Pre-analysis plan for "Effects of speed-schools in Niger"*

Anne Kielland[a] and Andreas Kotsadam[b]

Abstract

We study the effects of speed-schools for older children in Niger. In this plan we describe the hypotheses to be tested and how they will be tested. The description includes how the variables are coded, how we will deal with missing values, and the specification of the estimation equations. We also conduct a power analysis which suggests that we are able to identify relatively small differences. All deviations from the plan will be highlighted in the final paper.

*The research has been funded by the Norwegian Research Council (number 267550).
a Fafo Institute for Labour and Social Research.
b The Ragnar Frisch Centre for Economic Research, Gaustadalleen 21, N-0349 Oslo, Norway.

1. Introduction

The project is studying the effects of a two-year speed-school intervention in Niger that targets rural out-of-school children between 12 and 14 years of age that have never started or dropped out of primary education. The program covers the first six years of the primary school curriculum in two years. It is intended to be a low cost solution that provides a second chance to vulnerable children. We work closely with an NGO, the Strømme Foundation, that has been a pioneer in the speed-school field. Many other organizations have followed and speed schools are relatively common across Africa.

We study the effects of the speed-schools using a randomized controlled trial (RCT). We carried out a two-stage randomization, one at the community level, resulting in treatment and control communities, and the other at the individual level, resulting in "treatment" and "spillover" children in treatment villages, and "pure control" children in control villages. In this plan we describe how we will analyze the data from the experiment.

2. The field experiment

Speed-schools, also known as accelerated learning programs, are intended to let overaged children catch up on their primary education by offering a compressed syllabus. We study such a program in Niger that compresses grades 1-6 into two years. This program is labelled Stratégie de Scolarisation Accélérée 2 (SSA2). Due to an astonishing need in our areas and the limited spaces in the speed-schools, not all children that need and want a speed-school can attend one. Our main intervention consists of children with similar need randomly being offered a place in a speed-school. Additionally, we investigate the effects of a community receiving a speed-school. The experiment is randomized at two different levels:

Randomization across communities: Our experiment takes place in 85 communities in two regions. 30 communities were randomized to treatment and 55 to control. Initially we wanted to have 60 control communities but this was not possible due to too few eligible children in some communities initially listed. It was decided that there would be 15 treated communities in each of two regions covered (Dosso and Tillaberi) and therefore randomization to treatment and control communities were done within regions. In the control communities we listed 1,430 children and interviewed 1,159 children at baseline. In the follow up we will try to reach also children not interviewed at baseline.

Randomization across individuals: In the treatment areas we listed 1,786 children and interviewed 1,489 children at baseline. The schools wanted around 28 individuals per school and it was suggested that we randomized around 35 individuals to treatment as it was deemed likely that treatment compliance would be imperfect. When more than 45 children were listed in a community, we randomly selected 15 to be in the control group. In communities with fewer than 45 individuals listed we assign all non-treated individuals to the control group. Randomization was stratified by gender in the following way: In areas with more than 30 girls we had a quota of 50 percent girls in the treatment group. In areas with fewer girls on the list we randomly assigned treatment to girls in relation to the share of girls on the lists (35*share of girls). The remaining places were randomized to boys. All girls not receiving treatment were allocated to the control group and we filled the rest of the control group places with boys that were not randomly assigned to treatment.

3. Data and coding of main variables

We conducted a listing exercise to map the need in different communities in our two regions. Children were listed based on the following criteria: household within what is locally described as the "academic distance" from village center/prospective Speed-school location (5 km), age (around 12-14), willingness to go to school, parental support and consent to participate in this as a research project. Based on these lists we selected 85 communities and carried out the randomization.

We then collected baseline data in August to November 2018. Baseline interviews were blinded and treatment status was revealed after data collection in the village was finished. Treatment started shortly after baseline data was collected. The intervention was completed in June 2020 and the second round of data collection will take place during December 2020-April 2021.

Data was collected from community leaders, the household heads of the children on the lists, and the children on the lists. We here describe how we will use the survey data and how the variables will be coded.

Before going through the variables we start by describing some general principles in order to avoid repetition:

Standardized indices: When we say that we will use a standardized index we sum the values of the variables, subtract the mean of the control group and divide by the standard deviation of the control

group. The control group refers to control children in treated communities. Sometimes we use other types of indices but then that will be explicitly stated.

