For TechnoServe Ethiopia, Final Draft

# Measuring the effect of the Coffee Farmer College Program in Ethiopia

*Proposed evaluation design and analysis options*

Novermber 19th, 2014

www.laterite-africa.com

laterite TRANSFORMING RESEARCH IN AFRICA

Rwanda | Malawi | Burundi

# Framing of the report

## *Our understanding of the objectives and parameters of this assignment*

# Objectives of our assignment

## High level objective

To support TechnoServe to design a **robust and executable** impact evaluation strategy for its Coffee Farm College Program in Ethiopia (funded by Mondelez-IDH) aimed at measuring the impact of the program on the adoption of best coffee farming practices amongst "treated" farmers and giving TechnoServe the option to also measure the impact of the program on coffee productivity at a later date.

**The evaluation design should enable the following:**

- ❑ To estimate the effect of the program on Yields in Cohort 2015 / optionally 2016
- ❑ To estimate the effect of the program on Best Practice adoption in Cohort 2015 /2016

**The geographic area of focus:**

- ❑ The evaluation will focus on pre-selected Woredas in the Lekempti district only
- ❑ These Woredas include: Boji, Lalo Asabi, and Nejo

**Achieve a balance between operational objectives, budget and experimental validity:**

- ❑ The evaluation should be limited to about 11 Treatment Kebeles at maximum, in line with the 5000 farmer target (about 450 farmers "treated" per Kebele)
- ❑ Sample size should be limited to about +/-600 farmers, as has been the norm in the past
- ❑ The experiment should limit as much as possible interference with program operations

# Our understanding of TechnoServe's success criteria for the Mondelez program for Lekempti

## Outreach

- Train 5000 farmers in Cohort 2015
- Train 9000 farmers in Cohort 2016
- A trained farmer = a farmer that has attended more than 50% of sessions
- Farmers (even though they might be from the same household) are counted individually in the targets

## Yield

- To increase baseline yields by at least 50% in terms of kgs/ha
- This increase will only be visible four years after the baseline
- In the first few years, yields will decrease (due to rejuvenating), before they increase

## Best Practice Adoption

- 50% of trained households adopt at least half of best practices
- Best practice adoption will be measured as the sum of the adoption or not of 11 or 12 key best practices

# Our job: to propose pragmatic and well researched options, fallback options, and the tools to build a convincing story

Our proposed evaluation strategy will include 3 levels of analysis:

## Level 1: An experimental evaluation design

- This will be the core component of the evaluation
- The experiment will be based on randomization at the Kebele level
- The evaluation will aim to provide evidence of the impact (or not) of the program on yields and Best Practice Adoption

## Level 2: Quasi-experimental fall-back options

- Exploit program design to potentially introduce instrumental variables / regression discontinuity design
- Matching methods
- Looking at the potential for other creative solutions

## Level 3: The tools to build a good story

- Creating a link between Best Practice adoption levels and yields
- Using attendance data to create a link between attendance, yields, and BP
- Comparing BP adoption patterns to patterns of session attendance
- Using non-compliance to test for potential spill-over effects

# Our intention

1. **To provide TechnoServe with a very comprehensive study that can be replicated in Ethiopia and other country programs / other sectors**

- ❑ The proposed experimental strategy is flexible and is well suited for interventions where Treatment is assigned at the cluster (e.g. Kebele) level, rather than the individual level
- ❑ It is comparatively much less operationally disruptive than alternative methods

2. **To explain complicated concepts as best we can**

- ❑ Statistical analysis, inference, and related tools look and sound much more complicated than they really are
- ❑ We believe it is important for program managers to understand why certain things work and others don't from an evaluation perspective – we therefore take the time to explain things, step-by-step, breaking complicated looking methods into smaller manageable steps

3. **To base our recommendation on evidence and the latest available tools**

- ❑ Methods to deal with clustered data (data selected from a cluster, e.g. a Kebele) are just in the process of being developed and improved – it's a very active area of research
- ❑ The analysis options we propose are based on some of the latest methods available
- ❑ We test each of the tools using simulations based on models and real data provided by TNS

# Framing of the report

## *Summary of the main findings and proposals – a departure from previous TNS evaluation strategies*

# Summary of key recommendations (1)

1. **We propose that TNS implement a Pair-Matched Randomized Controlled Trial (PMRC) for this evaluation on no less than 10 Treatment and 10 Control Kebeles**

- ❑ Pair-Matched Randomized Controlled Trials are very valid randomization strategies, that enable efficiency gains, especially in the case of few clusters/Kebeles as is the case here (note that in this program Treatment is assigned at the Kebele level, not the individual level, so some form of cluster randomization is required).
- ❑ In this set-up, random assignment to the Treatment and Control group is done at the pair level: one Kebele is randomly assigned to the Treatment group, the other to the Control group.
- ❑ Individuals within each Kebele are randomly selected from a sampling frame (e.g. all coffee farmers in a Kebele or a village within a Kebele); random selection of farmers is conducted in exactly the same way in the Treatment and Control groups to avoid selection bias as much as possible.
- ❑ Pairing is done using the "greedy matching algorithm" – we provide TNS with a do-file to match Kebeles based on altitude, population and distance between Kebeles

2. **We propose that impact be measured at two levels, focusing in particular on estimating the Intention-to-Treat Effect (ITT) and the Complier Average Causal Effect (CACE)**

- ❑ The ITT is the effect of the program at the Kebele level that doesn't take into account whether a farmer actually took-up the training or not (it's the intention that counts)
- ❑ The CACE is an estimate of the treatment effect of the program on the treated

# Summary of key recommendations (2)

3. **We propose analysis tools that provide TNS with flexibility on: (i) the number of time periods required; and (ii) on whether to interview the same households at the baseline or endline**

❑ Inference is possible with just 1 time period (at the endline) and will be valid, albeit with lower power. This means that inference is still possible on the yield sample for Cohort 2015.

❑ Many tools rely on group-level estimates to derive inference about the impact of the program; in this case it is often not necessary to interview the same farmers at baseline and endline, an option that can be attractive from an operational perspective, even though this does in most cases come with lower statistical power.

❑ We leave it to TNS to decide on time periods and whether or not to interview the same people in each – the general rule is that the more time periods the better, and the best results are obtained when the same people are interviewed at baseline/endline.

4. **We recommend that TNS move away from using difference-in-difference techniques for coffee data - assumptions don't hold, potentially leading to very large estimation errors**

❑ We use the example of BP data from Cohort 2013 in Ethiopia to show that using difference-in-difference techniques can lead to an over-estimation of the impact of 178%!

❑ Difference-in-difference methods fail because the assumption of parallel trends does not hold: the baseline starting point is very strongly correlated to change at the endline

# Summary of key recommendations (3)

5. **Given methodological uncertainties and low power in the case of few clusters (i.e. Kebeles), we propose that TNS use several inference strategies to test the validity of results. Recommended inference techniques are grouped into three categories: (i) randomization inference; (ii) difference with lagged variables, using the wild cluster bootstrap; and (iii) synthetic control methods combined with randomization inference**

- ❑ We propose two types of randomization inference techniques, based on Fisher and Neyman inference. These methods are specifically adapted to the pair-matched cluster randomization scenario and base inference squarely on the randomization process.
- ❑ OLS estimation of baseline/endline differences, with a lagged dependent variable assuming unconfoundedness, using the wild cluster bootstrapping method to correct for clustered standard errors (this methods provides for accurate statistics under clustering).
- ❑ Synthetic Control methods, which enable us to improve the quality of the counter-factual, are combined with randomization inference to enable valid inference

6. **We provide do-files and detailed step-by-step explanations to enable TNS to implement these techniques in-house, along with monte-carlo simulations to estimate power**

- ❑ We estimate power in three steps: (i) model the behavior of the yield and best practice samples at baseline and endline assuming a certain level of impact; (ii) generate random hypothetical BP and yield samples; and (iii) apply the inference techniques to test whether we can detect an impact or not

# Summary of key recommendations (4)

7. **Low statistical power (partly due to the risk of high non-compliance, but in particular due to the few Kebele problem), is the biggest issue affecting this evaluation, so we put forward a number of ideas that might help increase power. These include:**

- ❑ A survey to collect individual, Kebele and Woreda level data at the baseline - controlling on covariates will increase the statistical power of any regression.
- ❑ We propose that the random selection of farmers be conducted at the village-level instead of the Kebele level, focusing on villages within treatment Kebeles where we know compliance will be high.
- ❑ We suggest that only households that: a) confirm they are coffee farmers; and b) say they would be likely to participate in a coffee training program if given the opportunity, be selected into the sample.
- ❑ We also recommend changes to the way yield and BP data is collected.

8. **We provide some ideas for quasi-experimental options, should the actual experiment fail or that could be used in addition to the core analysis.**

- ❑ One idea, is to exploit the fact that a Farmer Trainer can take on a limited number of groups/trainees. This means that in each Kebele there is an optimal coffee farmer population level. Above that level, inefficiencies creep in – e.g. group sizes are too large, farmers are not given the opportunity to participate. Distance from the optimal population level, could be a good instrument.
- ❑ We also provide some other ideas that could be tested.

# Summary of key recommendations (4)

9. **Finally, for ethical reasons, we recommend that if TNS receives additional funding to expand the intervention in Lekempti that the objective to measure the Treatment effect on the yield sample be dropped**

- ❑ By implementing a pair-matched randomized controlled trial, we are effectively preventing thousands of farmers from receiving the treatment over the next few years (approximately as many farmers are excluded, as farmers that receive the treatment)
- ❑ Given the current budget, TNS can only provide training to about 14,000 farmers in Lekempti over the next four years out of a total of about 150,000 potential coffee farmers across target Woredas in Lekempti, including Nejo, Lalo Asabi, Boji Dermaji, Haru, and Gulisso. The random assignment therefore doesn't prevent additional farmers from receiving the treatment.
- ❑ This equation would change however if TNS gets additional funding to expand the program in Lekempti.
- ❑ If this were to be the case, we would recommend that TNS drop the objective of measuring the effect of the program on the yield sample, given that: a) effects on the yield sample can only be measured after a period of four years (after coffee trees have rejuvenated), which is a very long period of time; b) statistical power in the yield sample is likely to be low, given the expected structure of the data; and c) the effect of the Treatment on BP adoption can be measured after the end of the program, thereby giving a strong initial indication of whether the program has achieved its targets or not.
- ❑ Endline data on the yield sample can provide evidence that yields have increased in Treatment Kebeles and instrumental methods could be used to prove causality

# Overview of proposed approach for each/sample and Cohort

**BP Cohort 2015**

Pair-matched cluster randomized trial, with baseline in early 2015 and endline in 2019, with ideally a midline in 2017 and other time periods if possible. We suggest using at least 10 Treatment and 10 Control Kebeles (bare minimum), more if possible. We would recommend sampling an equal number of farmers per Kebele, with the option of correcting for population levels using weights.  We would recommend targeting 30 farmers per Kebele in both the Treatment and Control groups.

**Yield Cohort 2015**

There is still a possibility of measuring impact for the yield Cohort in 2015. This would involve the same pair-matched cluster randomized trial as for the BP sample, with the analysis only focused on endline data collected in late 2018, controlling for baseline covariates and BP data. We would also recommend a minimum sample size of 30 farmers per Kebele. The risk is very low statistical power, but it is difficult to predict at this point.

**Yield and BP, Cohort 2016**

For Cohort 2016, we recommend implementing a very similar strategy, involving matched-pair cluster randomization. Target areas for Cohort 2016 have yet to be defined. Given the larger population target, we recommend that TNS implement this strategy with a greater number of Treatment and Control Kebeles, still targeting about 30 sampled farmers per Kebele. This would increase statistical power significantly.

# Will the methods we propose work?

… probably yes for the Best Practice sample
… possibly not for the Yield sample

**Despite a valid experimental design, statistical power to detect an effect in the yield sample risks being too low:**

- ❑ Even with high compliance levels, the power to detect an effect in the yield sample will be low, unless the effect is about a 70-80% increase in yields (which is not unfeasible, if the Rwanda case is to be believed).
- ❑ This is essentially because of the few clusters & intra-cluster correlation problem … many more clusters/Kebeles would be required to increase power.
- ❑ We propose a number ideas and fixes that might help increase power substantially, but we cannot predict whether these will yield sufficient statistical power or not.
- ❑ Should statistical power be too low, the objective of the evaluation will therefore not be to obtain statistical significance, but rather to build a convincing story about impact. There are many elements that can contribute to this story and we discuss them in this report.

# Why are the proposed changes a significant improvement on previous TNS evaluation methods?

| Main issues | This proposal | Current TNS methods |
|---|---|---|
| **Selection of Kebeles** | PMCR significantly reduces risk of selection bias at Kebele level – Kebeles are first matched, then randomly assigned to Treatment/Control | No randomization in allocation of Kebeles to Treatment/Control, hence selection bias. Moreover, control Kebeles sometimes selected because they were not suited for treatment, hence over-estimation of impact |
| **Selection of farmers** | Random selection of farmers at Kebele level reduces individual selection bias, and enables estimation of ITT and CACE | Comparison of compliers (or registered farmers) in the Treatment, to random group of control farmers, inevitably leads to selection bias (due to self-selection) and hence an over-estimation of impact |
| **Estimators** | Estimators specifically designed for PMCR or few clusters scenario, leading to un-biased inference. | Difference-in-difference estimation used, but assumptions don't hold, leading to potentially large over or under-estimations of impact |
| **Standard errors** | The selected analysis techniques were specifically designed to deal with inference in the case of few clusters, where asymptotic properties do not apply | Adjustments to clustered standard errors in the case of few clusters not made, leading to potentially very large over-estimation of the significance of results. |
| **Data collection** | Average yields calculated per plot, or only on one plot, rather than across plots. BP data collected from the exact same plot as the yield data. | Yield is calculated as the total production, divided by aggregate plot size across plots – leads to unrealistically large yield estimates (+/- 20% of sample dropped). BP data is not collected from the same plot as yield data. |

# Framing of the report

*List of deliverables and the structure of this report*

# Deliverables and structure of this power-point …

## Power-point

1. Section on understanding intra-cluster correlation and power calculations
2. Section on PMRC and the randomized allocation of Kebeles to Treatment and Control
3. Section on all the inference tools and their power
4. Section on how to improve our changes of success

## Do-files

1. 6 do-files to for inference tests and Monte-Carlo simulations for power
2. 2 do-files for balancing groups (greedy matching algorithm)
3. 6 do-files for modelling the yield and BP data
4. 1 do-file to study attendance patterns
5. 1 do-file for calculating Kebele distances
6. 2 do-files for basic power calculations

## Other deliverables

- Matching of Kebeles into pairs and random allocation of Kebeles to Treatment/Control
- Two Addis trips, in late August and early October
- Debriefs with John, Nupur, and Caroline (in Kigali)
- Support on running/adjusting do-files

# This power-point is structured along the following objectives …

*Slide 19*

**Explain the challenge of dealing with clustered data (section called “Understanding the issues”)**

- Understanding cluster-randomized controls
- Understanding intra-cluster correlation
- Understanding the parameters that contribute to power calculations
- Understanding power dynamics based on changes in underlying parameters

*Slide 42*

**Make the case for Pair Matched Cluster Randomized Trials (section called “Proposed Approach: PMCRT”)**

- Making the case for pair matched cluster randomized controlled trials
- Explaining the greedy algorithm for matching Kebeles into pairs
- The random allocation of Kebeles to treatment and control

*Slide 50*

**Explain inference options and guidelines for the analysis phase (section called “Inference options and guidelines”)**

- Modelling the behavior of the yield and BP datasets and associated insights
- Explaining 8 different inference strategies
- Estimating the power of these strategies using Monte-Carlo Simulations
- Explaining alternative quasi-experimental options and how to build a good story

*Slide 172*

**Suggest ways to increase power (section called “Increasing chances of success”)**

- Suggest ways to increase statistical power, including baseline survey and screening
- Suggestions to improve the way data is collected for the yield and BP samples
- Some ideas on questions for the baseline survey and for (optional) screening questionnaire

