Evaluation of a vocational school initiative to increase graduation

Fane Groes, Copenhagen Business School, (Principal investigator)

Anna Folke Larsen, ROCKWOOL Foundation Interventions Unit

Edith Madsen, Copenhagen Business School

Marcos Vera Hernandez, University College London, (Principal investigator)

Mikkel Stahlschmidt, ROCKWOOL Foundation Interventions Unit

Trial Information

Trial Title

Evaluation of a vocational school initiative to increase graduation and mental well-being through professional networks and authentic experiences with the profession

Country

Denmark

Keyword

Vocational schools; graduation; drop-out; mental well-being;

Abstract

This study examines the effect of an intervention at a 20 week's basic course at Danish vocational schools. The activities in the intervention are designed to give the students real-world experiences in the industry and a professional network. The intervention is delivered by existing personnel at the participating schools who participate in an on-boarding course prior to starting the intervention. The intervention is implemented in nine different educations at four vocational schools with multiple locations from spring 2022 to spring 2024. The intervention is rolled out using a clustered stepped wedge design. The starting time of the intervention is randomized at the education-location level within stratas. The primary outcome of the intervention is graduation from the basic vocational course. Secondary outcomes are continued education and mental well-being.

Trial Start Date

October 7th, 2021 (time of first randomization)

Sponsors and partners

The ROCKWOOL Foundation sponsored the development and implementation of the intervention and the evaluation. The evaluation is conducted by researchers at University College London, Copenhagen Business School and the ROCKWOOL Foundation Interventions Unit. As a part of the evaluation, the intervention is rolled out at three Danish vocational schools: Zealand Business College (ZBC), NEXT Uddannelse (NEXT) and Center for Erhvervsrettede uddannelser Lolland-Falster (CELF). We plan to recruit an additional vocational school for the evaluation. The National Center for Vocational Pedagogics (NCE) provides implementation support.

Rambøll Management Consulting administers the collection of survey data and conducts an implementation evaluation to complement the impact evaluation. The implementation evaluation is designed to assess the implementation capacity prior to roll-out of the intervention; the degree to which

the intervention is delivered as intended and any potential barriers to this; and from a qualitative perspective, which mechanisms that drive any impact of the intervention, and which subgroups of students that are expected to benefit the most from the intervention.

Study Design

Intervention(s)

The intervention is called OS I BRANCHEN (Us in the Industry) and represents a new approach to the 20 week's basic course at Danish vocational schools, called Grundforløb 2. Henceforth referred to with the common abbreviation GF2.

Background

The GF2 takes place at the vocational school as preparation for the main course. The main course consists of several apprenticeship and school modules. The intention of GF2 is to teach the students a range of vocational skills prior to their first apprenticeship.

The ROCKWOOL Foundation Interventions Unit has interviewed many GF2 students during the development of the intervention. Many students have chosen a vocational education because they want to be active in a workplace and use their skills to create value and make a difference for others. Quite a few have had bad experience with prior schooling and are eager to get out in the industry. Some students have limited experience with the industry and therefore need exposure to the industry to discover the possibilities and the everyday life in a company through practical experiences. Most students experience that the individual responsibility for their education and finding an apprenticeship is a heavy burden. The intervention is developed as a response to these student experiences. The intervention does not grant student apprenticeships, but elements of the intervention is meant to provide students with tools and experiences, which they can utilize in the process of getting an apprenticeship agreement with a company. Mostly, students carry the responsibility for getting an apprenticeship agreement with a workplace. However, a recent tri-party agreement on the vocational school area has shifted some of the responsibility of getting apprenticeship agreement away from the students towards the schools. The schools must then assist the students not able to get an apprenticeship agreement themselves. The tri-party agreement has not been fully implemented yet at the point of writing and therefore the effect of the responsibility shift is unknown, and there is uncertainty about the practical implementation of the tri-party agreement.

