Pre-analysis plan for "Offering scholarships to adult skilled workers: Evidence from a randomized field experiment", Version 2.0

Dapi, Bjorn
bjorn.dapi@fafo.no

Drange, Nina
n.e.drange@frisch.uio.no

Hauge, Karen E.
k.e.hauge@frisch.uio.no

Zhang, Tao
tao.zhang@frisch.uio.no

January 23 2025

Abstract

This document outlines the plan for analyzing the impact of providing an educational scholarship to adult skilled workers, using round 1 data. Note that this document was written up before receiving and analyzing round 1 data. Round 1 data are expected to be delivered in within June 2023. We will produce a similar document before the analysis of round 2 data, which are expected to be delivered in April 2024.

Updated in version 2: As stipulated in version 1 of this plan, we now update the plan for analyses of round 2 data. There are no changes in the research questions, hypothesis, or experimental design. This updated version adds a plan for defining the outcome measure completion testing hypothesis 2, which is now possible to analyse with the updated round 2 data. All changes made are highlighted in the text. List of changes made in version 2 of the pre-analysis plan:

- Section 2.1 Hypotheses
- Section 3.3 Coding of Round 2 outcome variables
- Section 4.3 Exploratory analyses

1 Experimental design

We designed a large-scale, randomized field experiment conducted in Norway in the educational years 2021/2022 and 2022/2023. The field experiment was in collaboration with The Norwegian Directorate for Higher Education and Skills (HKdir), the Norwegian State Education Loan Fund (a public bank that provides loans and scholarships to students, "Lånekassen"), and Statistics Norway. The experiment aims to find the causal effect of offering an educational scholarship on educational enrollment and completion, and if so, whether this gives any labor market consequences.

The Norwegian Directorate for Higher Education and Skills (HKdir) commissioned designing and evaluating an experiment for testing the effect of offering an educational scholarship to skilled adult workers.

Statistics Norway was responsible for selecting the study population based on data from the Norwegian population and sending out all information to participants on behalf of HKdir.

The Norwegian State Education Loan Fund administrated the scholarship's application and payment.

All information sent out as part of the experiment was developed in collaboration between HKdir, Lånekassen, Statistics Norway, and the research team. A smaller-scale pilot was conducted in 2021, and the design was improved somewhat before the main experiment was conducted on a larger scale in 2022. We will rely on data from both the pilot and the main in our analysis.

1.1 Definition of study population

Our definition of a skilled worker is a person who has either completed vocational secondary education, and either has received a diploma ("vitnemal" in Norwegian), attained a certificate of apprenticeship ("fag- brev"), or attained a journeyman's letter "svennebrev"). There are several routes to accomplishing this:

- completed three years in school, resulting in a diploma
- completed two years in school and two years of apprenticeship
- completed three years in school and a following apprenticeship

The main (pilot) study population consists of people who fulfilled the following three criteria:

1. Residents of Norway in December 2021 (December 2020),
2. Fulfill the criteria for being defined as a skilled worker, that is: Registered in Statistics Norway's statistics of education in December 2021 (December 2020) as having completed vocational secondary education and either have received a diploma ("vitnemål"), attained a certificate of apprenticeship ("fagbrev") or attained a journeyman's letter ("svennebrev") as their highest completed education, and
3. Aged as of 30 years old and up to and including 57 years old in December 2021 (December 2020).

Statistics Norway operationalized this by selecting individuals having the highest completed education classified as 354 in the international ISCED attainment standard.

1.2 The scholarship

The educational scholarship could be used for secondary, college, or university education in all fields of study and in courses that started between August and June in the following study year, given that the education is eligible for support from the Norwegian State Education Loan Fund. Following the common practice of the Norwegian State Education Fund (Lånekassen), the scholarship is paid out as a loan that is transformed into a scholarship when the receiver completes the education or passes the exams. The maximum size of the scholarship was NOK 50 000 for the equivalent of one year's full-time study. The scholarship is scaled for part-time studies lasting less than one year. For example, a full-time half-year study and a 50% part-time study over one year will give a scholarship of NOK 25 000.

Only individuals who received a personal offer were eligible for the scholarship.