# Understanding the issues

## *Cluster Randomized Controlled Trials and intra-cluster correlation*

# The only feasible path to a robust experimental design in this case is some form of a cluster-randomized controlled trial (CRT)

**What is a cluster-randomized trial (CRT)?**

- It's a randomized control trial (RCT), except that randomization is not done at the individual level but rather at the cluster level (here Kebeles)
- In this case it would mean that Kebeles in selected Woredas will be randomly assigned to the Treatment Group and the Control Group, following a specific randomization process (e.g. matched-pair design)

CRTs are an excellent way of conducting a valid evaluation, except that:

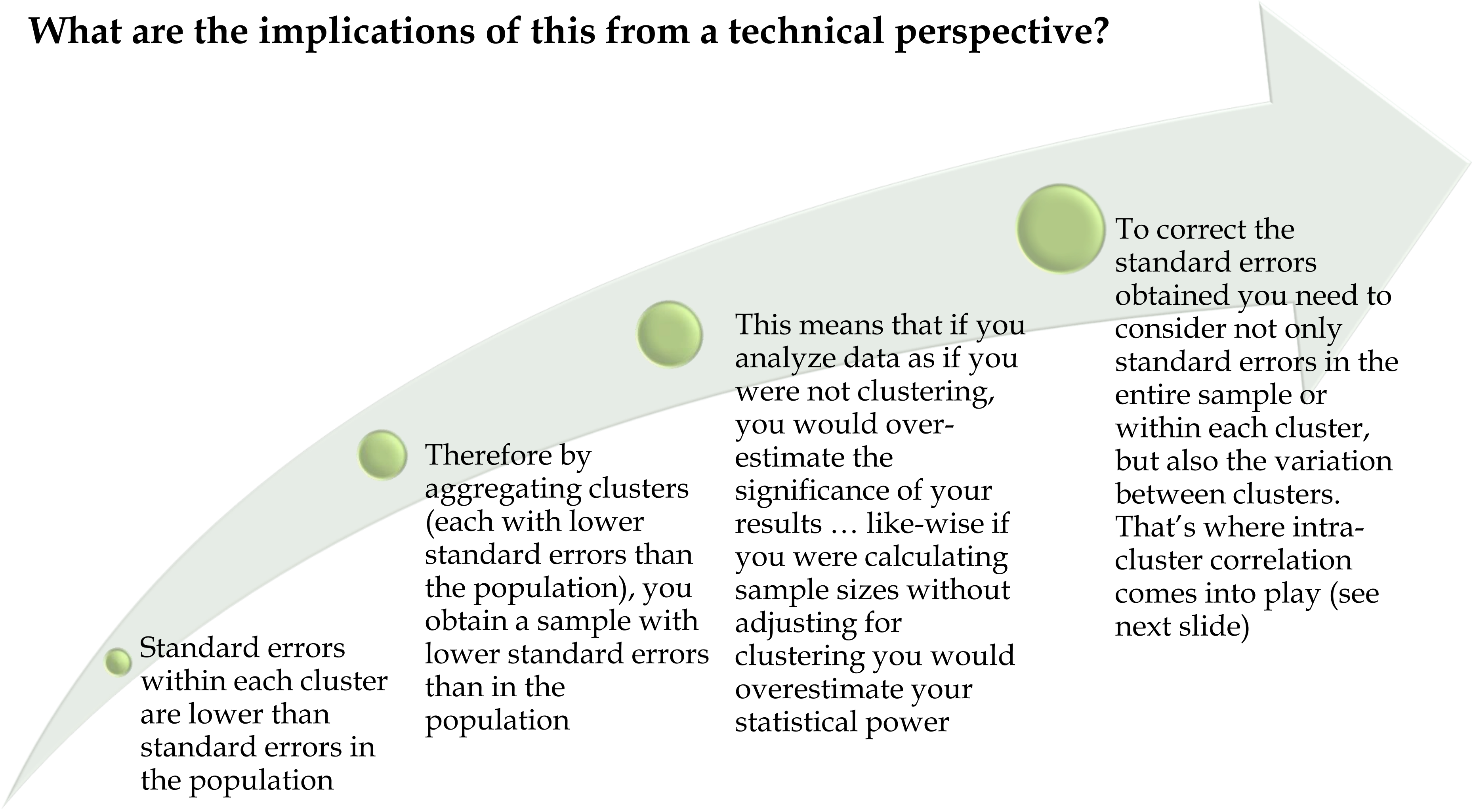
1. It might require many Kebeles to work because individuals within a cluster tend to be much more similar than individuals in different Kebeles, which violates the principle of independence of errors … you need to adjust for this similarity by taking into account what's called the intra-cluster correlation (see next slides)

2. There is a very high risk of imbalances between the treatment group and the control group due to cluster related factors (the fewer the clusters the higher the risk)

3. The analysis is significantly more complicated, especially in the case of few clusters (this case), where additional adjustments are required!

# CRTs are less efficient / more complicated than RCTs because people within clusters tend to be more similar than people across clusters

**What are the implications of this from a technical perspective?**



Standard errors within each cluster are lower than standard errors in the population

Therefore by aggregating clusters (each with lower standard errors than the population), you obtain a sample with lower standard errors than in the population

This means that if you analyze data as if you were not clustering, you would over-estimate the significance of your results … like-wise if you were calculating sample sizes without adjusting for clustering you would overestimate your statistical power

To correct the standard errors obtained you need to consider not only standard errors in the entire sample or within each cluster, but also the variation between clusters. That’s where intra-cluster correlation comes into play (see next slide)

# Some intuition about intra-cluster correlations (ICC) or what is called the Rho … here we replace the word "clusters" by "Kebeles"

**How is it measured and what does it mean?**

- It is simply the share of variance on a variable that is due to differences between Kebeles (another source of variance is differences amongst farmers within Kebeles)
- So if the Rho is very high, then most of the information (i.e. variation) comes from differences between Kebeles, rather than within Kebeles (and vice-versa)
- Intuitively, the Rho is a measure of how much information you stand to lose by not going to enough Kebeles in your design

If Rho = 0, then ….

- Individuals within a Kebele are no more similar that individuals in different Kebeles
- i.e. You have sufficient information in one Kebele to do your entire experiment

If Rho = 1, then ….

- All the individuals in a Kebele are exactly the same: i.e. interviewing one person in that Kebele is the same as interviewing 1000
- i.e. Your sample is only as good as the number of Kebeles

# Do we know what the Rho is likely to be for this experiment (example based on the yield sample only)?

**Estimating the Intra-cluster-correlation (ICC) or Rho**

- Usually, researchers do not have the luxury of knowing what the Rho will before they actually carry out a survey; hence they look at past or similar surveys to estimate what the Rho is likely to be
- In this case however, we have baseline data on yields from Jimma and Illubabor that can be used to estimate the Rho
- The good news, is that estimates of the Rho from Jimma and Illubabor are similar and consistent

**Estimates of Rho using data from Illubabor and Jimma for the yield sample**

| Variable used | Estimated Rho based on Illubabor data | Estimated Rho based on Jimma data |
|---|---|---|
| Yield (coffee per ha) | 0.218 | 0.196 |
| Total production | 0.299 | 0.263 |

# Interpreting these figures using design effects … the CRT design requires a much larger sample than if we were doing an RCT

## What are design effects?

- "Design effects" is the factor by which you need to increase your sample size, to compensate for the information lost given the Rho / the CRT design of the experiment
- The intuition: sample sizes are calculated in two steps – first you calculate the required sample size as if randomization had been done at the individual level (RCT); then you multiply that sample size by the "design effects" factor to adjust for the loss of information or statistical power that comes from the CRT design

CRT requires a sample at least 5.2 times larger than if we were doing an RCT

## "Design effects" corresponding to Rho (realistic example)

**Sampling assumptions**

- Desired power: 80%
- Mean (yield): 1200
- Target (yield): 1800
- Standard deviation: 1050
- Baseline/endline correlation: 50%
- Compliance rate=60%

| Rho | Design effects |
|---|---|
| 0.196 | 5.2 times larger sample required |
| 0.218 | 5.5 times larger sample required |
| 0.263 | 6.0 times larger sample required |
| 0.299 | 6.4 times larger sample required |

# It also means this CRT experiment will not be feasible below a certain number of Kebeles

**Rule of thumb to calculate the minimum number of Kebeles required**
The minimum number of Kebeles required to make a CRT possible can be calculated by simply multiplying the sample size required under an RCT (n) by the Rho

**Realistic example of what this means for the yield sample (same assumptions as in previous slide and assuming Rho=0.25)**

| | |
|---|---|
| If Target Power=80% | At least 25 Kebeles required in both Treatment and Control to make this CRT feasible (50 in total) |
| If Target Power=90% | At least 34 Kebeles required in both Treatment and Control to make this CRT feasible (68 in total) |

⚠ Note: this is just to give the reader a sense of the numbers we are talking about

# Understanding the issues

*The parameters of power calculations and parameter estimates for this evaluation*

# Calculating the expected compliance rate (1)

**What is the compliance rate and why is it important?**

- We define the compliance rate here as the share of treatment households in which at least one member has attended more than 50% of training sessions
- This definition assumes that no control group household will take-up the treatment
- The compliance rate is important because we will be randomly sampling at the Kebele level, regardless if a farmer/household ends-up taking the treatment or not. This means that there is a high risk that some members of the treatment group will not end up being treated. Sampling power reduces very rapidly when the non-compliance rate increases.

**Compliance rate = Registration Rate x Share of registered households that are "treated"**

- Estimating the registration rate: registration is driven by two dynamics, (i) the propensity of a farmer to register if given the opportunity; and (ii) the number of Farmer Trainers assigned to a Kebele, which puts a (fuzzy) upper limit on the total number of farmers that can be trained.

- Estimating the share of registered households that will be treated: here we can use data on attendance rates from previous TNS Cohorts in Ethiopia

# Calculating the expected compliance rate (2)

**Estimating the registration rate**

- The average registration rate of 9 treated Kebeles in the Lalo Asabi and Gulliso Woredas in Lekempti was 82% of coffee-farming households
- For design purposes we estimate that 90% of farmers who are given the opportunity will register, and that the opportunity to participate is provided to about 90% of farmers in each Kebele. Combined that equals 81%.
- The opportunity to register is determined by the number of coffee farmers in that Kebele vs the number of  Farmer Trainers (FTs) allocated to that Kebele (the average FT provides training to about 300-340 households)
- There will always be an efficiency loss (i.e. a mismatch between optimal population and number of FTs); the risk of such an efficiency loss increases the more Kebeles you go to given budget constraints

**Estimated average registration rate: 80%**
(note this is close to the estimate from Lalo Asabi and Gulliso)

# Calculating the expected compliance rate (3)

## Estimating the share of registered households in which at least one member has attended 50% of sessions

- To estimate this we use attendance data from Cohort 2013 interventions
- We adjust estimate by the "monitoring effect" (see Kigali work Laterite did) … farmers in the yield sample were much more likely to attend sessions

| Area | Share of registered households that are "treated" |
|---|---|
| Nespresso | 73.2% |
| Nestle | 74.8% |
| South | 69.9% |
| West Illubabor | 73.7% |
| West Wellega | 91.7% |
| **Average (excl. West Wellega)** | **73%** |

**Estimated % registered households that are treated: 73%**

**Estimated compliance rate: 80% x 73% = 58.4%**

# Estimating the Intra-Cluster Correlation (Rho)

**Intra-cluster correlation in the yield sample**

- We estimate the intra-cluster correlation in the yield sample using data from Jimma and Illubabor
- The estimated Rho in both samples is consistent at about 0.2
- Estimates are further refined using modelling (see section on inference techniques). These models yield an average Rho of about 0.17-0.18, but with high standard deviation. I.e. the Rho could lie anywhere between 0.1 and 0.3.

**Estimated Rho for Yields = 0.18**

**Intra-cluster correlation in the best practice sample**

- We estimate the Rho using baseline BP data from Cohort 2013. Estimated at 0.17
- Endline BP from Illubabor suggests much lower Rho
- Our models (see modelling section, under inference techniques), suggest Rho of 0.224. This Rho however has a large standard deviation.

**Estimated Rho for BP sample = 0.22**

# Estimating means and standard deviations in the yield sample

**Mean yields at baseline**

- We base yield estimates using data from the Jimma and Illubabor baseline
- In Jimma and Illubabor yield estimates were between 1160 and 1380 kg/ha on average
- We select the lower bound as our working assumption

**Estimated mean yield = 1160 kgs/ha**

**Standard deviation at baseline and endline**

- We estimate standard deviations at the baseline on data from Jimma and Illubabor
- Standard errors in both regions were very similar: between 1026-1061kg/ha

**Estimated standard deviation in treatment and control = 1044**

# Estimating means and standard deviations in the BP sample

**Mean Best Practice  at baseline**

- Our indicator of interest: Total number of best practices adopted
- We measure the mean value of this indicator using baseline data from Cohort 2013
- The mean on this variable of interest is about 3.2

**Estimated mean Total number of BPs adopted = 3.2**

**Standard deviation at baseline and endline**

- We estimate BP standard deviations at the baseline using data from Cohort 2013
- The standard error on this variable of interest is about 1.66

**Estimated standard deviation at baseline = 1.7**

# Estimating the coefficient of variation of cluster sizes

**What is the coefficient of variation of cluster sizes**

- If we want a sample that is fully representative of the population, then we have to take cluster sizes into account
- Having clusters of different sizes adds to the power loss due to clustering
- The coefficient of variation in this case is simply the standard error of cluster sizes, divided by the mean cluster size

**Estimating coefficient of variation of cluster sizes**

- We estimate the coefficient of variation of cluster sizes using data from the selected sample

**Estimated coefficient of variation of cluster sizes = 0.31**

# Sampling parameters that are not known and for which we do not have good estimates - baseline/endline correlation and attrition

**Estimating baseline endline correlation in yield and BP sample**

- We do not have a good estimate of the potential correlation between baseline and endline values either in the yield or BP sample (i.e. this is a high risk estimate)
- Data from Rwanda suggests it could be as low as 0.2 in the case of yields, but this was only for farmers in the treatment sample.
- Our models (see modelling section), suggest a baseline/endline correlation of about 0.31 in the yield sample, and -0,17 in the BP sample
- Low levels of correlation have a very negative impact on statistical power.

**Estimate of baseline/endline correlation = 0.31 (Y) and -0.17 (BP)**

**Attrition**

- Attrition will depend on: (i) whether we can trace farmers back at the endline (more of an issue in the control group); (ii) missing observations due to bad data quality
- We augment required sample sizes by 12.5% to account for potential attrition between baseline and endline (assumed 15% attrition in the Control group, and 10% in Treatment)

**Attrition rate = 12.5%**

# The minimum detectable difference target in the Best Practice and Yield samples

**What is the minimum detectable difference**

- It's the minimum difference (improvement in the outcome indicators) that we want to be able to detect with this experiment
- This can either be an input into the sampling calculations, or an outcome thereof
- You can sample with a given minimum difference target, sample size, or power target

Minimum detectable difference in yield sample (delta) = 50% increase

- TNS expects an impact of 50% on yield levels; so the maximum at which we can set this delta is 50%
- Sample size calculations suggest it will be difficult to get the delta well below the 50% mark, given the limited number of Kebeles we can include in the evaluation

Minimum detectable difference in the BP sample (delta) =2.8 BPs

- TNS targets that 50% of "trained" farmers adopt at least 50% of Best Practices – i.e. minimum of 6/11 BPs
- Given that the mean and median on the Total Number of BPs adopted variable are very close, the minimum detectable difference we require is the difference between 6 and the baseline level of BP adoption

# Understanding the issues

*Statistical power dynamics based on changes in some of these parameters (assuming asymptotic properties hold)*

# What are the main drivers of low statistical power in this evaluation? We provide examples using the yield sample

The main reason statistical power in this evaluation risks being low, in particular in the yield sample, is because of dynamics related to the following sampling parameters:

1. **The number of Kebeles (which depends on operational objectives and budget)**

- ❑ For Cohort 2015 we expect about 11 Treatment Kebeles, given operational objectives.
- ❑ The marginal power gains from adding one additional Treatment Kebele are much larger, than increasing actual sample size. The more Kebeles included in this evaluation, the better.