The intervention is being implemented and trialed at three vocational schools in Denmark. Zealand Business College (ZBC), NEXT Uddannelse København (NEXT) and Center for Erhvervsrettede Uddannelser Lolland-Falster (CELF). The intervention is being rolled out to the following nine educations or areas of education: Carpentry, Electrician, Bricklaying, Data & Communication, Car Mechanic, Plumbing, Social and Health Care (Social and health care assistants and social and health care helpers), Foods (Baker & confectioner, gastronome, waiter, nutrition assistant), and Business (Retail, trade, event coordinator).

The schools were recruited for the trial based on three parameters: They had a range of the educations which the intervention was designed for, the enrollment into these educations were large enough to utilize the network dynamics of the intervention, both in-between students and between the students and the companies, and lastly that graduation rates on the GF2 left room for improvement.

The intervention

The intervention rests on two principles:

- "The first day in school is the first day in the industry"

- "We succeed as a team"

The activities in the intervention are designed to give the students real-life experiences and a professional network. In all activities the students will try out the industry in a professional collaboration in a real world context. The intervention is delivered by existing personnel at the participating schools. The personnel participate in an on-boarding course prior to starting the intervention.

The intervention (hidden)

The intervention includes a range of activities throughout GF2 with a focus on the first two weeks, which have been found to be important for graduating GF2 (see Groes et. al. (2022)). The activities cover:

- Students are welcomed into the industry by meeting workplaces already on the first day of GF2.
- Students and workplaces meet during the first days of GF2 to establish connections and create new experiences for the students and workplaces.
- Doubts and uncertainties among students are normalized through exercises.
- Students watch inspirational film greetings that show different possible ways into the industry and how you can succeed despite initial challenges.
- Students are placed in network groups of 8-10 students to increase belonging and mutual support. Exercises and tasks that are part of the intervention are done in these network groups.
- Throughout GF2, students have a dialogue with workplaces through livestreaming to better connect GF2 curriculum to the coming apprenticeship.

Intervention Start Date

The first enrollment of students into the trial was January 2022.

Intervention End Date

The last enrollment of students into the trial will at the latest be June 2024[1].

Primary Outcomes

Short-term: graduation of the GF2 program at 8 months after enrollment

The intervention is at the GF2 program that lasts 20 weeks, excluding vacations. In a previous analysis of the timing of GF2-graduation (see Groes et al 2021), we have found that measuring graduation 8 months after starting the GF2 program is a time where we capture most of the graduates. At the eight months mark, we therefore measure student graduation from the GF2 program.

Secondary outcomes

The secondary outcomes originate from two data sources: students' answers to survey questions and data from the administrative record. Our first secondary outcome is the students' mental well-being, which we measure through survey questions. The second and third set of secondary outcomes are students' achievements and input factors, measured form the administrative records.

[1] This date can still change if updated power calculations point to the necessity of recruiting additional schools into the evaluation.

The secondary outcome for students' mental health is captured in a student survey. Rather than solely relying on the short-term effect to appear through graduation rate, we believe the intervention potentially can impact the mental well-being of students as students should feel more supported and motivated with their studies and experience that they belong and are able to add value to their new profession.

Mental well-being will be measured based upon seven questions from the Warwick-Edinburg Mental Well-being Scale (Clarke et al 2011, Koushede et al 2018, Powell et al 2013). Five of the questions are taking from the Short Warwick-Edinburg Mental Wellbeing Scale while the last two have been chosen because we believe that they measure what the intervention is trying to achieve. We will construct one mental wellbeing score by taking the sum of the seven questions.[2]

The other secondary outcomes are explorative and originates from the administrative records. First, they capture the students' potential of a better education match, the students' achievement, measured as the students' grades, and the students' probability of signing an apprenticeship. Second, they capture input factors measured as student and teacher attendance, which may change because of the intervention.

The intervention is meant to give students a better chance to graduate from the GF2 program. However, graduation from the GF2 program is not the final outcome for the vocational students' education because after the GF2 program comes the vocational main program of around 3 years after which the student receives a certification for the labor market. We therefore include a secondary outcome measure 8 months after initial enrollment to see if the students are enrolled in *any* education or have graduated from the GF2 program. This measure at 8 months after initial enrollment will capture both the students who graduated the GF2 program and it will also capture students who have started a different education.