Indicator variables: When creating a dummy variable from continuous variables, the following procedure is used unless otherwise stated. We first code all non-responses to missing. The binary variable will then be set in order to create 2 groups, as similar in size as possible, while respecting the order of the answers (e.g. from 1 to 4). For example, if 25% answer 1, 25% answer 2, 25% answer 3 and 25% answer 4, then the binary variable will be equal to 1 if the respondent answers 1 or 2 and zero otherwise. On the other hand, if for example 20% answer 1, 10% answer 2, 5% answer 3 and 65% answer 4, then the binary variable will be equal to one of the respondent answers 1, 2 or 3. We will only consider the answers of the respondents in the control group when creating the coding rules for the binary variables.

3.1 Primary and secondary dependent variables

A goal of this project is to provide a comprehensive picture of the impacts of speed-schools on the children, their families, and the communities. We will therefore collect data on a large number of outcomes and we will report the full breadth of the evidence. That being said, power considerations (see below) forces us to limit the number of main hypotheses tested. We group the outcomes into 6 groups and for each group we will have either an index or a main variable to be tested. We have chosen the following main groups:

1) Schooling and learning outcomes.
2) Marriage and fertility.
3) Support for violence.
4) Self-esteem and wellbeing.
5) Hazardous child labor and time use.
6) Gender attitudes.

3.1.1 Schooling and learning outcomes

The overarching program goal of SSA2 is for out-of-school children to transfer back into lower secondary level of the traditional schooling system in Niger. Lower secondary school is divided into academic and vocational tracks (CEG/CET and CFM).

Our main measure for the educational outcomes is *Starting lower-secondary education*. This variable is coded as zero for everyone answering No on the question "Are you enrolled in school or education program this fall?". It is coded as 1 if they answer yes and then answer CEG/CET or CFM on the follow up question "What type of education program is that?". We also replace the variable to be equal to one if the respondent answers lower secondary school on the question on the highest grade level achieved.

We will also create a secondary variable, *Qualified for secondary*, equal to 1 if the child answers yes to the question "Did you do the Government test to qualify for lower secondary this summer?" and answer CEG/CET or CFM on the question "What was the result of the Government test?". Other secondary outcomes will be years of schooling, highest grade level achieved, and results from reading, writing and math tests. We code the following:

Years of schooling will be coded as the continuous answers to the question, "For how many years did you go to school, not counting pre-school?". Missing values will be replaced by zero for individuals answering No to the screening question "Have you ever been to school?"

Highest grade level. This variable will be coded as ranging between 1-7 where 7 refers to all levels above 6. Together with years of schooling we can see how much they advance from a year of speed-school since we have both of these variables at baseline as well

We have several variables intended to measure learning. First we code a variable, *Illiterate*, to be equal to one if children answer no on the question "Can you read and write French?". If they answer yes we code them as zero if the manage to write their own name on a piece of paper we provide them and if they can read the sentence "Fatima likes going to school, but she also likes to help her father". We also create a *Numeracy index* ranging from 0 to 4 where one point is given for being able to solve each of the problems we give them on addition, subtraction, division, and multiplication.

Concerning life skills, we will look at a composite score for knowledge and practices related to nutrition (food with protein), environmental awareness (disposal of plastic bags), malaria prevention (know about bed nets) and Covid-19 (ability to list 4 preventive measures and 2 vulnerable populations, last washed hands).

We have many school related variables that can be used to investigate compliance in our setting.

We will code a dummy variable *Offered SSA2* to equal 1 if the child answers yes on the question: "After we were here two years ago, can you remember receiving the offer to go to the SSA2 in this village?".

We will also describe the answers to the questions, "Did you start the program right away, when the SSA2 first started?" and "Did you complete the program?". In investigating non-compliance we also have questions about why children did not start and why they quit.

There may also be imperfect compliance because the control group may attend the speed-schools. We will code a variable, *Attended the speed school*, which equals one if the child answers yes on the question "In the past two years, have you been attending any school or education program?" and SSA2 on the question "What type of school or program did you attend?".

3.1.2 Marriage and fertility.

Our main measure will be *Appropriate marriage age* which will be equal to the answer on the question "What do you think would be a good age for you to marry?". Non-numerical answers will be coded as missing.