1.3 Treatments and randomization

We designed an experiment to test whether providing an educational scholarship can increase adult education among skilled workers. The experiment consists of two treatments: Scholarship and information groups.

Date	Scholarship group	Information group
Jan 31 2022	Information about project and scholarship opportunity	Information about project
Feb 14 2022	Reminders about education application deadlines	Reminders about education application deadlines
Mar 1 2022	*Application deadline for all secondary and some tertiary education*	
Apr 4 2022	Reminders about education application deadlines	Reminders about education application deadlines
Apr 15 2022	*Application deadline for most tertiary education*	
May 20 2022	*Application results available for some vocational schools*	
May 30 2022	Information about specific scholarship and application form	General information about scholarship opportunities
Jul 22 2022	*Application results available for tertiary education*	
Aug 2022	*Semester start secondary and tertiary education*	
Nov 15 2022	*Application deadline: financial support from Lånekassen for autumn semester 22*	
May/June 2023	*Expected delivery of Round 1 data from Statistics Norway to research team*	

Table 1: Timeline of experiment with important dates and information content sent to scholarship and information group.

The scholarship group received emails offering a scholarship if they enroll in education during the following study year (August-June) and information about what possibilities exist regarding further education.

To rule out the possibility that the information provided alongside the scholarship offer drives a potential effect, we included **the information group** who received the same information regarding educational opportunities and general funding possibilities without the specific scholarship offer.

We did not contact **the control group**, implying they did not receive any emails. Emails can be found in the appendix.

Important dates and the timeline of the experiment are outlined in Table 1.

Statistics Norway used a script (in SAS) provided by the Frisch Centre for obtaining a random sample of the study population (defined in Section 1.1

Group	Pilot	Main
Scholarship	1500	61 800 (21%)
Information	1500	169 950 (57%)
Control	1500	67 252 (22%)
N	4500	299 002 (100%)

Table 2: Sample size by treatment groups

above) of approximately 300 000 people to the main study (and 4500 people to the pilot) and randomly allocating individuals into three groups. Table 2 illustrates the sample size by treatment groups. The sizes of the groups in the main experiment were based on simulations conducted after the pilot to maximize power; see Appendix 6 for further details.

Statistics Norway sent all emails to the study population (as defined in section 1.1 above) using the common contact register (in Norwegian "Kontakt og reservasjonsregisteret"). The common contact register is a national register of updated contact information used by the central and local government, including email addresses to more than 95% of the Norwegian population[1]. Although the common contact register has a very high coverage of email addresses within the Norwegian population, some addresses will be missing. In addition, we do not know with certainty whether the participants received the emails or whether they were read. The experiment, therefore, should have an intention-to-treat interpretation.

2 Research questions and hypotheses

We are interested in the impact of offering educational scholarships to adult skilled workers. The first data delivery, which we will term Round 1 data, will contain information about enrollment in education by October 2022. We expect to receive these data at the earliest in April 2023. The first impact of offering educational scholarships we thus can measure is the impact on enrollment in education. Later data deliveries, which we will refer to as Round 2 data (from April 2024 onwards), will contain information about

[1]In the pilot experiment, the sample consisted of 4500 people. Of these, 5 individuals were no longer in the Norwegian registers (either due to death or emigration), and of the remaining 4495 individuals, 4392 individuals (98%) were found in the common contact register. In total, emails were therefore sent to 4392 individuals.

educational completion and labor market consequences.

2.1 Hypotheses

We have chosen the following three main hypotheses:

1. **H1** Receiving an offer of a skilled worker scholarship of NOK 50 000 (for a year of full-time study) increases education **enrollment** among skilled workers.

2. **H2** Receiving an offer of a skilled worker scholarship of NOK 50 000 (for a year of full-time study) increases education **completion** among skilled workers.

3. **H3** Receiving an offer of a skilled worker scholarship of NOK 50 000 (for a year of full-time study) increases **earnings** among skilled workers 3-5 years after the offer is received.

Updated in version 2: This updated plan outlines our plan for analyzing H1 (unchanged) and H2 (new). We will produce a similar document outlining the plan for analyzing H3 before receiving the relevant data.