The intervention is meant to give students a better idea from the beginning of the program about whether the chosen education is the right one. This means that one consequence of the intervention could be that some students will drop out earlier when they realize that the specific education is not a good match. Some of these students who drop out because of a bad match may have improved skills due to the intervention that can help them choose another education than the one they drop out from (such as the ability to network, a better understanding of the industry, and understanding drop-out as learning rather than failure). These students should be better prepared to start a different education and stay with it. By measuring students' graduation and any other educational enrollment 8 months after initial enrollment, we capture both the restarters and the students who graduated from the main program.

The students receive grades in some of the subjects that are not practical (e.g. math). The practical exam is a pass/fail. If the intervention increases the probability of graduating from the GF2, we could expect the grade distribution to improve. The average grade of the graduates will depend on the composition of the graduates and it is therefore not clear if the average grade, conditional on passing the exam, will increase or decrease. However, as a fraction of all the enrolled students, we should expect the grades to improve. To measure whether the intervention has an impact on the high end of the grade distribution, we will measure the effect of the intervention on a variable that takes value one if the student achieves a high grade and takes value zero if the student passes but achieves a non-high grade, the student fails, or if the student drops out, which capture all the enrolled students.

We hope to include a measure of the probability of obtaining an apprenticeship at a firm and the date at which the agreement with the firm is made. If the intervention is successful at connecting the students more to the industry and that the network groups helps, we expect that the fraction of students finding an apprenticeship with a firm increases just as the fraction of students having an apprenticeship at the

[2] We also include survey questions on life satisfaction using the cantril ladder now and in five years, but in pilot-testing of the questionnaire we have found relatively high life satisfaction among the population of interest. Thus, we worry about ceiling effects and have therefore chosen not to include life satisfaction as a prespecified outcome.

school should decrease[3]. We will use the False Discovery Rate (Benjamin and Hochberg 1995, Andersen 2008) to obtain q-values for the hypotheses tests corresponding to the following two outcomes: probability of a high grade and probability of obtaining an apprenticeship.

Finally, we expect the students' attendance to increase if the students can see an added benefit to coming to class from the intervention. The teachers' attendance could also be affected if students' higher motivation leads teachers to be more motivated and less likely to suffer from stress or other sicknesses as a result of the intervention. We will use the False Discovery Rate (Benjamin and Hochberg 1995, Andersen 2008) to obtain q-values for the hypotheses tests corresponding to the following two outcomes: students' attendance and teachers' attendance.

Outcomes that will be used to explain the mechanisms at play

Through the student survey, which the student answers at week 3 and 13 of the GF2, we also ask questions covering topics such as: (1) self-efficacy, (1) feeling of "succeeding as a team", (3) sense of belonging to the profession, (4) perception of the curriculum, (5) feeling of having value, (6) Number of contacts with workplaces. These six concepts are tightly related to the underlying theory of change of the intervention.

(1) Self-efficacy[4] will be measured through an index which includes the answers to 2 questions which the research team has formulated. (2) Feeling of "succeeding as a team" will be measured using an index which includes the answers to 3 questions. (3) Sense of belonging is measured through an index including 2 questions uncovering if the student is uncertain about their belonging in the profession. (4) perception of curriculum is measured through 1 question uncovering if the student finds the acquired skills and experiences in GF2 to have value later on. (5) Feeling of having value is related to if the student feels like they can contribute to the workplace and vocation of their choice. (6) number of contacts with workplaces is measured as one question asking about this. We will use the False Discovery Rate (Benjamin and Hochberg 1995, Andersen 2008) to obtain q-values for the hypotheses tests corresponding to the following six outcomes: self-efficacy index, succeeding as a team index, sense of belonging on the profession index, perception of curriculum, perception of having value, and number of contacts with workplaces.

Subgroup analysis

In this subsection, we explain the subgroup analyses that we plan to perform. We expect heterogeneity in impact to stem from two different sources: 1) Heterogeneous effect across individuals within clusters, e.g. across students with different background characteristics. 2) Variation in impact across clusters stemming from variation in the degree of ex-ante intervention capacity as well as the degree to which the intervention is implemented as intended (fidelity).