2. **Potentially low levels of compliance (this is something we can influence, but not control)**

- ❑ The higher the compliance, the higher the power to detect an effect. Power is very sensitive to changes in compliance levels.

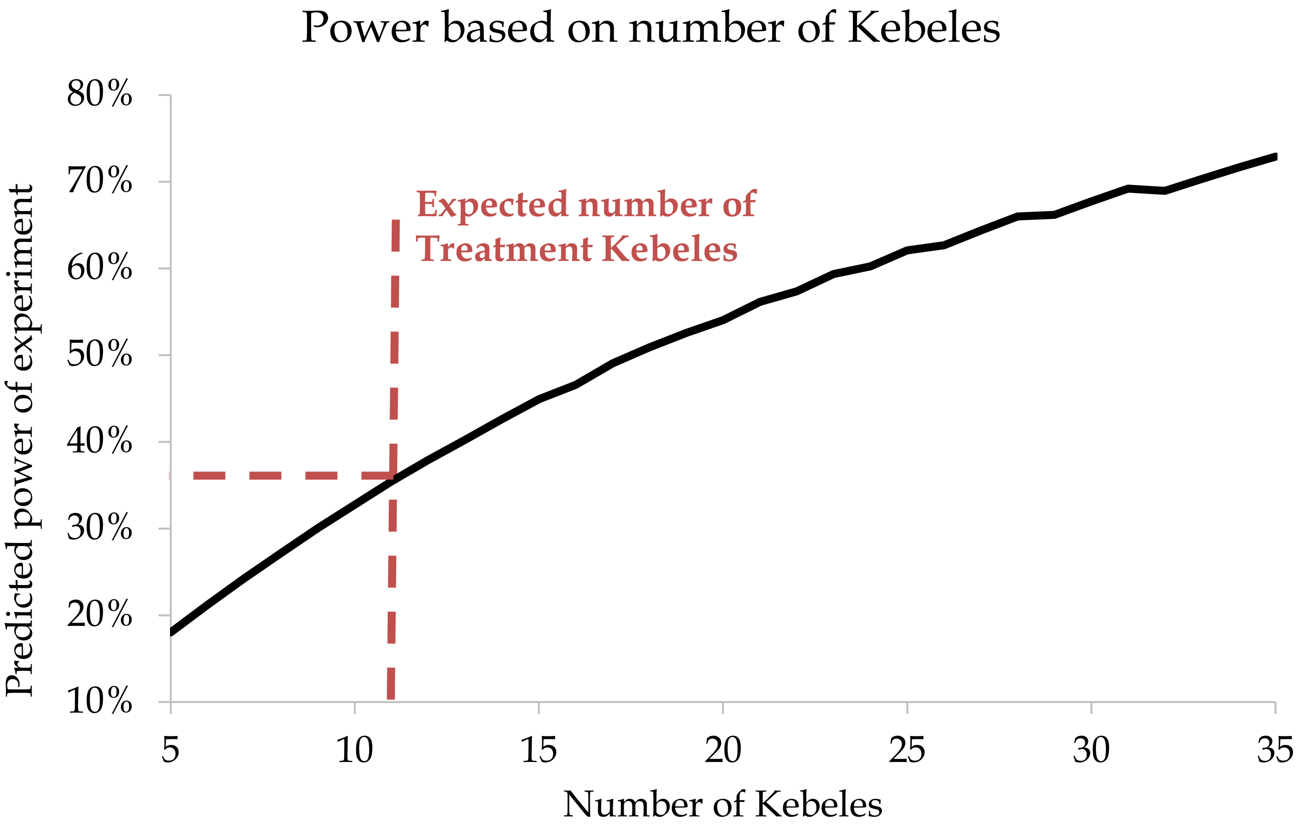
3. **Potentially high intra-cluster correlation (there is nothing we can do about this)**

- ❑ High intra-cluster correlation, has a very large effect on statistical power. In this evaluation the level of statistical power is unpredictable. The average predicted Rho in our simulations is about 0.18, but it could lie anywhere between 0.1 and 0.3.

# The number of Kebeles plays a key role: statistical power in the yield sample will be very low unless we have >35 treated Kebeles!

**Parameters for basic power calculations**

- Sample size: 600
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Rho=0.18
- Cov=0.31
- Coefficient of variation=0.31

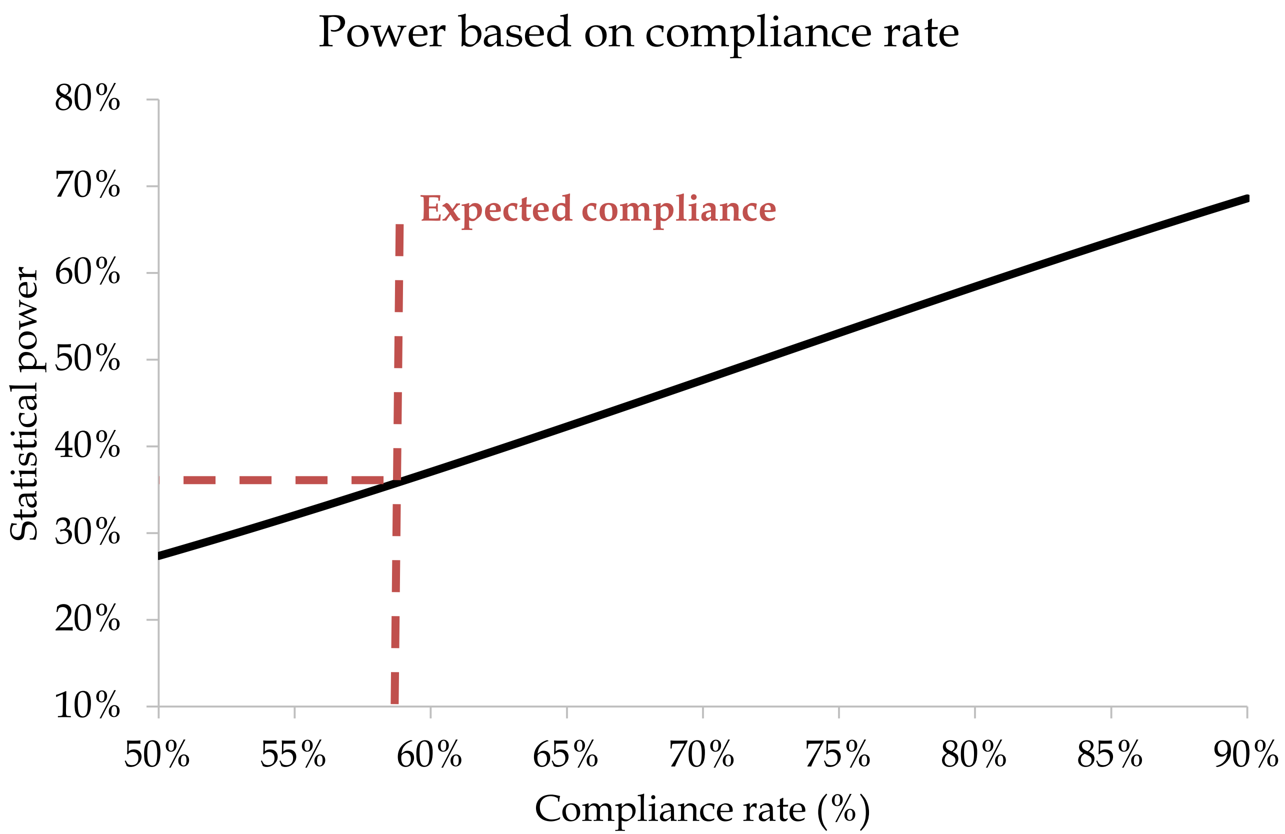


- With 11 Treatment Kebeles and based on these parameters we would estimate power to detect an effect in the yield sample to be about 35%, well below the 80% bar

# The compliance rate also plays a key role: the lower the compliance rate, the lower our ability to detect an effect

**Parameters for basic power calculations**

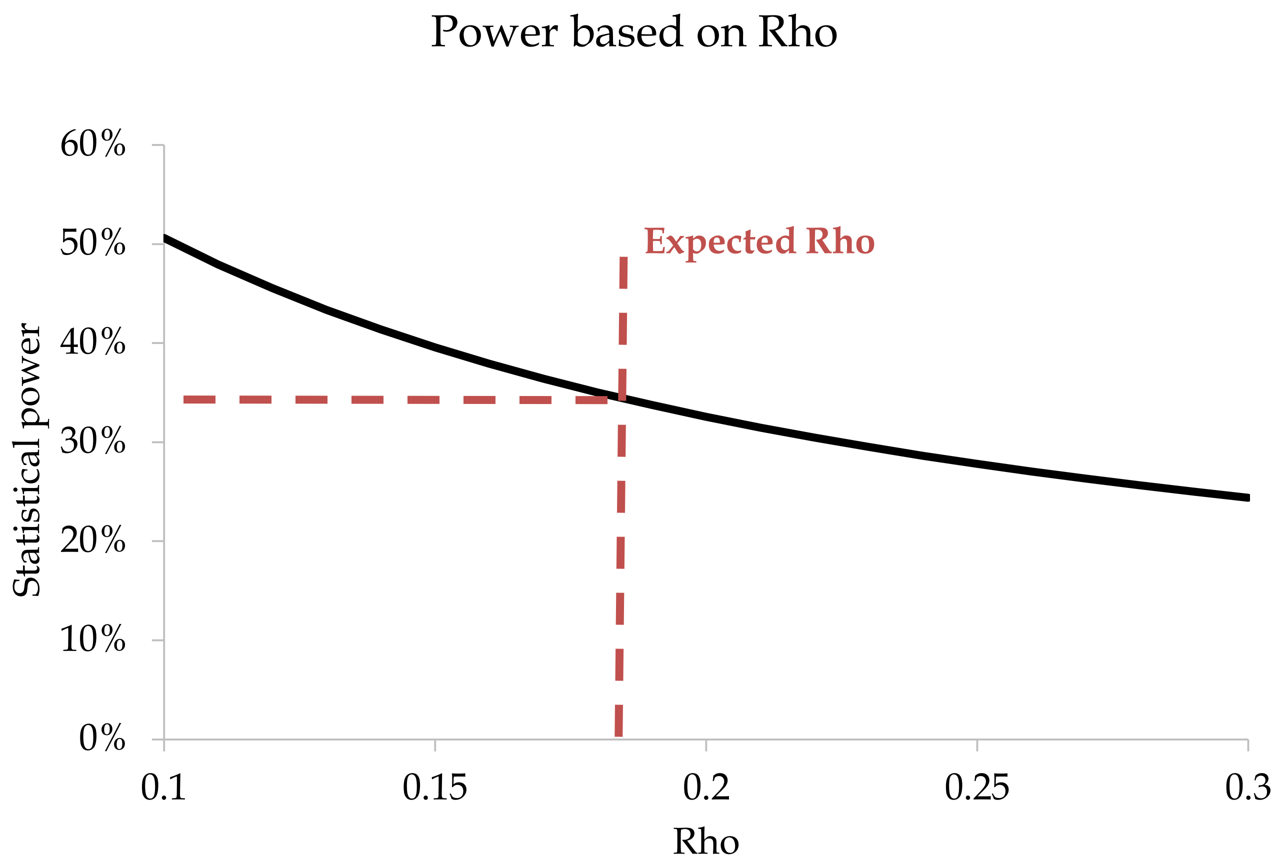
- Sample size: 600
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Rho=0.18
- Cov=0.31
- Coefficient of variation=0.31



# The intra-cluster correlation coefficient, which is outside our control, also plays an important role in determining power

**Parameters for basic power calculations**

- Sample size: 600
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=58%
- Attrition: 10% (T) and 15%(C)
- Rho=vaires
- Cov=0.31
- Coefficient of variation=0.31



# So what can we do in the case of few Kebeles, with potentially low compliance and hence low statistical power?

**The following three sections provide a detailed answer**

# The proposed experiment

## *Pair-Matched Cluster Randomized Controlled Trial (PMCRT)*

# What is a Pair-Matched Cluster Randomized Trial and why is it the best solution for TechnoServe's Ethiopia program?

## Pair-Matched Cluster Randomized Controlled Trial

It's a cluster-randomized controlled trial, except that randomization happens within pairs of similar Kebeles, rather than across all Kebeles. Pairs are formed using an algorithm that matches Kebeles based on pre-selected baseline indicators (i.e. Kebeles within pairs are as similar as possible). One Kebele in each pair is randomly assigned to the treatment group, the other to the control group. The advantage: (i) it's proven to lead to efficiency/power gains; (ii) it is operationally less burdensome to roll out.

The arguments for pair-matched cluster randomized controlled trials:

- ☐ **It does a good job at balancing covariates between treatment and control**
- ☐ **It is one of the most reliable ways to achieve greater experimental power and efficiency ex-ante,** given few clusters and high intra-cluster correlation.
- ☐ **It is still very much a valid randomization process**. There are $2^P$ different possible randomization outcomes, where P is the number of pairs. For example, with 11 pairs, there are exactly 2048 different possible assignments of Kebles to treatment and control. Only one of these possibilities is ultimately selected.
- ☐ **It provides greater operational flexibility than alternatives, while maintaining experimental validity**. For example: pairs can be dropped from the evaluation without affecting the validity of the experimental design (if for example TNS wants to treat both/neither Kebele, if implementation stops/fails in one Kebele, etc)

# How is the matching done? We propose that TNS use what's called the greedy matching algorithm …

## Matching using the greedy algorithm

Pair-matched randomization is the extreme of block randomization – rather than randomizing within strata or blocks of Kebeles, you randomize at the smallest possible unit, the pair. Kebeles are matched into pairs using an algorithm. One of the most popular is called the greedy algorithm, which minimizes what's called the Mahalanobis distance between Kebeles on certain variables. It's a way to ex-ante control for variables that you know will impact the variables of interest (yield and BP adoption). "Controlling" ex-ante for known predictors will increase power, because part of the variation between clusters is explained by these variables. We provide a do-file to run this.