3 Data and Coding of variables

We expect to receive the first wave of data from Statistics Norway in April 2023. The study relies on two main data sources: 1) the randomization data set and 2) the outcome and covariates delivered from Statistics Norway's administrative register data. The variables and datasets from Statistics Norway can be granted by application to Statistics Norway and following their guidelines, which can be found here: https://www.ssb.no/en/data-til-forskning.

3.1 Coding of treatment variables

The randomization data set includes data about the outcome of the random draw, that is, which treatment group each participant ended in. We construct the dummy variables $Scholarship$, $Information$, and $Control$, which each equal one if the individual is in the scholarship, information, or control group, respectively.

3.2 Coding of Round 1 outcome variables

Uptake in education can be measured on the extensive margin (whether or not the person is enrolled in a study program) and the intensive margin (the magnitude of the planned study progression). We choose extensive margin as our main outcome variable. The continuous intensive margin could increase power, but as we do not know the quality of this variable nor the mean for our population, we have chosen not to use this variable as our main outcome variable. Variables on educational outcomes are from Statistic Norway's dataset "Utdanning"[2].

As our main education enrollment measure, we will construct the dummy variable *Enrolled* that measures enrollment in education. Specifically, we base this on the variable for ongoing education from Statistics Norway called "igang_2022", which records if a person is registered in education in October of any given year, in our case, in October 2022 (2021 for the pilot sample). The dummy variable will take the value 1 if the person has a value different from missing on this variable and 0 otherwise.

As an alternative measure of enrollment, we will use the percentage of planned study progression based on the Statistics Norway variable "studieprogresjon_h" measuring the planned study progression in the fall semester and "studieprogresjon" measuring the planned study progression in a given semester. Statistics Norway use this variable to define a student as a full-time student when the planned study progression is higher than 70% and a part-time student when the planned study progression is 70% or lower. We will define the variable *studyprogression* as the share of study progression from zero (not studying) and one (studying full-time) reported by October 2022.

3.3 Coding of Round 2 outcome variables

Updated in version 2: To measure study completion, we will use data from Statistics Norway and the Norwegian education database (Norsk utdanningsdatabase, NUDB) file F_UTD_KURS also called the course table ("kurstabellen"). The Round 2 data contains data from NUDB updated as of October 1, 2023. When a person starts a new education, this is captured in the course table by a new start record ("tilgangsrecord") with "$KODE = 1$".

[2]https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/utdanning

There are unique IDs for a course, but a course can be everything from an integrated bachelor's degree to a single course in a college. When a person completes a course or an education, this is captured with an exit record ("avgangsrecord") with "$KODE = 0$" and the variable "$UTFALL = 8$", defined as completed and passed. We will define completion as when there is a new start record and exit record after the experiment started, with the same course ID.

More specifically, we will create a dummy variable that takes the value 1 if two criteria are met: 1) the start record ("tilgangsrecord") for a course ID is set to "$KODE = 1$" in the school year of 2022-2023 (2021-2022 for the pilot), that is after June 1 2022 (June 1 2021 for the pilot), and 2) the exit record ("avgangsrecord") for the same course ID is set to "$KODE = 0$" and the variable "$UTFALL = 8$".

In an exploratory analysis, we will also describe whether the different groups work to the same extent as before throughout the year when they can receive the subsidy. For this, we will use monthly data from A-ordningen in Statistics Norway, where we observe the percentage of scheduled work hours. We will use their primary source of employment and show descriptive results in an event model for the information, scholarship, and control groups, respectively, from August 2022 (2021 for the pilot) to June 2023 (2022 for the pilot). The main source of employment will be defined as the firm where they have their main position as of June 2022 (2021 for the pilot).

3.4 Coding of covariates

If not specified otherwise, control variables will be measured in the year before the intervention.

Covariates are from Statistic Norway's dataset Population ("Befolkning")[3], Income ("Inntekt")[4], Employment ("a-ordningen")[5] and Education ("Utdanning")[6].