To assess heterogeneous impact across individuals, we will estimate the effect of the intervention on those students who were "Not in Education, Employment or Training" (NEET) before they enrolled in one of the education-locations of the study. This is a group which is particularly relevant from the policy point of view.

[3] The apprenticeship outcome will potentially be affected by a reform that will be introduced fully from 2023, which alters the Vocational Schools' economic incentives of finding apprenticeships. In our analysis, we will consider how to adjust for the reform, depending on the final implementation.

[4] Self-efficacy refers to an individual's belief in his or her capacity to execute behaviors necessary to produce specific performance attainments. Self-efficacy reflects confidence in the ability to exert control over one's own motivation, behavior, and social environment.

Furthermore, as a part of the implementation evaluation conducted by Rambøll, they will qualitatively assess which group of students that appear to benefit most from the intervention and what their background characteristics are. Based on this knowledge, we will use administrative data to form a subgroup with these background characteristics and estimate the effect for this particular group. Within this group elicited from the implementation evaluation, we will correct for multiple hypothesis testing using the False Discovery Rate.

To estimate other individual level heterogeneous effects, we will use a machine learning algorithm that endogenously defines the relevant subgroups and is compatible with our estimation method.

We will also estimate heterogeneous effects according to fidelity in the implementation of the intervention and ex-ante implementation capacity.

Fidelity in the implementation of the intervention is assessed every semester through a questionnaire to school staff followed by a discussion where nuances are captured in qualitative notes and through a few questions in the student survey. These data can be used to group education locations according to their level of fidelity when implementing for the first time. We will use the grouping as done by Rambøll as they will conduct qualitative interviews which we expect will bring valuable information. Since all education locations in the evaluation sample will implement the intervention sooner or later it is also possible to group control observations according to future fidelity. However, we acknowledge that the measure may be endogenous to implementation timing, e.g. education locations starting later having higher degrees of fidelity.

We will also estimate heterogeneous effects according to implementation capacity. Prior to rolling out the intervention, Rambøll carried out a phone survey on implementation capacity to all education leaders of all education locations. We will group the education locations according to their initial implementation capacity using the information included in the phone survey conducted by Rambøll.

Data sources

We plan to use seven different data sources, four types of administrative data and three surveys.

The first and main administrative data source is from Statistics Denmark and is a not yet completely developed register data set that contains detailed information on students and teachers in the population of Danish Vocational Schools. This data set, which we will refer to as our VET data, is data extracted directly by the IT developer of the schools administrative systems and processed by Statistics Denmark. In the contract with the IT developer, Statistics Denmark has ordered student data on start- and end date, program of enrollment, graduation status as pass/fail or with a grade on the courses enrolled in, courses taken during enrollment, attendance and absence at the hourly level for all classes during a courses, location of the classes by date and hour, and the scheduled teacher for the class by date and hour. Statistics Denmark has further requested all teachers' attendance and absence at the daily level. We would ideally like to use this VET data for all register data outcomes and indicators for whether a student belong to the treatment group or not, which can be found by the education-program-location and semester. The VET data will according to Statistics Denmark be available by December 31, 2023. Therefore it should be possible for us the use the VET data in our evaluation of the intervention. However, since this data is only currently being collected, processed and validated, we are concerned that the data will contain some measurement error (e.g. inaccurate id for the education-location or start and end date), which will affect our outcomes and assignment to treatment. We therefore include two other data sources in the hope to reduce measurement error on the treatment assignment and the outcomes and to be able to include background characteristics of the students and perhaps the teachers.

The second data source is data that the evaluation team has requested directly from the three intervention schools. This data is collected at the beginning of every semester and contains data on all

students enrolled in GF2 at the given school. The data includes information on students' id, education, program, and location for the given semester. We will use this information on semester-education-program-location to verify which students are assigned treatment at a given time. It is possible to link this data to the Statistics Denmark VET data by student id.