We will also explore the following variables as secondary variables:

Married: Dummy equal to one if answering yes on "Are you married or engaged?"

Parent: Dummy equal to one if answering yes on "Do you have any children?"

Desired fertility: The numerical answer to "How many children would you like to have?" The value will be set to equal to 15 if the numerical value is above 15. We expect quite a few children to answer that it is in the hands of God or similar fatalistic answers and we will code a dummy variable for fatalistic desired fertility as well.

3.1.3 Support for violence

We hypothesize that randomization to SSA2 reduces participants' own use of violence and also the accept for the use of violence more in general. Participants' own inclination to use of violence is

assessed by score on the variable "In general, I am able to resolve my problems without using violence". We label the indicator "*Own violence*".

Participant accept for the use of violence more in general is proxied by an index based on the following variables:

In certain cases, it is acceptable to use violence to defend health and security of my family
In certain cases, it is acceptable to use violence to defend my private property
In certain cases, it is acceptable to use violence to defend the honor of my family
In certain cases, it is acceptable to use violence to defend political principles
In certain cases, it is acceptable to use violence to defend religious principles

All variables are scored on a scale from 0 (Strongly disagree) to 10 (Strongly agree). We will dummy code the variables based on our general rules and then create the index indicator "*Violence accept*".

Niger, like the other countries in the region, is increasingly experiencing armed attacks, notably from misguided individuals citing religious motives, often labelled jihadists. It is worth noticing that the way Islam is practiced in Niger, no matter religious affiliation, inspires moral and self-discipline, not the use of violence.

For the purposes of the study, our interest is if being randomized to SSA2 may lower probability of jihadist sympathies, represented by the combination of religious intolerance combined with an acceptance for violence. With regards to acceptance for violence, our proxy is a continuous variable for *Accept religious violence*: "In certain cases, it is acceptable to use violence to defend religious principles", which is coded on a scale from 0 (Strongly disagree) to 10 (Strongly agree).

With regards to *Religious intolerance*, we use a continuous variable based on: "If my friend abandons the faith, I will avoid him", also coded on a range from 0 (Strongly disagree) to 10 (Strongly agree).

Since our religious radicalism is a function of (1) religious intolerance and (2) acceptance of violent and/or illegal means (Ozer and Bertelsen 2018) we create a composite dummy variable equal to 1 if the sum of the two variables is > = 10, and each indicator gets at least a score of 5. To make a more precise proxy for jihadist sympathies we recode to zero respondents who on the question "Why do you think some young people these days chose to join an armed group" select response category a; "to gain personal power", b; "to enrich themselves" or j; "they are bandits". Similarly, we code to zero those

who, to the question "What should government do to prevent young people from joining armed, violent groups" respond a) prohibit radical preaching. We call this variable *Radical* and it will be our main measure.

We will also explore possible variation in the combination on violence acceptance and high scores on other indicators of religious intensity such as religious identification, number of times prayed, number of days of religious studies, religious studies a s main activity over past 30 days, experience of religion as strength in life, and acceptance of female missioning.

We will also explore the answers to the following questions, which will be dummy coded:

I'd like to have friends with other religions than mine.
If you think about yesterday, how many times did you pray?
Religion and faith are sources of strengths to me
I'd like to have friends with other religions than mine
I'd like to have friends from other ethnic groups than mine
If my friend becomes pregnant outside of marriage, I would avoid her
People who go to live non-muslim countries often get detached from their religion and their traditions

In addition we will look into answers to the questions: "Why do you think some young people these days chose to join armed groups" and "What should government do to prevent young people from joining armed, violent groups".

3.1.4 Self esteem and wellbeing

Our main variable will be *Self-esteem.* We use a validated French translation of the Rosenberg self-esteem scale (Vallieres and Vallerand 1990) which contains the following 10 survey questions:

I am generally satisfied with myself.
Sometimes, I think I'm no good.
I think a have several good qualities
I'm able to do the same things as other people
I don't have much to be proud of
Sometimes, I certainly feel useless
I feel that I'm as valuable as anyone else
I wish I had more self respect
I generally think that I'm a failure.
I have a positive attitude to myself

The answer alternatives are: 0 Strongly agree, 1 Agree, 2 Disagree, 3 Strongly disagree. Items 2, 5, 6, 8, 9 are reverse scored. We therefore give “Strongly Disagree” 0 points, “Disagree” 1 point, “Agree” 2 points, and “Strongly Agree” 3 points on these items. We then sum the scores for all ten items and keep scores on a continuous scale. Higher scores indicate higher self-esteem. A score below 15 (out of the total 30) is usually interpreted as low self esteem and we will also create a dummy for this.