[3]https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/befolkning

[4]https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/inntekt

[5]https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/a-ordningen

[6]https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/utdanning

Variable name	Source dataset	Source variable	Variable definition
Age	Population	$Fødselsår$	Continuous variable
Female	Population	$Kjønn$	Dummy **female**=1 if female
Two foreign-born parents	Population	$invkat$	=1 if $invkat = C$ or $invkat = B$
Children	Population	$barn3\_i\_fam$	continuous variable, the numb of children in household
Lives with partner	Population	$sambo\_snr\_åååå$	=1 if $sambo\_snr\_åååå! = .$
Mother education	Education	$NUS2000\_MOR\_16$	
Father education	Education	$NUS2000\_FAR\_16$	
In education	Education	$IGANG\_XXXX$	
Unemployment benefits	Income	arbledtrygd	=1 if $arbledtrygd > 0$
Disability benefits	Income	uforetrygd	=1 if $uforetrygd > 0$
Employed	Income	wyrkinnt	$wyrkinnt > 2G$
Income	Income	wyrkinnt	Continuous variable
Debt	Income	Gjeld	Continuous variable
Partner's income	Income	wyrkinnt	Continuous variable

Table 3: Control variables

4 Empirical approach

4.1 Balance tests

To test for balance, we will compare background characteristics measured one year (per December in t-1) prior to randomization for each of our two treatment groups relative to the control group and the information treatment relative to scholarship treatment. First, we will test for balance between the Scholarship group and the Control group:

$$Scholarship_i = \alpha + \beta_1 X_{i1} + u_i \tag{1}$$

where i indexes individuals, t is time (where t=1 implies before the treatment and t=2 implies after the treatment), X_{it} is a vector of controls variables listed in Table 3. Further, individuals with $Control = 1$ are set to 0, and individuals with $Information = 1$ are excluded.

Second, we will test the balance between the Information group and the Control group:

$$Information_i = \alpha + \beta_1 X_{i1} + u_i \tag{2}$$

where individuals with $Control = 1$ are set to 0 and individuals with $Scholarship = 1$ are excluded.

Finally, we will test the balance between the Scholarship group and the Information group:

$$Scholarship_i = \alpha + \beta_1 X_{i1} + u_i \tag{3}$$

where individuals with $Information = 1$ are set to 0 and individuals with $Control = 1$ are excluded.

We will judge whether the randomization was successful by conducting an F-test of whether the control variables jointly predict each treatment status.

4.2 Hypothesis testing

To test H1, whether a scholarship offer increases education enrollment, we will test whether education enrollment is higher in the scholarship group relative to the control group. We also include the information group to rule out the possibility that the information provided alongside the scholarship is a confounding factor.

The estimation equation can be summarized as follows:

$$Y_{i2} = \alpha + \beta_S * Scholarship_i + \beta_I * Information_i + X_{i1} + \epsilon_i \tag{4}$$

where i indicates individuals and t is time (t=1 then implies before the treatment and t=2 after the treatment), and X_{it} is a vector of control variables (see below).

The above specification estimates the following three effects:

- β_I gives the isolated effect of receiving information about further education.
- β_S gives the combined effect of receiving information and a scholarship.
- (β_S–β_I) gives the isolated effect of receiving a scholarship.

Assuming successful randomization, the coefficients β_1 and β_2 will have a causal interpretation. We assume that receiving a scholarship offer will not reduce the probability of taking further education. Therefore, we will use a one-sided test for the test concerning β_S. The effect of received information about further education is less clear-cut, so we will perform two-sided tests

when testing our hypotheses concerning β_I. We will follow the commonly used inferences by using significance levels of 0.05; if p-values are under 5%, the coefficients will be interpreted as evidence that the treatment(s) has (have) an impact.

We will include variables that fail the balance test as controls (irrespective of the result of the F-test of joint significance). We will also conduct analyses where we select control variables through a double robust LASSO procedure (Belloni et al. 2014) to increase power. This is a data-driven approach for including a small set of influential confounds from a large set of potential variables.

Standard errors: As the sample is a random draw of the entire population of skilled adult workers in Norway, and the unit of random assignment to treatments is at the individual level, it is unnecessary to cluster the standard errors (Abadie et al. 2023). We use robust standard errors in all estimations[7].

TOST: In the case of null results, we will test for equivalence using TOST (two one-sided tests) to rule out effects larger than the smallest effect size of interest, and rule out effects larger than the minimum detectable effect size outlined in section 5 below, suggested by Lakens (2017).