The third data source is the traditional education data from Statistics Denmark that contains information on students' education spells, education and program enrolled in, graduation status of a given spell, and information on dropout without completing a course. This data set is called KOTRE. We hope not to use this data for the main outcome, which is graduation status 8 month after enrollment in the GF2 program, since we prefer to use the information from the VET data where we also have information on absences and teachers. However, in the case that the VET data turns out to be unreliable, we will use graduation status from KOTRE. For the secondary outcome on graduation or enrollment in other educations, we will need to use the KOTRE data, since the outcome is enrolled in ANY education, not just the VET educations.

The fourth data source is also traditional Statistic Denmark data that we will include for background characteristics. This data includes information on demographics (e.g. student age, parents' id, immigration status, municipality), prior education (e.g. completed a high school degree prior to enrollment or already dropped out once from the GF2 program), status prior to enrollment (e.g. employed, in school, on benefits, or without any income), and parents' education and income at a given age of the student. See section on background characteristics.

Three surveys are being conducted and made available by Rambøll Management Consulting. The first survey is a student survey. The second survey is a phone interview with educational leaders uncovering the implementation capacity before the intervention begins. The third survey is fidelity measurement among teachers and relevant school personnel uncovering whether the intervention is delivered-as-intended.

The first survey is a survey among students. The survey is distributed and conducted during the 3rd the 13th week of GF2. The survey consists of 20 questions covering the following topics: Life satisfaction, mental well-being, fidelity of the intervention, self-efficacy, feeling of "succeeding as a team" (One of the two principles behind the intervention), sense of belonging and having value and lastly the student's perception of the curriculum. The survey is rather short but uses two validated instruments (Cantril's Ladder for Life Satisfaction and a (modified) short Warwick-Edinburg Mental Well-being Scale for the mental well-being questions) and 11 self-made questions untangling the mechanisms behind the theory of change of the intervention. The survey can be linked to the register data using a student id. The survey also works as a triple check on the education-location of the students in each semester to ensure that we assign the correct students to the treatment and control. We attach this survey.

The student surveys are scheduled to be done in-class by the vocational teachers across treatment and controls groups in each semester from fall 2022 to the end of the trial period. The in-class setup is chosen because we expect this to give a higher response rate compared to just inviting students to participate in the survey online outside the school program. Alongside this, we will send text messages to students as a reminder to fill out the survey if they haven't done it. The propensity to respond might be correlated both with treatment status and with the answers to the survey. To deal with selective response, we will induce random variation in the response rate by randomizing the number of reminders each student receives. The students are randomized into three categories, which will determine the numbers of reminders they get: 0, 1 or 2. The randomization for this is done at the student level. The random variation in the number of reminders received will allow us to correct for the non-random response rate using a Heckman selection model.

A second survey is also conducted by Rambøll and serves as an implementation capacity measure on how ready the schools are to implement the intervention. Rambøll conducts phone interviews with leaders of education at the participating education-locations once before the education-location begins implementing the intervention. Questions spans topics such as: the educational leader's vision for GF2, relationship between educational personnel, conditions for undertaking new actions and initiatives, past experiences with changes, if the students have experiences already which are closely connected to the

intervention and the underlying theory of change, cooperation with firms and workplaces, absenteeism and dropout, organizational relationships, leader's experiences from earlier employment. Some questions are closed-ended, and some are more open-ended. If possible, these implementation capacity measures will be supplemented with information from the register data, such as teacher turnover. The investigator team will construct a measure of the ex-ante implementation capacity of each education-location. The goal is to pursue heterogeneity analyses and examine whether specific implementation conditions are needed to achieve a large treatment effect. The implementation measure is also important for information about the possibility of scaling the intervention.

A third survey conducted is a fidelity measurement, where teachers and other education personnel answer questions on whether the intervention was delivered-as-intended. The intervention instructions include three purposes for each activity. The implementing staff are allowed to adjust activities as long as they follow these three purposes. We measure fidelity as the degree to which the staff consider these purposes fulfilled. This survey is conducted every semester towards the end of GF2. Questions cover topics such as: seniority in position, whether the task of giving the students the desired experiences was accomplished (experiencing the industry, feeling a sense of belonging, normalized uncertainty, students being in network groups and supporting each other etc..), the teacher's perception of curriculum (whether the goal of the intervention is clear and meaningful) and if the intervention/curriculum has created new connections to co-workers around the intervention and lastly if their feeling of job satisfaction has increased. The survey is followed by a facilitated discussion among school staff on how to improve the implementation of the intervention. Rambøll will construct a measure of fidelity based on the survey responses and qualitative notes from the following discussion.