We will also explore effects on happiness:

Happiness: We will reverse code the answers to the question, How happy are you? (1 Very happy 2 Quite happy 3 Not very happy 4 Not at all happy).

As well as the following variables:

In the past two weeks, how often have you felt a low interest for or little pleasure in doing things?

In the past two weeks, how often have you felt sad, depressed or filled with hopelessness?

Since the last time we were interviewing for this study, two years ago, would you say that your life has become better, worse or is about the same as then?

Today, how well would you rate your life conditions on a scale from 0-10, where 0 is very bad, and 10 is perfect?

In general, how would you describe your life conditions compared to those of other young people in the village?

In general, how would you describe your future prospects compared to those of other young people in the village?

How would you rate your own health on a scale from 0-10, where 0 means very bad, and 10 means excellent?

3.1.5 Hazardous child labor and time use

Our main variable will be a dummy variable, *Hazardous child labor*, which is equal to 1 if the child answers that they during the past two years have worked in at least one of the following types of work places:

A household of someone who were not a relative
A mine
A quarry
The streets in a city
On a farm handling chemicals
On a farm where one could sense the smell of chemicals
Carrying loads that one felt were painfully heavy

We will also code a variable, *Working*, equal to one if the child does answer anything but No to the question "Last week, did you work and get paid in cash or kind?". The other response categories are: Yes for cash, Yes for in kind, or Yes for both cash and in kind payment.

We will also create other types of work variables by investigating answer to the question "What would you say was your main activity in the past 30 days?". In particular we will create dummies for domestic work, farm work, market work, business, other street work.

Acknowledging that most child labor in Africa is non-remunerated, we will also explore effects on time use by investigating effects on the different time use variables that come after the following in the survey: "We would like to learn about how you spend your time. If you think about the 7 days of the past week: I will read out some activities, and you can tell me if you have done that during the last week, and if so, how many times".

3.1.6 Gender attitudes

We have a section in the survey with attitude questions after the prompt: "I will read out some different statements to you, and I would like for you to tell me if you agree or not. You should answer exactly what you'd like. There are no right or wrong answers." The answer categories are from 0 to 10 where 0 is strongly disagree and 10 is strongly agree.

Our main outcome will be a gender equality index based on dummy coding of the following variables:

It is more important to send boys to secondary school
Girls should be allowed to study in higher secondary school even if it is far away
When remunerated work is rare, men should have priority to the jobs

The dummy codings follow our general rule for such codings where we split the variables so that we get as equal sized groups as possible but we will code the dummies so that 1 is always the more gender equal answers.

We will also explore answers to the question about gendered social norms regarding moving away to study. This is based on the question "In this community, most people think girls should be allowed to study at the higher secondary level, even if it is far away"

We will create a variable, *Share of cross sex friends*, with the help of the following two variables: "How many friends do you have, that you would say are close friends?" and "How many of those close friends are girls?".

In addition we will explore answers to the questions "In certain cases, it is justifiable for a man to beat a woman", and "Women should be allowed to leave their villages to go on mission/Spread the word of God."

3.2 Main independent variables from the baseline survey

All main independent variables are from the baseline surveys. The value codings and response categories are often different from baseline to endline but as we use ancova specifications (see below) and have general coding rules for dummy variables, this is not a problem.

We will always have the strata variables (female and community fixed effects for the individual analysis and region fixed effects for the community analysis) as independent variables.

The other main independent variables are:

<u>Continuous numerical values of:</u>

Years of schooling: Already defined.

Desired fertility: Already defined.

Self-esteem: Already defined.

HH respondent age: Numerical value of the age of the household respondent.

Asset index: Based on the principal component analysis of the following items:

Does the household have/own… (recoded as Yes=1, No=0)?

1. *Radio*
2. *Fridge*
3. *Cellphone*
4. *Bicycle*
5. *Moped*
6. *Motorbike*
7. *Car*
8. *Electricity*
9. *Solar panel*
10. *Generator*
11. *Push cart*
12. *Cistern or water reservoir*

The asset index will be standardized by subtracting the mean of the control communities and dividing by the standard deviation of the control communities.