**Proposed variables on which we can conduct the paired-matching:**

- ❑ ***Number of coffee farmers:*** this will play a very important role in the determining the outcome of the project, because: a) the number of Farmer Trainers allocated per Kebele is linked to it; and b) because it will determine registration rates (which will be influenced by difference between optimal training population / full capacity and the actual population of the Kebele)
- ❑ ***Altitude:*** is a very strong predictor of coffee yields
- ❑ ***Distance***: Kebeles that are geographically closer to eachother are likely to be more similar

# We run the greedy algorithm (see do-file), that creates pairs of Kebeles and ranks pairs based on the quality of the match in each Woreda

**Example using Kebeles in the Lalo Asabi Woreda**

| Pair Rank | Kebele 1 | Kebele 2 | Coffee households (Kebele 1) | Coffee households (Kebele 2) | Altitude (Kebele 1) | Altitude (Kebele 2) | Distance between Kebeles (km) |
|---|---|---|---|---|---|---|---|
| 1 | Horda Daleti | Betrochekosa | 610 | 549 | 1666 | 1672 | 3.5 |
| 2 | Hatis Siben | Werebabosiben | 306 | 288 | 1705 | 1633 | 3.9 |
| 3 | Haroji Horowa | Garjo Siban | 272 | 290 | 1774 | 1831 | 7.3 |
| 4 | Jarso Damota | Tosiye Mole | 805 | 587 | 1603 | 1685 | 6.4 |
| 5 | Halehuwa | Amuru Gara | 364 | 480 | 1639 | 1591 | 8.2 |
| 6 | Dinihuwa | Werego Arsema Ab. | 464 | 307 | 1606 | 1743 | 11.5 |
| 7 | Dengoro Disi | Dengoro Gebo | 271 | 282 | 1828 | 1724 | 3.7 |
| 8 | Kele Birbir | Wera jiru Becho | 724 | 506 | 1568 | 1845 | 14.8 |
| 9 | Gombohuwa | Haroji Agemsa | 306 | 350 | 1661 | 1860 | 5.7 |

**See do-file called:**
**matching_algorithm (implements the greedy algorithm)**

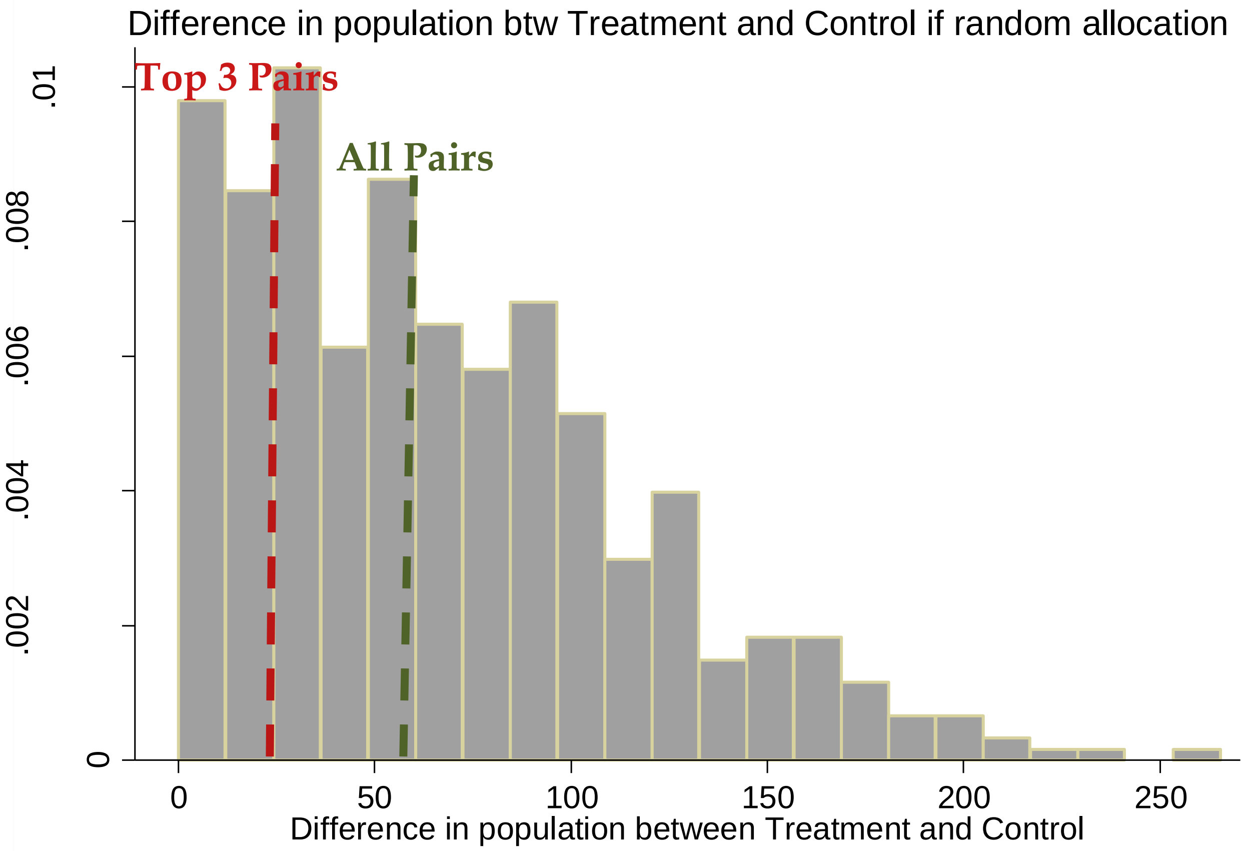
# The quality of the match is best for the top pairs … we recommend that the top pairs of Kebeles be selected into the sample in each Woreda

**Example using Kebeles in the Lalo Asabi Woreda**

| | Indicators | Treatment | Control | Difference |
|---|---|---|---|---|
| **Top 3 Pairs** | **Coffee households** | 396 | 376 | 20 |
| | **Altitude** | 1715m | 1712m | 3m |
| | **Distance between Kebeles within pairs** | 4.9km | | n/a |
| **Bottom 6 Pairs** | **Coffee households** | 489 | 419 | 70 |
| | **Altitude** | 1651m | 1741m | -91m |
| | **Distance between Kebeles within pairs** | 8kms | | n/a |

# For population levels, the top pairs lead to a significant improvement in balance compared to a fully random allocation of Kebeles

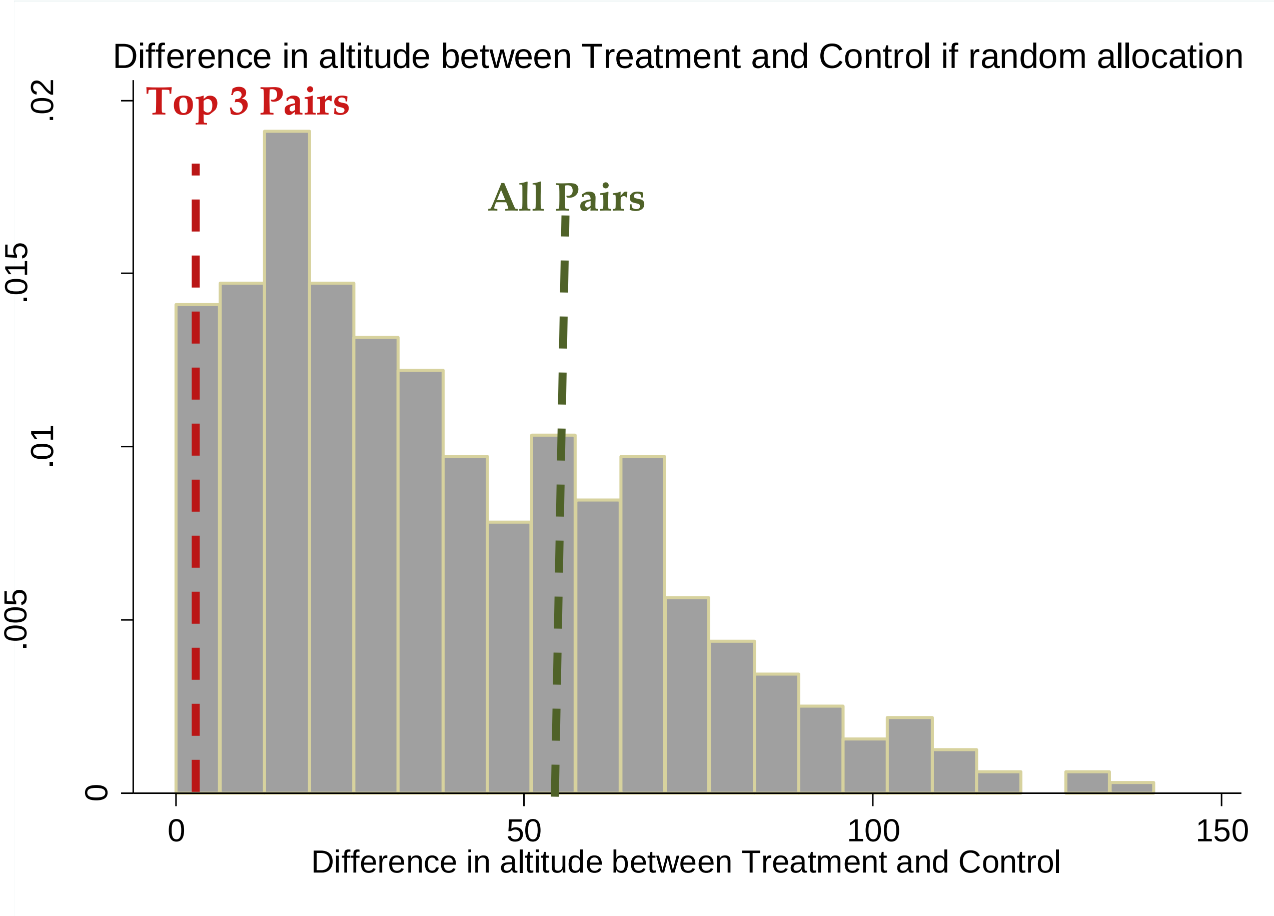
**Example using Kebeles in the Lalo Asabi Woreda – 1000 possible random allocations of Kebeles to Treatment and Control**



**Stats**
Difference top 3 pairs = 20
Difference all pairs = 54
Difference random mean=67

# This is also true for altitude, although on average a random allocation is likely to perform better than the case where all pairs are included

**Example using Kebeles in the Lalo Asabi Woreda – 1000 possible random allocations of Kebeles to Treatment and Control**



**Stats**
Difference top 3 pairs = 3m
Difference all pairs = 59m
Difference random mean=39m

# Conclusion on the benefits of pair-matched cluster randomization for TNS

**Benefits**

❑ PMCRTs provide operational flexibility: pairs of Kebeles can be excluded from the evaluation, others included, without affecting the validity of the experiment

❑ When there are multiple Woredas, and the best pairs of Kebeles are selected from each, the balance between the Treatment and Control groups is optimized on the variables of interest

❑ The greedy matching algorithm enables us to balance on multiple variables at once – so the balance is optimized not only along one dimension, but multiple dimensions

❑ If all Treatment and Control Kebeles in a Woreda are selected into the program, then an alternative to pair-matching can be considered … we provide one such balancing algorithm we developed for future use by TechnoServe

❑ PMCRTs, as we will see in the next chapter, allow for several specifically tailored analysis techniques

# Inference options and guidelines

## *Framing the discussion about inference in the case of few clusters*

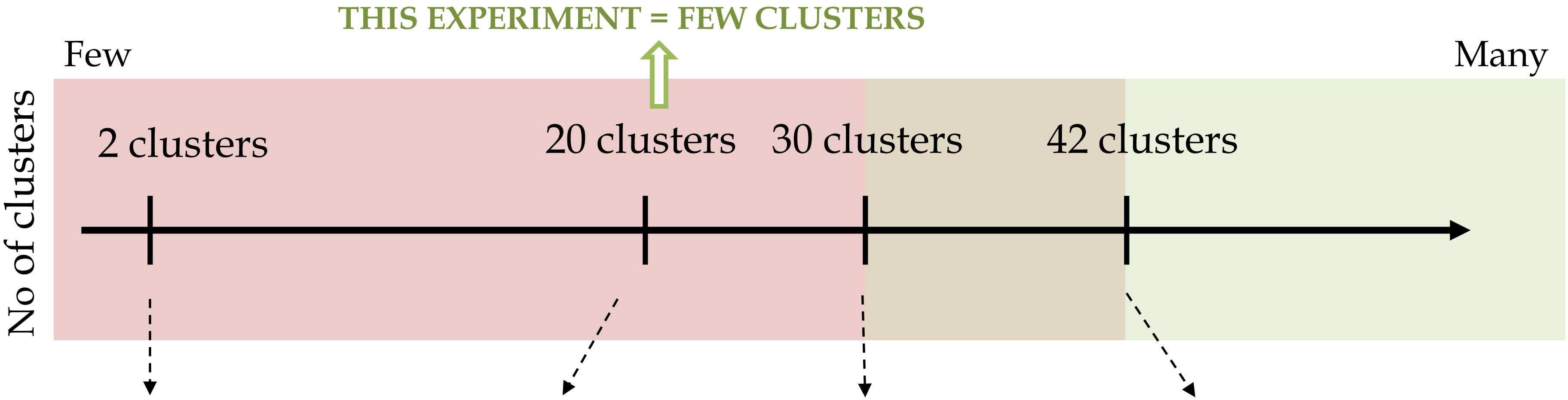
# Determining the right analytic approach depends on where we situate this experiment on the fuzzy "Few vs Many" clusters scale

**Few clusters**

- ❑ Analytic tools to adjust standard errors and account for clustering still being developed
- ❑ This is at the moment a very very active area of research in program evaluation

**Many clusters**

- ❑ Analytic tools very well established
- ❑ Variety of options including difference in difference estimators with cluster robust standard errors, feasible GLS estimation, etc



One treatment, One control

Carter et al (2010) find that if "effective number of clusters" < 20, the robust clustered standard errors fail (note effective #clusters << #actual)

30 is usually the benchmark where the asymptotic properties of test statistics start kicking in

Angrist (2008) and others refer to about 42 clusters as the minimum required for reliable inference using standard cluster robust variance estimators

# Why is it complicated to analyze the results of a cluster-randomized experiment in the case of few clusters (i.e. in this case)?

**There are 3 main problems:**

1 **The estimators – in particular the difference-in-difference estimator – can be inconsistent**. Inconsistency comes from:

- ❑ The fact that we cannot distinguish what change is due to the program effect vs what is due to the cluster effect
- ❑ There can be spatial correlation between clusters
- ❑ And, in the case of non-compliance, non compliance can be related to cluster specific effects (i.e. not just individual self-selection), further complicating the analysis

2 **It is very difficult to calculate standard errors and hence conduct inference in the case of few clusters** (this is because asymptotic properties haven't kicked in). In fact, this is an issue that has not yet been fully resolved in the literature. It's a very active area of research, with a variety of competing models being set forward.

3 **Few clusters lead to low statistical power.** This is mainly because of intra-cluster correlations (see section on ICC) and because of the additional adjustments to cluster-robust variance estimators that are required in the case of few clusters.

# Given analytic uncertainties with the few clusters scenario, we propose that TNS calculate the impact of the program using several approaches

**We propose three distinct strategies aimed at estimating impact:**

3 options

**Randomization inference**. Randomization inference provides exact statistics, by exploiting the random allocation of Kebeles to the Treatment and Control groups to conduct inference. Methods are well developed, and provide solutions for measuring the intention-to-treat effect (ITT), the treatment effect on the compliers (CACE) and quantile treatment effects. We look at two types of techniques: **Fisher**-based and **Neyman**-based techniques that have been developed over the past few years.

**Difference with a lagged dependent, assuming unconfoundedness given the lagged outcomes, and the wild cluster bootstrap**. The wild cluster bootstrap is one out of many possible techniques to correct test statistics to enable inference in the case of few clusters. It is a technique that is proven to result in reliable p-values.

**Synthetic controls**. Synthetic controls is a technique that enables the creation of "synthetic control" regions ex-post (i.e. a synthetic linear combination of control regions). Inference is conducted at the group level. Requires at least two time periods; only enables estimate of Intention-to-Treat effect.

# Combined these strategies provide options with different analytical and in particular operational implications for TNS

## Two key points to keep in mind ….

Inference can be conducted both at the individual and/or group levels

- In the case of few clusters, methods involving cluster-level inference are better developed
- There are a variety of techniques that allow for mixed inference, taking both individual and group data into account. They tend to offer greater power, but not in all cases.
- **If TNS decides to conduct inference at the group level only, then it is not necessary to collect baseline and endline data on the same individuals**. This can have large implications on cost and operations.

