and Soc.* 25:239–50.

Tuinstra, Cheri, and Diana Hiatt-Michael. 2004. "Student-Led Parent Conferences in Middle Schools." *School Community J.* 14 (1) 59–80.

Vansteenkiste, Maarten, Eline Sierens, Bart Soenens, Koen Luyckx, and Willy Lens. 2009. "Motivational Profiles from a Self-Determination Perspective: The Quality of Motivation Matters." *J. Educ. Psychology* 101 (3): 671–88.

Vansteenkiste, Maarten, Joke Simons, Willy Lens, Bart Soenens, Lennia Matos, and Marlies Lacante. 2004. "Less Is Sometimes More: Goal Content Matters." *J. Educ. Psychology* 96 (4): 755–64.

Waddell, Glen R. 2006. "Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth." *Econ. Inquiry* 44 (1): 69–97.

Wehmeyer, Michael, and Margaret Lawrence. 1995. "Whose Future Is It Anyway? Promoting Student Involvement in Transition Planning." *Career Development Exceptional Individuals* 18:69–83.

Weiss, Michael J., J. R. Lockwood, and Daniel F. McCaffrey. 2014. "Estimating the Standard Error of the Impact Estimator in Individually Randomized Trials with Clustering." MDRC Working Paper on Research Methodology, Manpower Demonstration Research Corporation, New York.