Dummy variables for:

Religious intolerance: "If my friend abandons the faith, I will avoid him/her."

Accept religious violence: "It is acceptable to use violence to defend religious principles."

Gender Equality[1] *school*: "It is more important to send boys to secondary school"

Gender Equality work: "When remunerated work is rare, men should have priority to the jobs"

Gender Equality school: "It is justifiable for a man to beat a woman if she: burns the food"

HH respondent male: Equal to 1 if the household respondent is male.

3.3 Other outcome variables

These variables will be used as exploratory outcomes and to investigate mechanisms.

3.3.1. S*ocial desirability index.*

[1] Note that the gender equality variables are coded such that 1=gender equal.

We will create a *Social desirability index* to investigate whether experimenter demand effects are upward biasing the estimated program impacts, The index is based on responses on a Marlowe-Crowne module which measures a person's general tendency to give socially desirable answers. We use a validated French translation (Daigle 2019). The index is based on the following questions where individuals can answer agree or disagree:

i. It is sometimes hard for me to go on with my work if I am not encouraged.
ii. I sometimes feel resentful when I don't get my way.
iii. On a few occasions, I have given up doing something because I thought too little of my ability.
iv. There have been times when I felt like rebelling against people in positions of authority even though I knew they were right.
v. No matter who I'm talking to, I'm always a good listener.
vi. There have been occasions when I took advantage of someone.
vii. I'm always willing to admit it when I make a mistake.
viii. I sometimes try to get even rather than forgive and forget.
ix. I am always courteous, even to people who are disagreeable.
x. I have never been irked when people expressed ideas very different from my own.
xi. There have been times when I was quite jealous of the good fortune of others.
xii. I am sometimes irritated by people who ask favours of me.
xiii. I have never deliberately said something that hurt someone's feelings.

We can also test for heterogeneous treatment effects based on the social desirability score. The worrisome pattern would be if the treatment effects were driven by kids with a high propensity to disingenuously give socially desirable answers and vanished for those with a low such tendency.

For some of the children we have administrative data on the results from the Government test to qualify for lower secondary school. We will compare these to the answers the children give on *Qualified for secondary* and also relate the differences to the social desirability index.

In addition, there are some variables that are extra interesting to see how they correlate with social acceptability, such as "In general, I am able to resolve my problems without using violence" and the attitudes toward domestic violence.

3.3.2 *Migration*

We will measure *Intend to migrate,* which captures international migration and will be coded to equal 1 if the child answers yes to "Do you think you will actually move away from the village in the future?" and "Do you think you may travel to far away destinations, such as Europe?".

We will also measure national migration by investigating the answers to the questions:

Do you ever think about that you would like to travel to live outside this vilage?
If you should go live outside this village, where would you like to live? (you can chose several)
Are you—alone or with friends or family—seriously considering migrating to another country?

For those that have already moved we will also describe where they have migrated to the extent possible.

3.3.3. Other outcomes

We will also explore the effects on other variables.

3.4 Variables used to investigate spillovers and for LASSO

Some variables will be used as additional controls when using a LASSO procedure, and some to dig into mechanisms.

3.4.1 *Additional variables to investigate spillover effects*

In investigating spillover effects we will additionally investigate if there are treatment effects on the following variables:

We would now like for you to think about what these siblings were doing in the past two years.
Did any of your siblings start school in the fall of 2019 (last year)?
Was that a boy or a girl or both?
Did any of your siblings quit school in the past two years?
Was that a boy or a girl or both?
Did any of your siblings under the age of 18 leave the village to earn money in the past two years?
Was that a boy or a girl or both?
Did any of your siblings under the age of 18 leave the country during the past two years?
Was that a boy or a girl or both?
Did any of your siblings under the age of 18 marry during the past two years?
Was that a boy or a girl or both?
Did any of your siblings under the age of 18 go to work in a mine or a quarry during the past two years?

Was that a boy or a girl or both?

The answers will be dummy coded. The question about gender will be explored to see if the spillovers are gendered.

In addition we will use the household survey questions about what kids have been doing during the last two years as well as the household questions about gender equality.