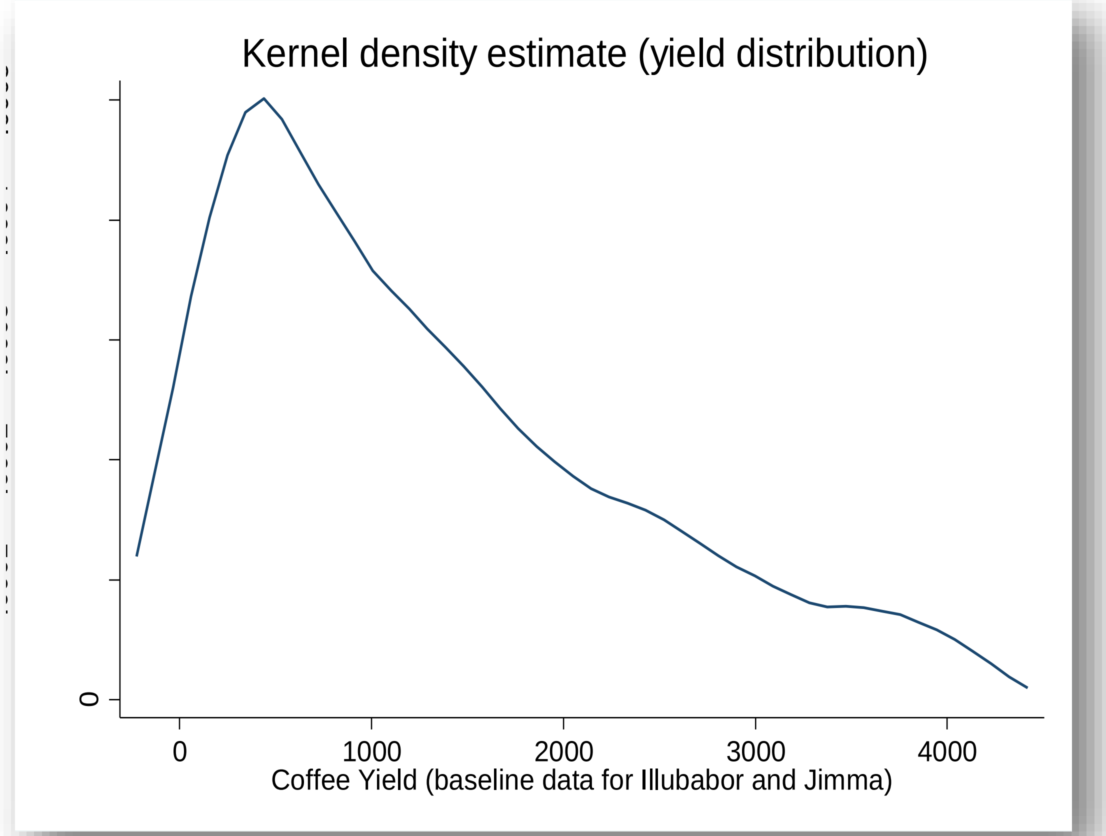
Conduct inference based on one or multiple time periods?

- Some of the techniques proposed here enable perfectly valid inference with only endline data (albeit with significantly lower power) or multiple time periods.
- **This means that TNS still has the option of conducting a valid evaluation for the yield sample for Cohort 2015**.
- We recommend a minimum of two time-periods, but all techniques proposed can be adapted to multiple time periods
- The more time-periods available, the greater the chances of achieving sufficient statistical power

# Remember that the best strategy for the yield sample is not necessarily the best for the BP sample, because data is distributed differently (1)

**Distribution of Individual Yield Data.**

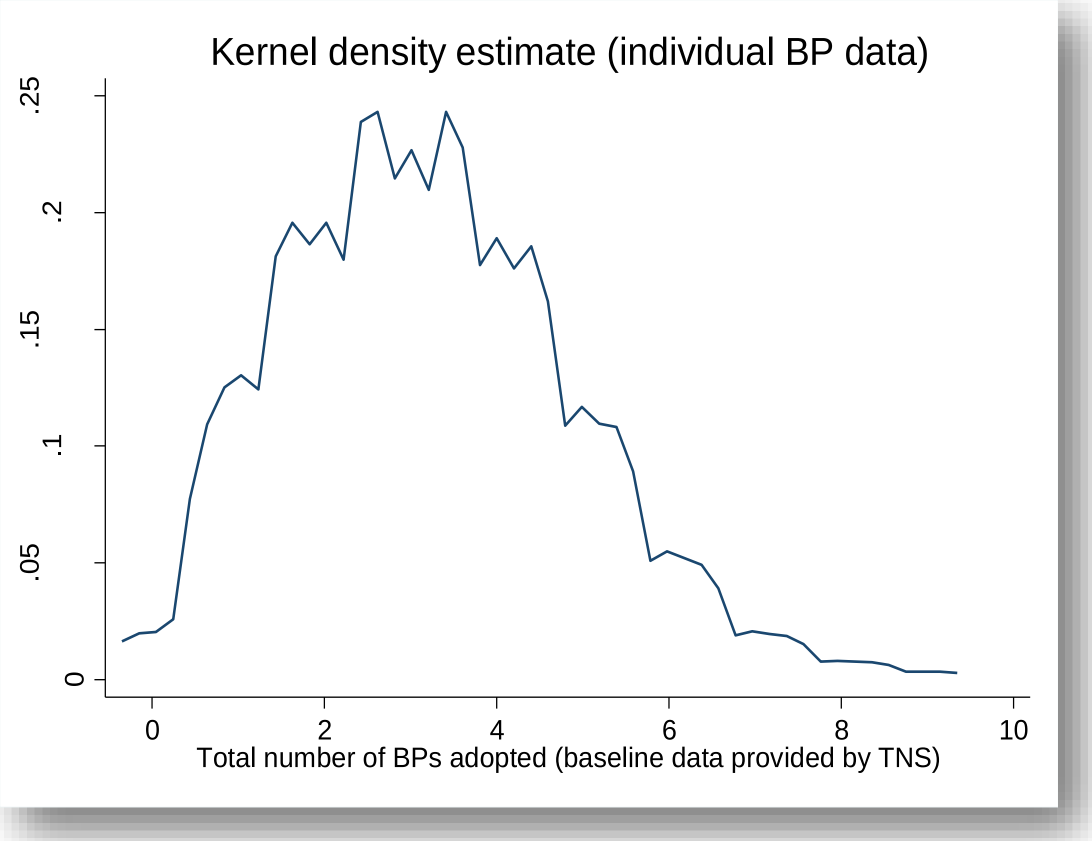
At the individual level the yield sample does not follow a normal distribution. As can be seen in the graph, the density of the yield function decreases is highly skewed to the right. This suggests that the impact is also likely to be non-normal – there is a cap on the  maximum feasible yield levels and low-yield farmers should in theory be able to benefit the most from the intervention. Models that assume normally distributed underlying data would have low power levels do detect an effect in the case of the yield sample (this is because the shape of the data makes standard deviation high … leading to lower power).



# Remember that the best strategy for the yield sample is not necessarily the best for the BP sample, because data is distributed differently (2)

**Distribution of Individual BP data.**

Individual BP data, while not fully normally distributed looks like it could be estimated quite efficiently by a normal distribution. The structure of the data, suggests the impact might also be normal. Hence, models that assume normality might be much more efficient choices at the individual level for BP data.



# So how do we determine which strategies might work and which are likely to fail for both samples? (1)

**Our proposed approach ….**

1. **Model the likely behavior of the yield and BP samples**

TNS has sufficient data on the yield and BP samples enabling us to design realistic models of how the data will behave at the cluster and individual level over time. The modelling we propose (see next few slides) have in-build randomization, which makes it possible to take multiple random draws of the data … i.e. create multiple “model datasets” that we can use to test how well inference models are likely to perform.

2. **Estimate the power of various inference tests under certain parameters using Monte-Carlo simulations based on these models and compare**

Standard power calculation formulas, that are based on asymptotics (i.e. simplifications when the number of individuals or groups is large), do not apply in the case of few clusters. Moreover, remember that each test requires a different sort of power calculation … there is no inherent power in a dataset, it really depends on what test you want to do. Finally, the idiosyncratic structure of the data – that has a huge influence on power – is not easily captured using a few parameters (e.g. mean, standard error) that you can plug into a power formula. Modelling makes it possible to run Monte Carlo simulations (many many random draws) to estimate the power of some of these tests based on certain parameters and the (projected) structure of the data.

# So how do we determine which strategies might work and which are likely to fail for both samples? (3)

**Our proposed approach ….**

3. **Provide some insights and advice on how to implement these tests and their likelihood of success**

We describe and recommend 8 different tests that TechnoServe could implement to estimate the effect of the program on the yield and Best Practice samples. The Monte-Carlo simulations, which we implement where possible, provide some pointers as to how successful these tests are likely to be under certain parameters.

**However, please remember:**

- **The results presented in this section are based on simulations, that might end up being very different from reality**. It's the best informed guess we can make at the moment. Therefore, do not discard potential alternative methods just yet, as they might end up being useful further down the line.
- Note that there are many other possible tests that we have decided not to include, mostly because they do not add any particular advantage over the methods we propose. Other tests that we would like to have added – instrumental variable estimators or quantile effects, under cluster robust inference with few clusters, are still in the process of being developed.
- **All the power calculations presented here are conservative and do not account for potential large gains in efficiency/power from: (i) pair-matched clusters; and (ii) covariate controls.**

# Inference options and guidelines

## *Modelling of yield and BP data at the individual and cluster levels*

# Why model the yield and BP samples?

**Modelling yield and BP samples**

- ❑ The purpose of modelling here is for us to estimate the power and validity of various tests, specially designed for small number of cluster scenarios. For these tests typical power calculations don't work; deriving power calculations from the equations can be complicate, especially given the structure of the data in the yield sample.

- ❑ From TNS data collected in Ethiopia and Rwanda on both yields and best practice adoption, we have very good pointers as to the potential structure of the data at the individual and cluster levels, for both baseline and endline data. This enables us to create relatively realistic "model data" for both the yield and best practice samples.

- ❑ Modelling, involves assigning random distributions to the data at the individual or cluster level (normal, chi-squared, etc), thereby making it possible to make thousands of random draws – i.e. create thousands of different datasets that model the real behavior of the BP and yield data.

- ❑ In our case, **Monte Carlo simulations** basically involve creating thousands of random "model datasets" for the yield and best practice samples, and based on these estimating how likely we are to find a statistically significant impact of the treatment, given certain parameters.

# What do these models include?

**Modelling parameters**

❑ Modelling will result in a distribution of individual and Kebele level data at the baseline and endline that is consistent with previous yield and BP data from Rwanda and Ethiopia, yet will include some level of randomization.

❑ Compliers are identified randomly, given a user-specified level of compliance in the treatment group. We assume only compliers benefit from the program (an assumption that will not fully hold in reality, but simplifies the modelling) and that non-compiers behave in the same way as control group members.

❑ The model includes random missing variables at the endline, reflecting likely attrition levels and bad data collection. We assume different levels for treatment and control. This is a useful feature to include as it creates realistic imbalances in the data that are likely to affect the power of inference tests.

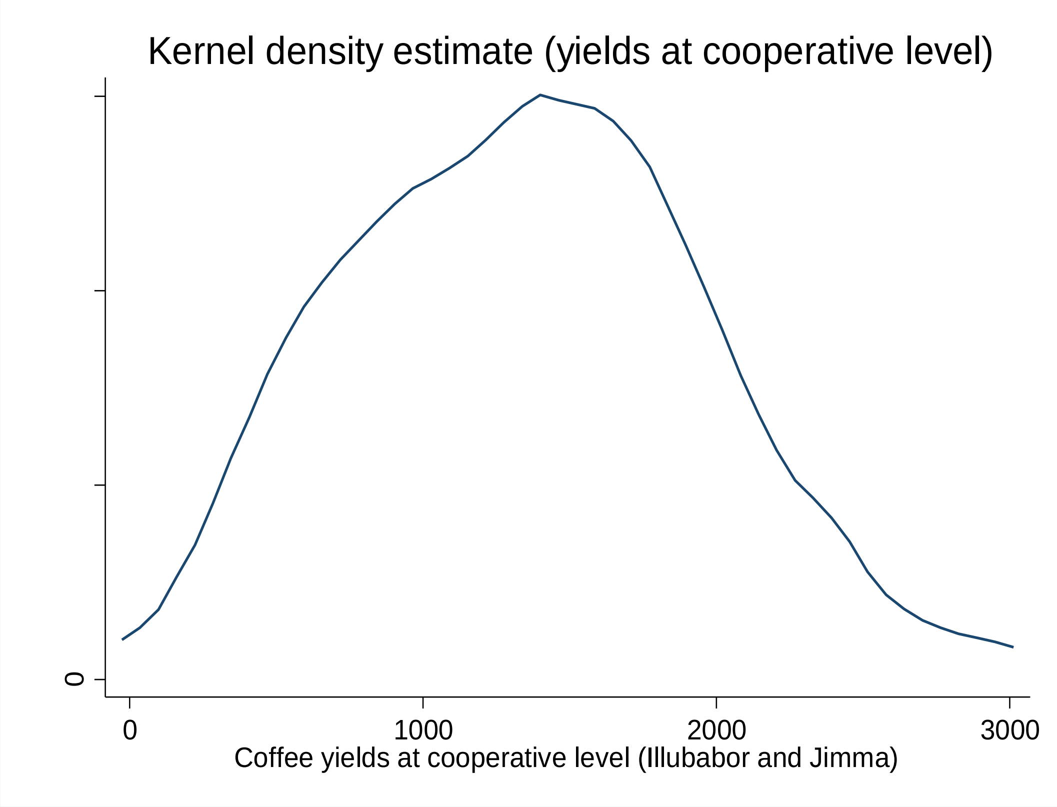
❑ Modelling is flexible to different levels of: (i) target sample size; (ii) expected change between baseline and endline (or minimum detectable effect desired); (iii) different number of pairs; (iv) different levels of compliance; and (v) different levels of attrition.

# Modelling yield data: step 1, assign Kebele averages at baseline following a normal distribution

**Modelling yield data – step 1.**

- First we randomly assign Kebele averages following a normal distribution, with means and standard deviations obtained from TNS data from Illubabor and Jimma

- As can be seen in the graph (right), based on Illubabor and Jimma data, yield levels follow a normal – looking distribution at the cooperative level (i.e. not the individual level!) at the baseline.



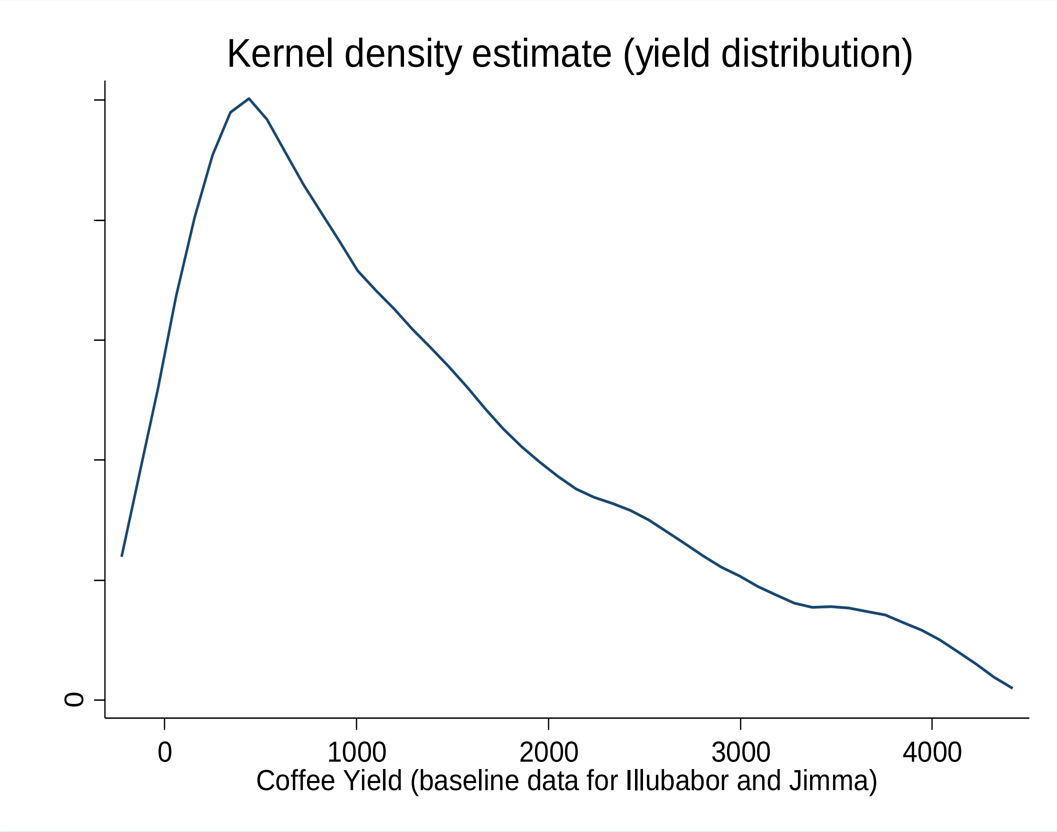
# Modelling yield data: step 2, assign individual yields following chi-sq distribution, with adjustments, and scaled using Kebele averages

**Modelling yield data – step 2**

- To keep it simple in Stata, the distribution of individual yield data is modelled using a chi-squared distribution with 2 degrees of freedom (actual distribution is better modelled by a gamma or poisson distribution, but more complicated).