Background Characteristics

From the standard administrative data available at Statistics Denmark, we can extract a detailed set of background characteristics. From Groes et. al. (2021) we have found the following background characteristics to be important: Information on age, sex, immigration status, and parental identifiers. For individuals who graduated 9th grade after 2002, we observe the grades in the year. Highest completed education at time of enrollment, and parental education at the time of enrollment. We also include the most recent education prior to GF2 enrollment. We also include parents' labor market status and labor market status of students during the year prior to enrollment. We also have the possibility to include health data, such as visits to the local general practitioner and purchase of prescription medicine related to mental health. The prescription data is however uncertain if we can include in the analysis, due to a long lag in obtaining the data from the Danish Health Agency.

We plan to include data from student's mental wellbeing as measured in a nationwide student survey at year 9. In principle, including this information in the model should help to improve precision, especially for the mental well-being outcome. However, this data will only be available for those who are relatively young, as the survey is recent.

Population Selection

The population affected by the intervention is students enrolled at GF2 at the education-locations in the evaluation at the three vocational schools in the period January 2022 to June 2024. We select the student population from the administrative data. It is possible to include periods before the intervention started. We will be guided by the chosen estimation method.

Most students start in a class in January or August, while Social and Health care students also have full class intake in April and October. A smaller share of students starts at odd times and is enrolled in existing GF2 classes, and we do not expect these students to benefit fully from the intervention. Hence, we only include January and August enrollment, plus April and October classes for the Social and Health care educations.

There are different ways to enroll into GF2. We only include enrolled students in "skolevejen" or "praktikvejen". They will follow ordinary G2 courses at the vocational school. This means that we exclude students who take apprenticeship-based GF2 programs from the analysis. We will also disregard online students to the extent that we can identify them in the data. We further remove students who are 14 years or younger when starting GF2 (these students cannot give consent for the survey).

Identification

The identification of the treatment effect of the intervention on the outcomes will come from the stratified random staggered roll out of the intervention. See below for further details on the randomization.

Estimation

The econometrics literature that exploits the staggered roll out of a policy has expanded very significantly in the last four years (see for instance Borusyak et al 2022, Chaisemartin and D'HaultfŒuille 2018 and 2022, Freyaldenhoven et al 2021, Gardner 2021, Goodman-Bacon 2021, Roth et al 2022, Sun and Abraham 2020, Roth et al 2022). However, most of the literature has focused on non-random staggered rollout, in which identification of the treatment effects relies on the common trends assumption.

Recently, both Athey and Imbens (2022) and Roth and Sant'Anna (2022) have focused on the random staggered rollouts. Athey and Imbens (2022) assumes that the treatment effect is homogeneous with respect to the adoption date (invariance to history). Both papers assume a non-clustered design, in which the adoption date of a unit *i* is randomized, and the outcome for such unit *i* is observed during the entire time period. However, in our set up, we observe the outcomes of repeated cross sections (the students of a particular education-location in each semester, who will be different from semester to semester). Some of the ideas proposed by Athey and Imbens (2022) seem relatively straightforward to adapt to the repeated cross-section case. However, in order to use the methods proposed by Roth and Sant'Anna (2022), we would need to average the data at education-location-semester level, but this would lead to loss of power unless the intracluster correlation was one. Because the literature on staggered roll outs is evolving very quickly, we prefer not to select an estimation method at the moment of writing the pre-analysis plan.[5]

Some secondary outcome variables are measured using a student survey. We anticipate the response rate to be different depending on whether the intervention is being implemented or not at the time of the survey. To deal with possible selection bias, we will use a Heckman Selection model in which the randomized reminders to answer the survey will be used for the exclusion restriction.

Randomization method

The randomization, or the order in which the education-locations can begin implementing the intervention, was done in-person at the ROCKWOOL Foundations premises.