3.4.2. *Additional variables to be used in the LASSO control set:*

As there is no limitation on how many variables that can meaningfully be included in the LASSO regressions we will include baseline variables and household variables for all variables described in section 3. In addition we will explore whether the variables on building material for the roof, floor, and walls help us in precision beyond the asset index.

4. Empirical strategy and hypotheses

Our empirical strategy is intended to capture individual treatment effects, spillover effects, and effects at the community level.

4.1 Individual level intention to treat effects of treatment

For the analysis of individual level treatment effects we restrict the sample to the treated communities and estimate the following regression:

$$Y_{2i} = \beta Treated_i + \chi X_{1i} + \varepsilon_i \qquad (1)$$

where 2 indicates follow up and 1 indicates baseline, i indexes individuals. The vector X will always include fixed effects for the strata variables gender and community. In addition we will see if we can improve precision in the estimates by including the controls listed in 3.2 and by picking optimal controls from the total list of controls (see 3.4.2) using LASSO (Belloni et al. 2014; Ahrens et al. 2018). If we have missing values on explanatory variables we will code the variables as zero and include dummy variables controlling for missing status so that we do not lose observations. We use robust standard errors in all estimations unless otherwise stated. Unless there is imbalance across

treatment and control (see below), the estimation with only the strata variables will be our main specification.

4.2 Spillovers

The specification in (1) will be a biased estimate of the overall effect in a community if there are spillover effects within villages. The question of spillovers is also interesting in itself. To estimate spillover effects we will restrict the sample to control individuals and estimate the following regression:

$$Y_{2i} = \beta Treated\_Community_i + \chi X_{1i} + \varepsilon_i \qquad (2)$$

where the treatment community is randomly assigned at the community level. As treatment at the community level is randomly assigned within regions we will always include *Region* as a strata variable in these regressions. We still estimate this regression using individual level data as power may be increased by including individual level controls. We do, however, cluster the standard errors at the community level in these regressions. As the sample includes only non-treatment individuals, β identifies within-community spillover effects by comparing control individuals in treatment villages to control individuals in pure control villages. Note that we have not randomly assigned the treatment *share* in different communities so we can not investigate how the spillovers change with different intensities of the treatment without further assumptions (Baird et al. 2018).

We will also investigate the questions described in section 3.4.1 to measure spillover effects.

4.3 Community level effects of schooling

We can also use specification (2) but without excluding any individuals to estimate the total effect of having a speed-school in your area. We can further add a dummy variable for being in the control group and interact Treated_community with being in the control group. This would lead to the following estimating regression:

$$Y_{2i} = \beta T\_Community_i + \eta C\_child_i + \phi C\_child_i * T\_Community_i + \chi X_{1i} + \varepsilon_i \qquad (3)$$

where T is short for Treatment and C is short for control.

The coefficient β now identifies the effect of having a speed-school in the area on the individual kids that were randomized to attend the school. This regression would include region fixed effects, female, and the standard errors would be clustered at the community level.

4.4 Balance and attrition

To test for balance we will regress treatment on the main independent variables described in section 3.2, both individually and together, while controlling for the Strata variables (female, community). We will judge whether the randomization worked by conducting an F-test of whether the control variables jointly predict treatment status.

We will also test for balance in the spillover sample by regressing *Treated_Community* on the control variables one by one and together, again controlling for the strata variables (Region fixed effects in this case). In the balance tests using randomization at the community level we will cluster the standard errors at the community level.

We will probably not manage to reach all the respondents initially sampled. We will check whether attrition and missing outcomes are correlated with treatment. If there are statistically significant differences in attrition or non-response between treatment and control (controlling for the strata variables), we will follow the correction proposed by Lee (2009).

4.5. Instrumental variables

To account for imperfect compliance we will also estimate an IV model of the following form:

$$Education_{2i} = \beta Treated_i + \chi X_{1i} + \varepsilon_i \quad (4)$$

$$Y_{2i} = \text{Predicted}\,(Education_{2i}) + \chi X_{1i} + \varepsilon_i \quad (5)$$

Where we predict education with the randomization and use the predicted values for education in the second stage to calculate the local average treatment effect of education on on our outcomes. Education can be measured as *Starting lower-secondary education* or by *Highest grade level*. We can also use the various learning measures in the first stage to estimate the effects of learning on the various outcomes. These analyses will not be treated as our main analyses but it may be useful to scale the ITT estimates to gain more insight into potential mechanisms.