- We correct for too low or too high yields using cut-off points, and "Normal" variation around the cut-off

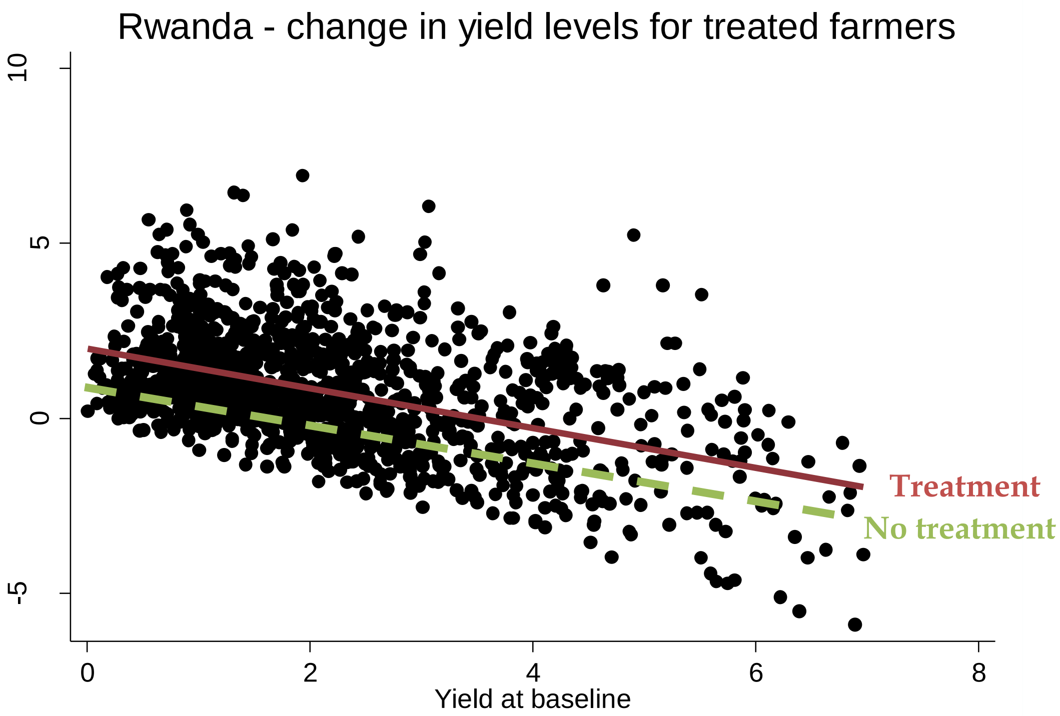
- Data is scaled using Kebele averages, so that individual averages match Kebele-level data



# Modelling yield data: step 3, estimate endline data for compliers, non-compliers, and control group members using model (1)

## Modelling yield data – step 3

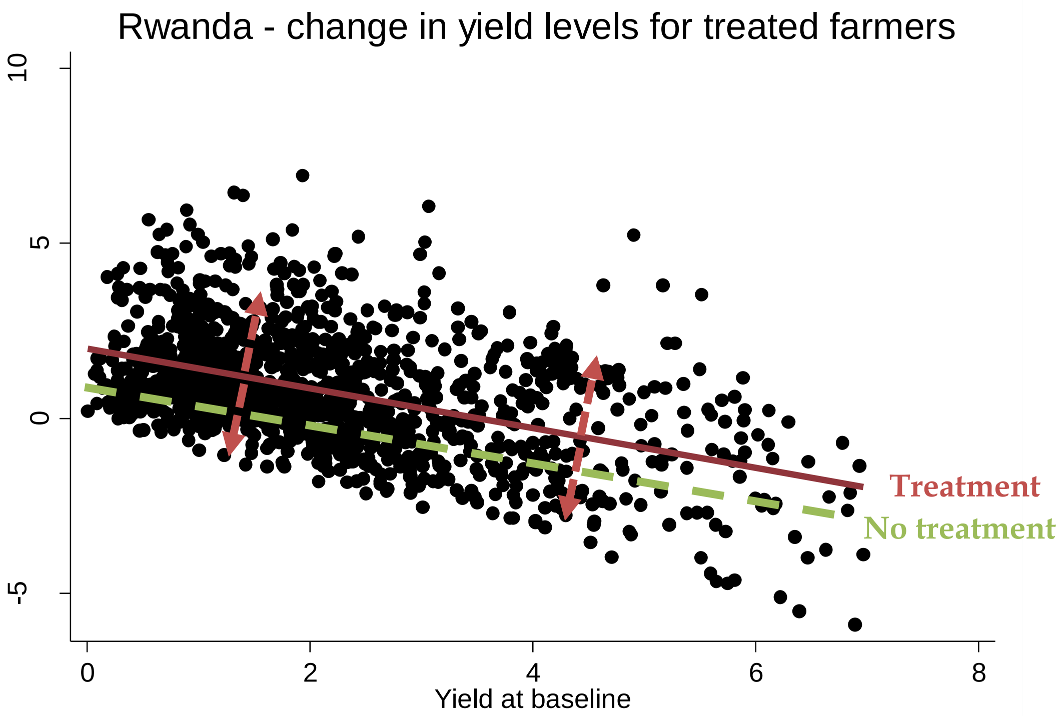
- From Rwanda data, we can see there is a negative, almost linear relationship, between endline/baseline data for treated farmers (data is two years apart, so takes cyclicality into account). We speculate that this is due to: (i) a constant treatment effect; and (ii) a negative association between baseline and endline data, regardless of treatment.
- We therefore assume the change between baseline and endline would follow a similar linear relationship in the case of no-impact.



# Modelling yield data: step 3, estimate endline data for compliers, non-compliers, and control group members using model (2)

**Modelling yield data – step 3 (contibued …)**

- We assign endline values using the intercept obtained from the regression of the change in yields (observed in Rwanda) on baseline yield levels. Call this the "potential yield".
- We then let yields vary around the "potential yield" following a normal distribution of mean 0, with standard deviation derived from the regression error terms and adapted to the Ethiopia case where yield = kg/ha.
- We correct for target sample means by adjusting the constant



**Model:** $\boldsymbol{Change_i = -0.6 * Baseline_i + C + \tau Z_i + \varepsilon}$

where C is a constant estimated at the cluster level, $\tau$ the additive treatment effect, $Z_i$ a dummy for whether individual $i$ was treated or not (i.e. complied), and $\varepsilon$ a normally distributed error term

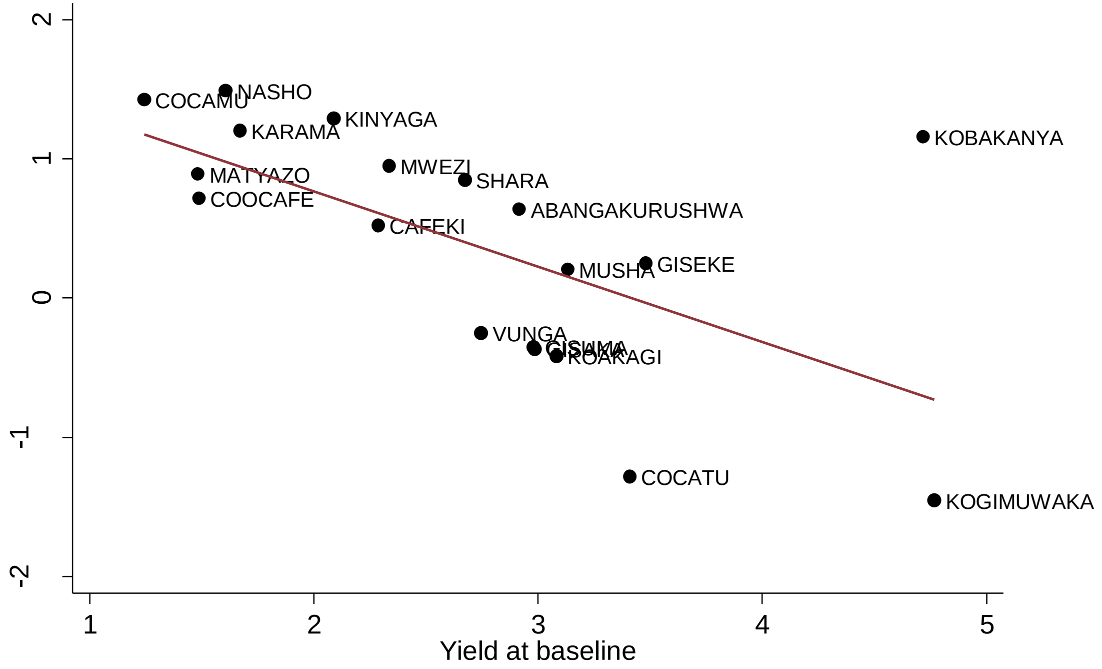
# Modelling yield data: step 3, estimate endline data for compliers, non-compliers, and control group members using model (3)

## Modelling yield data – step 3 (contibued …)

- Note, that this pattern is consistent both at the individual and cooperative levels (at least based on Rwanda data). So while on average yields at the cooperative level increased, they decreased for cooperatives that had high yield levels at the baseline. This could be due to measurement error that is corrected over time, or reflect the cyclicality of the coffee (even though data is two years apart).

- This structure of the data has a very negative impact on power calculations

Rwanda - change in yield levels for treatment cooperatives



# Some immediate implications based on the model … (1)

**In general, what does the structure of the yield data imply about potential evaluation strategies?**

❑ If the model is valid, then typical difference-in-difference estimates will lead to biased results, unless the match between the treatment and control is perfect (which is very unlikely in the case of few pairs/clusters). This is because the key hypothesis of parallel trends between control and treatment in the absence of treatment will be violated. Consider this scenario to see why:

❑ The treatment sample consists of two individuals (T1 and T2) – one with a baseline yield of 500 kgs/ha, the other with a baseline yield of 1500 kgs/ha, hence a mean yield at the baseline of 1000ksg/ha. Assume for simplicity that the constant term C=1000, the error term equals 0, and that all treated farmers see their yields increase by $\tau = 500$ kgs/ha compared to the counter-factual. Then at the endline we would observe a mean increase in yield of 900 kgs/ha for this treatment group (not 500!), given that:

$$\Delta T_1 = -0.6 * 500(B) + 1000(C) + 500(\tau) = 1200 \text{ kgs/ha}$$
$$\Delta T_2 = -0.6 * 1500(B) + 1000(C) + 500(\tau) = 600 \text{ kgs/ha}$$

$$\Delta Treatment = \frac{1200+600}{2} = 900\text{kgs/ha}$$

❑ If at the baseline the control group also consisted of two individuals, with exactly the same levels of baseline yields (or exactly the same mean yield at the baseline), then …

# Some immediate implications based on the model … (2)

**Continued …**

… then a difference-in-difference estimate would result in an average treatment effect estimate (ATE) of exactly 500 (just take the $\tau = 500$ out of the equations for T1 and T2).

- ❑ Let’s now imagine a different scenario where the control group has a slightly different mean at the baseline of 800kgs/ha, hence a baseline difference of about 200kgs/ha with the treatment group (so the balance of treatment/control is not fully achieved at the baseline, a very realistic scenario in the case of few clusters). Here individual 1 (C1) has a baseline of 300kgs/ha and Individual 2 (C2) a baseline of 1300 kgs/ha – hence for an average of 800kgs/ha. In this case we would observe an average increase in yields of about 460kgs/ha in the control between baseline and endline and obtain a difference-in-difference estimate of 380kgs/ha, i.e. 120kgs/ha lower than the actual effect of 500kgs/ha.

$$\Delta C_1 = -0.6 * 300(B) + 1000(C) = 820\text{kgs/ha}$$
$$\Delta C_2 = -0.6 * 1300(B) + 1000(C) = 220\text{kgs/ha}$$

$$\Delta Control = \frac{820+220}{2} = 520\text{ kgs/ha}$$

$$Difference\ in\ difference = 900(\Delta\text{Treatment}) - 520(\Delta\text{Control}) = 380\text{ kgs/ha}$$

# Some immediate implications based on the model … (3)

**Continued ….**

- ❑ **What this simple example tells us, is that unless pair-matching is excellent (which with few clusters is very unlikely to be the case), typical difference-in-difference estimates will yield biased average treatment effects**. The direction of the bias will depend on the baseline difference between treatment and control (a positive bias if control is higher at the baseline; and a negative bias if the control is lower). This bias is likely to be proportional to the size of the difference between baseline and endline estimates. We believe the bias in difference-in-difference estimates will be about 0.6 times the size of baseline differences between treatment and control (if Rwanda data apply to the case of Ethiopia).

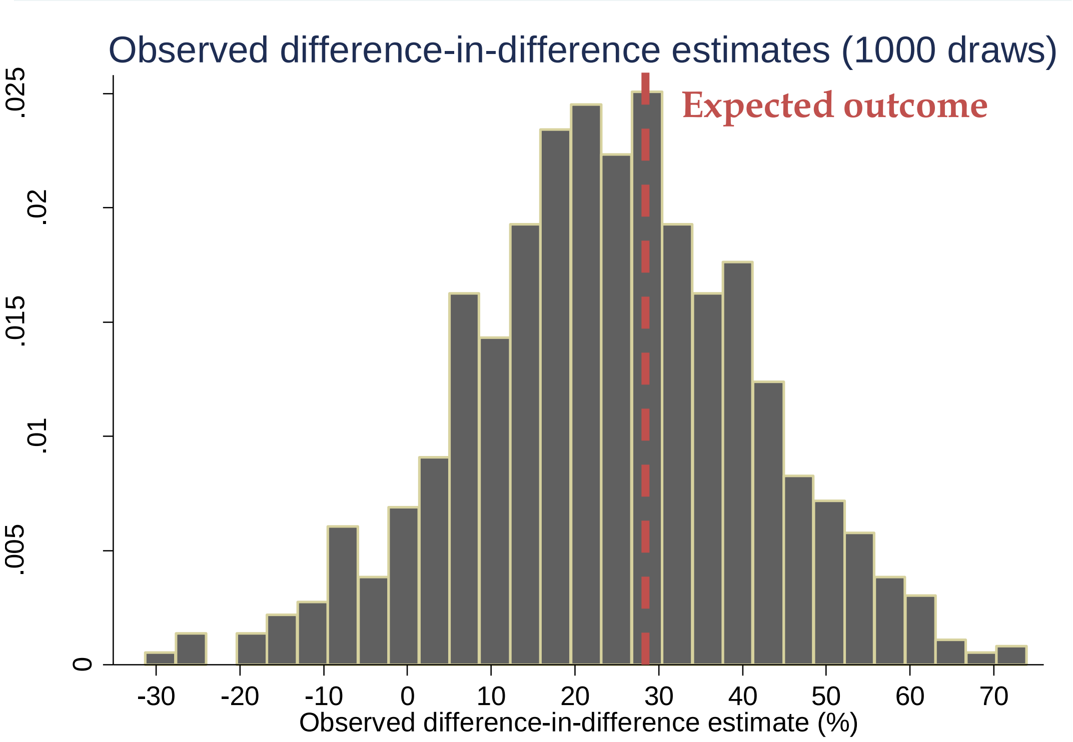
- ❑ **This strongly suggests that non-parametric models such as randomization inference, quantile treatment effects, or Change-in-Change models, are likely to work best for the Yield Sample**. We also recommend the synthetic controls approach. The downside is that many of these evaluation techniques are very complicated and are only just being adapted to the case of clustered data in the case of few clusters.

- ❑ Note also that the estimated parameters for the yield sample, in particular a Rho of about 0.18 and a baseline/endline correlation of about 0.31, suggest that typical impact evaluation techniques will have very low power (see next two simulation slides).

# Simulations (yield model): observed difference-in-difference if program actually leads to 50% mean increase for treated farmers

**Parameters for Monte Carlo Simulations of Difference-in-difference estimates**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Yield follows ChiSq distribution with 2 degrees of freedom
- Delta (yield): 580 (50% increase)
- Compliance=58%
- Pairs:11



- Positive Intention-to-Treat Effect observed in 91% of cases
- Expected ITT or above (=delta*compliance) observed in about 40% of cases
- Mean impact observed 24%, expected is 29%.