4.3 Exploratory analyses

We will also explore treatment heterogeneity across all baseline variables described in Table 3. The treatment heterogeneity analysis will use the honest causal forest method (Wager & Athey 2018). In addition to the machine learning analysis described above, we will investigate the following pre-defined subgroups in an exploratory analysis:

- Company size (less than 20 employees vs. others)
- Immigrant background (first and second-generation immigrants pooled)
- Gender

Update in version 2: In addition to the variables in Table 3, we will add the following in the honest causal forest analyses: pre-treatment field of education (see categories below) and pre-treatment industry (based on the international standard NACE, see categories below).

Fields of education:

[7] using the Stata command reg y x, vce(robust)

- Studies qualifying for higher education ("almenne fag")
- Health, welfare and sport
- Humanities and aesthetics
- Teaching and pedagogy
- Nat.sciences, vocational and techn. subjects
- Primary industries
- Transport, communications, safety, security, other services
- Social science and law
- Business and administration

Industries:

- NACE Section A - Agriculture, Forestry and Fishing
- NACE Section B - Mining and Quarrying
- NACE Section C - Manufacturing
- NACE Section D - Electricity, Gas, Steam and Air Conditioning Supply
- NACE Section E - Water Supply, Sewerage, Waste Management and Remediation Activities
- NACE Section F - Construction
- NACE Section G - Wholesale and Retail Trade
- NACE Section H - Transporting and Storage
- NACE Section I - Accommodation and Food Service Activities
- NACE Section J - Information and Communication
- NACE Section K - Financial and Insurance Activities
- NACE Section L - Real Estate Activities

- NACE Section M - Professional, Scientific and Technical Activities
- NACE Section N - Administrative and Support Service Activities
- NACE Section O - Public Administration and Defence, Compulsory Social Security
- NACE Section P - Education
- NACE Section Q - Human Health and Social Work Activities
- NACE Section R - Arts, Entertainment and Recreation
- NACE Section S - Other Service Activities
- NACE Section T - Activities of Households as Employers
- NACE Section U - Activities of Extraterritorial Organisations and Bodies

5 Power considerations

Our estimate of the proportion who will enroll in education in the control group is 3.1%. To test the hypothesis that a scholarship offer affects the proportion who enrolls in education, and using a p-value of 0.05 and 80% power, we have a minimum detectable effect size (MDE) of 0.00027[8]. This means that the MDE we can detect is an increase in the proportion who enrolls in education as small as 0.0027, that is, an increase in enrollment proportion from 3.1% to 3.28%. The proportion of 3.1% enrolment would thus constitute 1915.8 people enrolled in education (61800*0.031), while 3.28% would constitute 2027 people, which is an increase of 111 persons. When testing one hypothesis, the experiment has enough power to detect an increase in enrollment of 111 people in the treatment group.

In addition, we want to rule out the effect of information (but here, we also have an additional group of people to include in the tests). To test the hypothesis that information affects the proportion who enroll in education,

[8]Stata command: power twoproportions 0.031, test(chi2) power(0.8) n1(67252) n2(61800). Here the assumption is that there is approx. 61800 people in the scholarship group and 67252 in the control group

and using a p-value of 0.05 and 80% power, we have an MDE of 0.0022[9]. This means that the MDE we can detect is an increase in the proportion who enrolls in education due to the information sent as small as 0.0022, an increase in enrollment proportion from 3.1% to 3.22%.

6 Appendix

Insert simulations done prior to the main experiment to calculate sample sizes and division between treatments.

References

Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. M. (2023), 'When should you adjust standard errors for clustering?', *The Quarterly Journal of Economics* **138**(1), 1–35.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'High-dimensional methods and inference on structural and treatment effects', *Journal of Economic Perspectives* **28**(2), 29–50.

Lakens, D. (2017), 'Equivalence tests: A practical primer for t tests, correlations, and meta-analyses', *Social psychological and personality science* **8**(4), 355–362.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

[9]Stata command: power twoproportions 0.0301, test(chi2) power(0.8) n(237202) nratio(2.53). Here the assumption is that there is approx. 170000 people in the information group and 67000 in the control group, and thus a ratio of 2.53 in group size between the two groups.