5. Heterogeneity and further exploratory analyses

We will use machine learning techniques to automate the search for heterogenous treatment effects. There are many different types of machine learning algorithms and we have decided to use the "Generic ML" approach by Chernozhukov et al. (2018). If the omnibus test suggests that there is no heterogeneity we will not present further results from the method. As this field is moving rapidly, however, it is possible that there will be other techniques that are relevant for us once we start analyzing the data.

We will also explore heterogeneity directly with respect to some aspects and we will then estimate the following equation:

$$Y_{2i} = \delta Z_{1i} + \beta Treated_i + \lambda Treated_i * Z_{1i} + \chi X_{1i} + \varepsilon_i \qquad (6)$$

Z= Variables in 3.2. For some outcomes, such as the gender attitudes, it will also be interesting to explore heterogeneity across the dimension of household level gender inequality.

6. Power calculation, reporting, and discussion of null findings

Assuming a sample of 1300 individuals in the treatment areas and 80 percent power we can detect effects as small as 0.16 standard deviations.

We will also adjust the p-values for the fact that we are testing the impact on 6 outcomes. We follow the recommendations of Fink, McConnell, and Vollmer, (2014) and use a method developed by Benjamini and Hochberg, (1995) and Benjamini and Yekutieli, (2001) to minimize the false non-discovery rate. The main advantage of the method is that it is limiting the risk of false discoveries while only adjusting the critical values based on other true hypotheses. The false discovery rate method developed by Benjamini and Hochberg (1995) implies that the m p-values of the i hypotheses are ordered from low to high and that the critical value of the p-value is then p(i) = a*i/m. To illustrate, with 6 hypotheses and a significance level (a) of 0.05, the critical p-value would be 0.0083 for the one with the lowest p-value (0.05* 1/6, which is the same as a Bonferroni correction). For the second hypothesis, the critical p-value is 0.017 (0.05*2/6) and for the sixth it is 0.05 (0.05*6/6). As we have 6 main hypotheses, the most significant effect would have to have a p-value of less than 0.0083

and the minimum detectable effect becomes 0.19 with 1300 individuals (0.18 with 1500 individuals).

Note that we apply this correction to the main dependent variables only; when discussing individual variable results within particular outcome groups, we use conventional significance levels. We use this approach because the purpose of studying individual variables within the outcome groups is to understand mechanisms, rather than to single out particular variables for general conclusions.

We may not publish all of the findings in the same paper but we will create a working paper where we present all results outlined in this plan. This working paper will then be referred to and we will mention all main outcomes studied in each published paper.

Some of our findings are likely to be null results. It is often difficult to judge whether such results are showing a meaningful lack of effect or whether they arise due to low power. To investigate if the effects are meaningful null findings we will conduct equivalence tests with two one-sided t-tests (TOST) and show how large positive and negative effects we can reject. The tests are one sided in equivalence testing as one tests whether effects are larger than a highest value and lower than a lowest value. In practice, the procedure is equivalent to presenting the bounds of a 90 percent confidence interval.

7. IRB approval

This study was reviewed and approved by the Norwegian Centre for Research Data (number 386893).

8. Archive and sharing of replication data

The pre-analysis plan is archived before any follow-up data is collected. We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: https://www.socialscienceregistry.org/ on December 10, 2020. We will start the data collection later in December and it will be ongoing until April 2021.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. (2018). PDSLASSO: Stata module for or post-selection and post-regularization OLS or IV estimation and inference.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler. "Optimal design of experiments in the presence of interference." *Review of Economics and Statistics* 100, no. 5 (2018): 844-860.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608-650.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." In: *Annals of statistics*, pp. 1165–1188

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

Daigle, A. (2019). *Impact des représentations du vieillissement et du sentiment de contrôle sur l'ajustement à la retraite dans une population de nouveaux retraités* (Doctoral dissertation, Université du Québec en Outaouais).

Lee, David S. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76, no. 3 (2009): 1071-1102.

Ozer, Simon, and Preben Bertelsen. (2018) "Capturing violent radicalization: Developing and validating scales." *Scandinavian Journal of Psychology*.

Vallieres, E. F., & Vallerand, R. J. (1990). Traduction et validation canadienne-française de l'échelle de l'estime de soi de Rosenberg. *International journal of psychology*, *25*(2), 305-316.