# Simulations (yield model): estimates of rho and baseline/endline correlation based on 1000 runs

**Parameters for Monte Carlo Simulations of Difference-in-difference estimates**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Yield follows ChiSq distribution with 2 degrees of freedom
- Delta (yield): 580 (50% increase)
- Compliance=58%
- Pairs:11

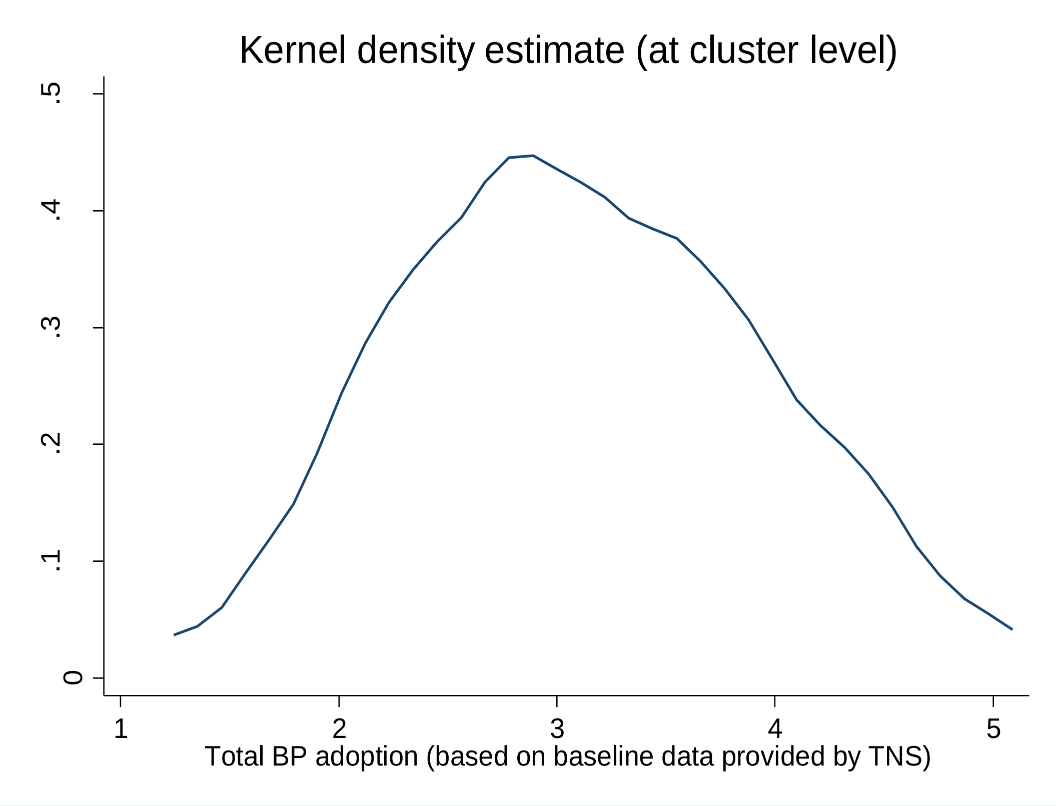
| Parameter | Runs | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|
| Rho baseline | 1000 | **0.180** | 0.053 | 0.035 | 0.356 |
| Rho endline | 1000 | **0.169** | 0.050 | 0.036 | 0.392 |
| Correlation baseline/endline | 1000 | **0.318** | 0.091 | -0.049 | 0.577 |

- These levels of Rho are consistent with data from both Ethiopia and Rwanda – and are likely to be anywhere between 0.8 and 0.28, so the range of possibilities is quite large.
- Note that the baseline/endline correlation of results at an average of about 0.318 is quite low. This suggests that standard power calculations will yield very low power levels.

# Modelling BP data: step 1, assign Kebele averages at baseline following a normal distribution

**Modelling BP data step 1**

- Kebele level BP data looks very much like a normal distribution, based on baseline data provided by TNS from previous Ethiopia based Cohorts.
- The first step in the modelling process is therefore to create a normal distribution of baseline BP data at the Kebele level, following mean and standard deviation estimates obtained from that same dataset.
- Note that we use the sum of Best Practices (out of 11) as the dependent variable of interest.



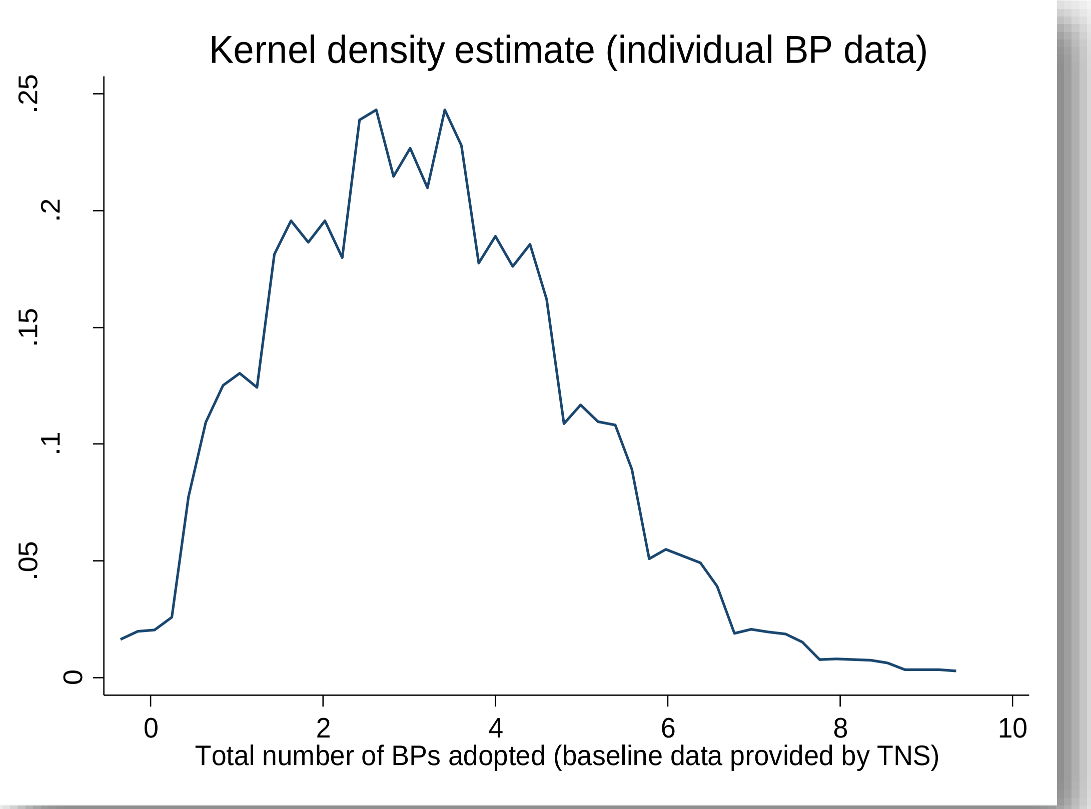
# Modelling BP data: step 2, assign individual baseline data following a normal distribution within each Kebele

**Modelling BP data step 2**

- Unlike yield data, BP data at the individual level also follows a normal distribution.

- We therefore assign individual level baseline data following a normal distribution, that has as mean the Kebele-level BP adoption score. Standard deviations within each Kebele are estimated as a function of the Kebele mean, using real data.

- We translate the data into integers bounded by 0 and 11 (the maximum number of BPs permitted)

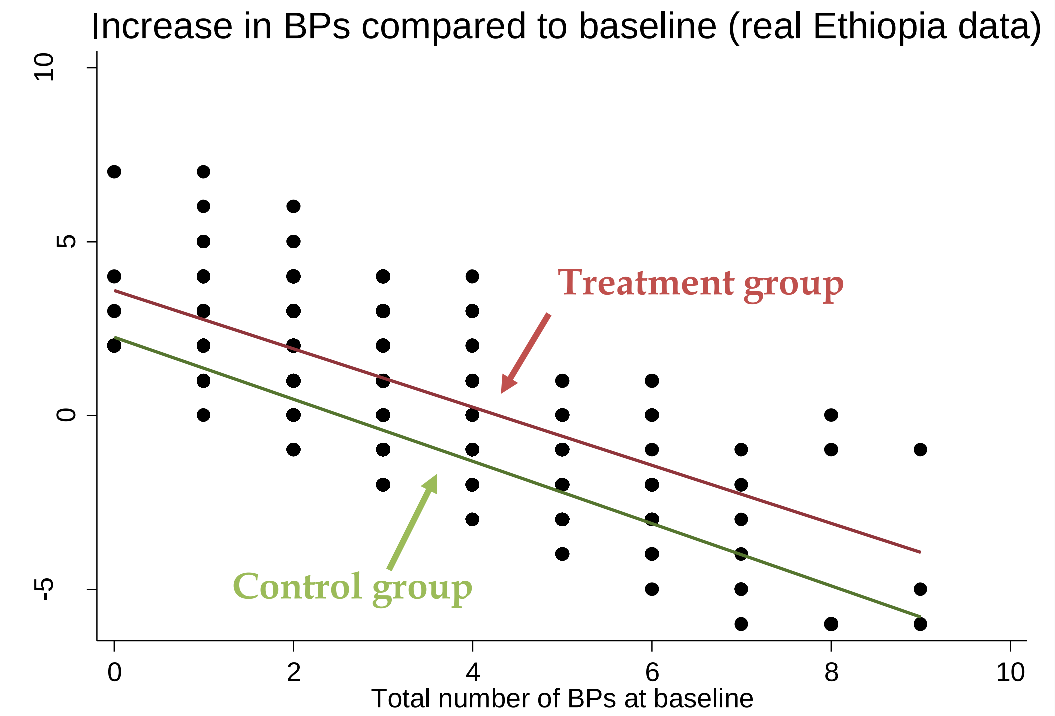


# Modelling BP data: step 3, estimate endline data for compliers, non-compliers, and control group members using model

**Modelling BP data step 3**

- **Remarkably, Ethiopia data confirms for BPs exactly the same hypothesis we had developed for yield data**: i.e. there is an inverse relationship between the potential increase in BPs and baseline BP levels for both treatment and control group farmers (the lines in the graph are actual regression lines).

- What this result suggests is that the higher your baseline BP levels, the greater the inherent downward pressure on potential future outcomes, regardless of treatment.



**Model:** $\boldsymbol{Change_i = -0.86 * Baseline_i + C + \tau Z_i + \varepsilon}$

where C is a constant estimated at the cluster level, $\tau$ the additive treatment effect, $Z_i$ a dummy for whether individual $i$ was treated or not (i.e. complied), and $\varepsilon$ a normally distributed error term

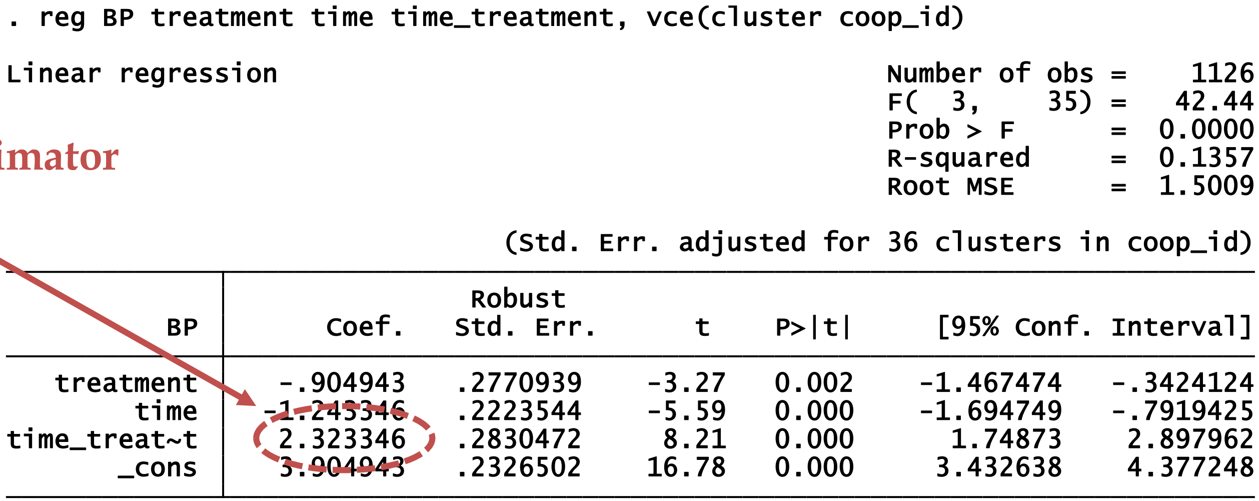
# Some immediate implications using the real example of BP data in Ethiopia … (1)

**In general, what does the structure of the BP data imply about potential evaluation strategies?**

- ❑ If the model is valid, then for BPs as well typical difference-in-difference estimates will lead to biased results, unless the match between the treatment and control is perfect (which is very unlikely in the case of few pairs/clusters). Here is an example of how DiD estimators fail, using the real example of baseline and endline BP data provided by TNS on Cohort 2013 in Ethiopia.

- ❑ A DiD estimation of the treatment effect would yield an impact estimate of 2.3 – ie**. TNS would conclude that on average treatment led to a 2.3 increase** in the number of BPs of treated farmers (ignore the standard errors estimates here, they are wrong).

This is the DiD Estimator



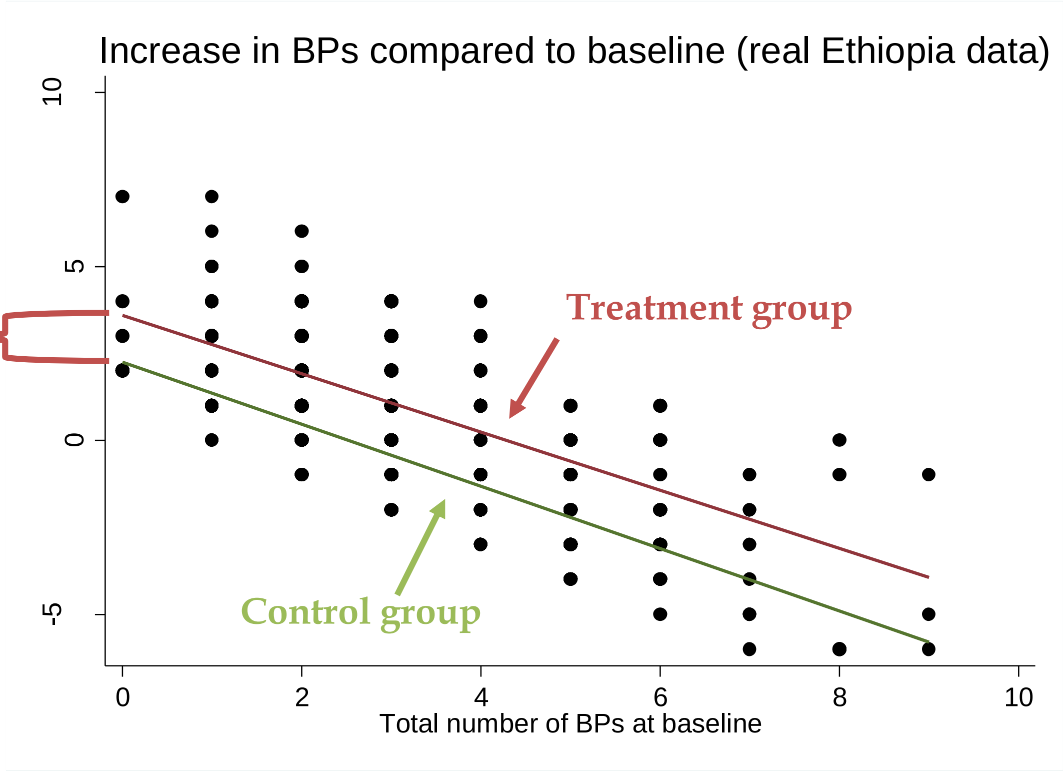
```
. reg BP treatment time time_treatment, vce(cluster coop_id)

Linear regression                               Number of obs =     1126
                                                F(  3,    35) =    42.44
                                                Prob > F      =   0.0000
                                                R-squared     =   0.1357
                                                Root MSE      =   1.5009

                                (Std. Err. adjusted for 36 clusters in coop_id)
------------------------------------------------------------------------------
             |               Robust
          BP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   -.904943   .2770939    -3.27   0.002    -1.467474   -.3424124
        time |  -1.243346   .2223544    -5.59   0.000    -1.694749   -.7919425
time_treat~t |   2.323346   .2830472     8.21   0.000      1.74873    2.897962
       _cons |   3.904943   .2326502    16.78   0.000     3.432638    4.377248
------------------------------------------------------------------------------
```

# Some immediate implications using the real example of BP data in Ethiopia … (2)

**Continued …**

- ❑ In fact, the real impact is the difference between the two regression lines in the graph – i.e. which is essentially the difference between the two intercepts. **Using the intercepts as a proxy for impact, we find a revised estimate of about 1.35 BPs (which is not even included in 95% confidence intervals of the previous estimate!)**

**The real estimated impact=+/-1.35 BPs**



# Some immediate implications using the real example of BP data in Ethiopia … (3)

**Continued …**

- **So how is it possible that TNS would overestimate the actual impact of the program by 173% using DiD methods?**

- Because the assumption of parallel trends, which is the basis of difference-in-difference calculations fails! This is due to the inverse association between baseline and endline data – in this case with or without treatment, BP levels in the control group would have decreased more than BP levels in the treatment group. We can understand this using the model we propose and using real data from TNS Ethiopia.