[5] If the chosen evaluation method was underpowered, we would explore using vocational schools outside the experimental sample as comparison clusters.

The randomization was stratified. Education locations were grouped into stratas according to the type of education that they provide (i.e. carpentry) as well as logistical constraints. The stratification allows to minimize the chance that all locations of a given education, say all locations which teaches Carpentry, started the intervention at the same time. When the number of education-locations was too low for a given education, we grouped several different educations in the same strata.

Each strata had pre-defined sequences which determined the roll out scheme over a five semester period. The pre-defined sequences took into account logistical constraints such as not starting too many education-locations at the beginning of the trial. The education-locations where then randomly picked and applied to a sequence, which determined when the education-location location was going to implement the intervention.

The randomization was done by randomly picking downward-facing cards, which identified the individual education-locations. The order of the draw determined the roll out scheme for the stepped wedge trial.

The directors of education at the partnering schools participated in the draft to increase local support for the roll-out scheme among the implementation teams. ROCKWOOL personnel also attended the draft. The randomization was recorded and saved for documentation.

Randomization unit

The unit of randomization is an education-location, e.g. Carpentry in the city of Slagelse. Within some areas of education, different educations are taught together in GF2: All food educations are taught together, the same are all business educations and also the two social and health care educations. To avoid spillovers, we grouped these educations together that take lessons together during GF2.

One school can have several education-locations for each education. In some cases, the same vocational teachers work at multiple locations, and here we grouped the education-locations together into one cluster to avoid spillovers.

Was the treatment clustered?

Yes, the timing of the treatment was randomized at the education-location level.

Sample size: Planned Number of Clusters

At the time of writing the pre-analysis plan, we have recruited 3 school which provide 33 education-locations. We are planning to recruit an additional school into the study which will provide additional education-locations.

Sample size: Planned Number of Observations

The number students participating in the trial, will depend on future enrolment into vocational schools and specifically the education-locations participating in the trial.

Based on intake from historical administrative data, we estimate around 7600 students, which wil increase when we recruit an additional school.

Sample size (or number of clusters) by treatment arm

Because of the staggered rollout design, all clusters end up receiving the intervention at some point.

Power for the main outcomes (accounting for sample design and clustering)

Following Athey and Imbens (2022), we have used the following model to compute the power of our experiment:

$$y_{ijst} = \alpha_0 + \beta_1 D_{jst} + \beta_2 (Strata)_s * (Semester)_t + \beta_3 X_{ijst} + \theta_j + \varepsilon_{ijst},$$

where y_{ijst} is a dummy variable that takes value 1 if student *i* from education-location *j* of stratum *s* graduates in semester *t* from GF2 and 0 if not, D_{jst} takes value 1 if the intervention is being implemented in education-location *j* of stratum *s* at semester *t* and 0 otherwise, $(Strata)_s * (Semester)_t$ are Strata-Semester dummy variable interactions, X_{ijst} are covariates, θ_j are education-location fixed effect, and ε_{ijst} is an error term. The parameter of interest is β_1, and we clustered the standard errors at the education-location level.

We compute the power of our design by block bootstrap (drawing the existing data from the administrative registry for the students enrolled in the period August 2015 to June 2020, in which each education-location is a block) and imposing an effect size constant between January 2018 and June 2020. We confirm that this leads to consistent estimates of β_1, and with a power of 0.70 for an effect size of 5 percentage points, and 0.83 for an effect size of 6 percentage points in two-tail tests at 5% significance. We expect the power level to improve as we are planning to recruit another school into the study.

Potential future analysis

A different type of analysis is related to peer effects. If the intervention works through the network groups, we could expect to see higher effects of, for example, being in a class with a student with parents that have a vocational school degree, or students with higher GPA from lower secondary school. It is also possible analyze the peer effects of being in a class with students with a mental health diagnosis and how this is affected by the intervention. Information for the peer effects is for a second analysis, subject to detecting a direct effect on the students from the intervention.

Supporting Documents and Materials

Ethics approval

Was Ethical approval obtained?

Ethical approval was received by the Ethics Council at Copenhagen Business School, and the UCL Research Ethics Committee (Project ID: 1827/008).

References

Anderson, Michael L. 2008. 'Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects'. *Journal of the American Statistical Association* 103, no. 484 (December 2008): 1481–95. https://doi.org/10.1198/016214508000000841.

Athey, Susan, and Guido W. Imbens. 2022. 'Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption'. *Journal of Econometrics*, Annals Issue in Honor of Gary Chamberlain, 226 (1): 62–79. https://doi.org/10.1016/j.jeconom.2020.10.012.

Benjamini, Yoav, and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, no. 1 (1 January 1995): 289–300.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. 'Revisiting Event Study Designs: Robust and Efficient Estimation'. arXiv:2108.12419. arXiv. https://doi.org/10.48550/arXiv.2108.12419.

Chaisemartin, C de, and X D'HaultfŒuille. 2018. 'Fuzzy Differences-in-Differences'. *The Review of Economic Studies* 85 (2): 999–1028. https://doi.org/10.1093/restud/rdx049.

Chaisemartin, Clément de, and Xavier D'Haultfoeuille. 2022. 'Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey'. Working Paper 29691. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w29691.

Clarke, Aileen, Tim Friede, Rebecca Putz, Jacquie Ashdown, Steven Martin, Amy Blake, Yaser Adi, et al. 2011. 'Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): Validated for Teenage School Students in England and Scotland. A Mixed Methods Assessment'. *BMC Public Health* 11, no. 1 (21 June 2011): 487. https://doi.org/10.1186/1471-2458-11-487.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse Shapiro. 2021. 'Visualization, Identification, and Stimation in the Linear Panel Event-Study Design'. Working Paper 21–44. Federal Reserve Bank of Philadelphia. https://econpapers.repec.org/paper/fipfedpwp/93518.htm.

Gardner, John. 2021. 'Two-Stage Differences in Differences'. mimeo. https://jrgcmu.github.io/2sdd_current.pdf.

Goodman-Bacon, Andrew. 2021. 'Difference-in-Differences with Variation in Treatment Timing'. *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 254–77. https://doi.org/10.1016/j.jeconom.2021.03.014

Groes, Fane; Madsen, Edith; and Sandoy Tróndur, 'Completion from Vocational Educations: A Register Based Analysis'. Mimeo Copenhagen Business School 2021. https://fanegroes.net/onewebmedia/Completion%20from%20Vocational%20Educations_Final.pdf

Groes, Fane; Madsen, Edith; and Sandoy Tróndur, 'Absenteeism and Completion among students in Vocational Education'. Mimeo Copenhagen Business School 2022.

Koushede, Vibeke, Mathias Lasgaard, Carsten Hinrichsen, Charlotte Meilstrup, Line Nielsen, Signe Boe Rayce, Manuel Torres-Sahli, Dora Gudrun Gudmundsdottir, Sarah Stewart-Brown, and Ziggi Ivan Santini. 'Measuring Mental Well-Being in Denmark: Validation of the Original and Short Version of the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS and SWEMWBS) and Cross-Cultural Comparison across Four European Settings'. *Psychiatry Research* 271 (1 January 2019): 502–9. https://doi.org/10.1016/j.psychres.2018.12.003.

Powell, John, Thomas Hamborg, Nigel Stallard, Amanda Burls, Jaime McSorley, Kylie Bennett, Kathleen M. Griffiths, and Helen Christensen. 2013. 'Effectiveness of a Web-Based Cognitive-Behavioral Tool to Improve Mental Well-Being in the General Population: Randomized Controlled Trial'. Journal of Medical Internet Research 15, no. 1 (9 January 2013): e2240. https://doi.org/10.2196/jmir.2240.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. 'What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature'. Paper. arXiv.org. https://econpapers.repec.org/paper/arxpapers/2201.01194.htm.

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2022. 'Efficient Estimation for Staggered Rollout Designs'. arXiv:2102.01291. arXiv. https://doi.org/10.48550/arXiv.2102.01291.

Sun, Liyang, and Sarah Abraham. 2020. 'Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects'. *Journal of Econometrics*, December. https://doi.org/10.1016/j.jeconom.2020.09.006.