- We use the following estimates to complete the model, derived directly from the data: C=2.247, $\tau$ = 1.35, baseline in the treatment group = 3 BPs, baseline in the control = 3.9 BPs

- Starting with the treatment group, let's measure actual outcomes based on the model (they match reality)

$$\Delta Treatment = -0.86 * 3(\mathrm{B}) + 2.247(\mathrm{C}) + 1.35(\tau) = 1.017 \text{ (in reality 1.08)}$$

# Some immediate implications using the real example of BP data in Ethiopia … (4)

**Continued …**

- The expected change in the treatment group in the absence of treatment following the model, would have been -0.33, not 0:

$$-0.86 * 3(\mathrm{B}) + 2.247(\mathrm{C}) = -0.33$$

- Now focusing on the Control group, the expected outcome in the case of no-treatment was about -1.107 not 0 or -0.33. Hence the assumption of parallel outcomes in the case of no treatment does not hold.

$$\Delta Control = -0.86 * 3.9(\mathrm{B}) + 2.247(\mathrm{C}) = -1.107 \text{ (in reality -1.124)}$$

- The adjusted impact estimate can be calculated as follows, which is close to the estimated 1.35 obtained (differences are due to slightly different slopes in treatment and control):

$$\mathrm{I} = (\Delta Treatment_{\text{actual}} - \Delta Treatment_{\text{no treatmnet}}) - (\Delta Control_{\text{actual}} - Control_{\text{expected}})$$

$$I = (1.08+0.33)-(-1.124+1.107)$$
$$I = 1.43$$

# Some immediate implications using the real example of BP data in Ethiopia … (5)

**Continued …**

- ❑ Conclusion: **The DiD estimator in the case of slight differences between baseline and endline results is biased in this case as well**. The bias in the difference-in-difference estimate is positive, because baseline data in the control group was higher than in the treatment group (i.e. there was greater downwards pressure to start with in the control group). Second, the size of the bias is proportion to the size of the difference between baseline and endline (the proportion is determined by the slope of the regression).

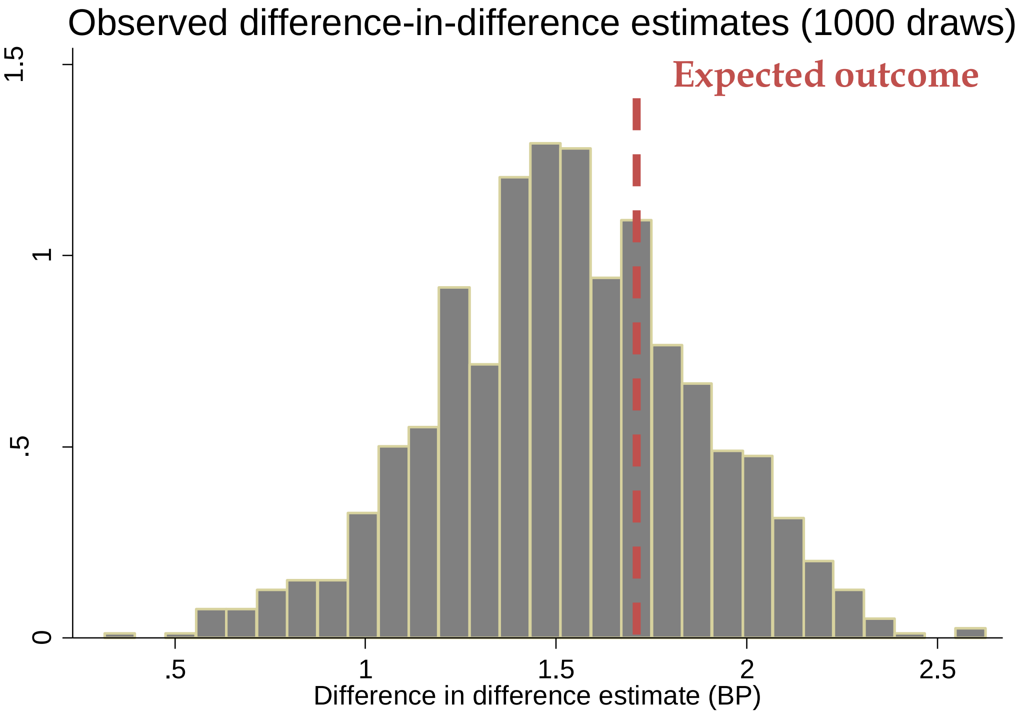
- ❑ **This implies that difference-in-difference strategies are not well adapted to the case of BP data either**. Because of the small number of clusters and high intra-cluster correlation, there are bound to be baseline differences between the treatment and control groups. These small baseline differences, can lead to gross over or under-estimations of the actual effect (here 173%).

- ❑ **Instead of using difference-in-difference estimates we propos to use differences with lagged outcome variables**. Differences with lagged outcomes can yield valid inference based on the assumption of unconfoudedness with the lag (see Imbens and Wooldridge, 2008). Unconfoundedness assumes that "beyond the observed covariates – in this case the lag - there are no (unobserved) characteristics of the individual associated with the potential outcomes and the treatment.

# Simulations (BP model): observed difference-in-difference if program actually leads to desired increase for treated farmers

**Parameters for Monte Carlo Simulations of Difference-in-difference estimates**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11



- Positive Intention-to-Treat Effect observed in 100% of cases
- Expected ITT or above (=delta*compliance) observed in about 36% of cases
- Mean impact observed 1.5, expected impact 1.67.

# Simulations (BP model): estimates of rho and baseline/endline correlation based on 1000 runs

**Parameters for Monte Carlo Simulations of Difference-in-difference estimates**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11

| Parameter | Runs | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|
| Rho baseline | 1000 | 0.224 | 0.058 | 0.072 | 0.388 |
| Rho endline | 1000 | 0.175 | 0.033 | 0.078 | 0.322 |
| Correlation baseline/endline | 1000 | -0.200 | 0.062 | -0.367 | 0.037 |

- These levels of Rho are consistent with data from Ethiopia's 2013 Cohort, where baseline rho=0.2 and endline rho=0.130, which fall within the 95% confidence interval)
- Note that the baseline/endline correlation of results at an average of about -0.2 is very low. This has a very negative impact on power calculations. The actual estimates from the Ethiopia case are at about 0.01, but note that this does not include non-compliers.

# Inference options and guidelines

## *Fisher based randomization inference, applied to pair-matched cluster randomization*

# Some intuition behind Fisher based randomization inference (or what's called "exact inference") before we proceed …

- Here the random allocation of Kebeles to the treatment and control groups is the basis of inference on whether the program had an impact or not
- This randomization leads to only one out of many possible combinations of Kebeles into the Treatment (T) and Control (C) groups
- With 20 Kebeles, there are 184,756 different possible allocations of Kebeles into a T & C group (each consisting of 10 Kebeles)!!!
- There are also 1,024 different possible allocations of pairs of Kebeles, when Kebeles are paired and then randomly assigned to T and C as is the case here
- Each of these 1,024 combinations has an equal chance of being selected, but only one possible combination ends up being implemented
- Let's call each of the 1,023 combinations that weren't selected a "Placebo" combination
- Randomization inference, consists in looking at what would have happened if any of the 1,023 other "Placebo" combinations had been selected
- This means we look at the actual results as if the allocation to T&C was different
- If there was no program impact, then regardless what the combination of T & C Kebeles is, we should find little or no impact
- If the program did have an effect, then we should find that the actual selected combination leads to an usually high result compared to the "Placebos". This is called exact inference because we look at all the possible alternative allocations of Kebeles to T & C (it's not based on estimates, but the exhaustive list of possibilities)

# What makes the following 4 “randomization inference” based tests attractive for this evaluation? (1)

- ❑ All the tests presented here are based on Rosenbaum (2002-2010) and some of his co-authors, including Small (2008) and Ten Have. Rosenbaum’s randomization inference techniques are derived from Fisher, a statistician of the early 20th century. How does randomization inference differ from the difference-in-difference type evaluations we have been accustomed to and why are they useful in this case?

- ❑ First of all, inference is not based on asymptotics (i.e. approximations of probabilities when the number of individuals or groups go to infinity, i.e. >30 at least), but rather alternative histories if randomization had turned out differently (i.e. permutations). Statistics in this case are called exact, because the number of alternative permutations (possible random allocations of Kebeles to treatment and control) are known and finite. These methods are also non-parametric, in that they do not assume that the data has any particular underlying structure/distribution (e.g. unlike linear regressions). This presents two key advantages: (i) it enables us to overcome the clustered standard error problem; and (ii) it enables us to conduct inference regardless of the structure of the BP/yield data, which we have shown is problematic in the DiD case.

- ❑ These methods test a sharp null hypothesis of an “additive” treatment effect. The sharp null, means that inference is not conducted looking at whether the average mean is different than 0 (or a given null hypothesis), but rather by testing whether all individuals/units experienced a given constant treatment effect or no effect at all.

# What makes the following 4 "randomization inference" based tests attractive for this evaluation? (2)

- ❑ "Additive" means that all farmers experience a similar increase in yield or BP levels compared to the (unobserved) counterfactual, regardless of their baseline yield or BP levels. This is a very attractive feature of these techniques in this case - as we have seen in the yield and BP models, the effect appears to be "additive".

- ❑ These models and specific test statistics, have been developed for the pair-matched cluster randomization case. Hence they enable us to fully exploit the benefits of PMCRT. Variants of the model, allow for tests in the case of non-compliance and quantile effects.

- ❑ Inference can be conducted either at the group level (in which there is no need to interview the same people at the baseline and the endline) or a the individual level (which is the preferred scenario as panel data is always better, but not necessarily required to conduct valid inference).

- ❑ Lastly, given that inference is fully based on permutations of alternative outcomes of the randomization process, inference can be conducted with 1, 2, 3 … as many time periods as you want. This means that TNS can still conduct valid inference on the yield sample for Cohort 2015 (albeit with reduced statistical power).

- ❑ Remember that all the power calculations presented here are conservative: they do not take into account the potential benefits of successful pair-matched cluster randomization and covariate controls, which could be large … unfortunately we don't know how large in advance.

# Test 1: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, at the group level (1)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, one-sided test
- Method and test statistic specifically adapted to pair-matched cluster randomization
- References: Rosenbaum (2002, 2010)

## Why?

This test will enable us to determine (statistical power permitting) whether the program had an Intention to Treat Effect of τ (a number) on yield / best practice adoption levels. Inference is conducted at the group level, with group level controls. By varying the τ that we test for, we can identify a minimum level of impact and a point estimate of the impact, called the **Hodges-Lehmann estimate**. We propose to use what's called the straightforward to implement **Wilcoxon matched-pairs Signed Rank Test**, which is an exact inference test, to test for the significance of results.

Inference in this case is exact. There is no doubt / approximations on the standard errors which use permutations of random allocations of Kebeles to T & C as the basis for inference. This method also enables us to control for covariates, thereby increasing power. Nevertheless, in general, power will tend to be lower than alternative methods that impose restrictions/assumptions (i.e. are not exact) in order to gain efficiency.

# Test 1: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, at the group level (2)

**Test parameters**

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, one-sided test
- Method and test statistic specifically adapted to pair-matched cluster randomization
- References: Rosenbaum (2002, 2010)

## How?

For detailed explanation, see Rosenbaum (2002). The intuition: to take into account covariance, inference is conducted not using the outcomes, but the residuals after controlling for covariates.

Step 1: Calculate the "observed response" for each pair

- Calculate the difference-in-difference increase of cluster-averaged BP/yield levels within each pair (we call this $Y_p$)

Step 2: Calculate the "adjusted response" for each pair under τ

- The adjusted response is the "observed response" minus τ (this is our null hypothesis about the impact of the program)
- $A_p$=$Y_p$- τ  (where p is the pair number)

Step 3: Calculate residuals assuming impact τ, after controlling for covariates of interest (at the cluster level) in regression

- Select covariates of interest at group level (call them X)
- Select a regression … suggest using a non-parametric machine learning method called Kernel Regularized Least squares (KRLS)

# Test 1: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, at the group level (3)

**Test parameters**

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, one-sided test
- Method and test statistic specifically adapted to pair-matched cluster randomization
- References: Rosenbaum (2002, 2010)

## How? (continued …)

Step 3 (continued)

- Estimate following regression: $A = \beta X + \varepsilon$, where X are all the covariates of interest  (there should be no constant term)
- We are interested in capturing the error term $\varepsilon$ (i.e. the residuals), which will form the basis of the inference
- In practice, download KRLS package for stata (ssc install krls) and run the following command: krls a x1 x2 x3
- Follow-up with the command: kpredict a_pred (this predicts the adjusted A). Then create $\varepsilon$=a-$\beta$ ×a_pred (where $\beta$ is the coefficient of the regression)
- This can be done using other types of regressions as well (we like KRLS because it's easy to use and is non-parametric)

Step 4: Calculate the test statistic for a Wilcoxon's signed ranked test

- Simply use the Stata signtest command to test this statistic, by typing: signtest resid=0
- Make sure to look at the one-sided p-value statistic

# Test 1: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, at the group level (4)

**Test parameters**

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, one-sided test
- Method and test statistic specifically adapted to pair-matched cluster randomization
- References: Rosenbaum (2002, 2010)

## How? (continued …)

Step 4 (continued)

- The test value (W) will be calculated as follows (it's a sum of ranks :

  $W = \sum_{i=1}^{P} Rank_i \times S_i$ with p ∈

  $[1, \dots, P]$ where P is the total number of pairs

- The test statistic is as follows:

$$\text{Test statistic} = \frac{W - \frac{P(P+1)}{4}}{\sqrt{\frac{P(P+1)(2P+1)}{24}}}$$

- We can reject the null hypothesis that the impact = τ if the p-value corresponding to this one-sided test statistic is <0.05.
- Note that by setting τ =0, this would enable us to test the hypothesis of no impact as well.

# Test 1: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, at the group level (5)

**Test parameters**

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, one-sided test
- Method and test statistic specifically adapted to pair-matched cluster randomization
- References: Rosenbaum (2002, 2010)

## How? (continued …)

Step 5: Loop on different values of τ to obtain confidence intervals

- Repeat steps 1-4 using different values of τ. Values for which the null is rejected form part of the confidence interval

Step 6: Final step – calculate point estimate of impact

- To estimate the Intention-to-Treat effect, we equate the obtained test-value W to the expected value of W if impact was indeed τ (i.e. if the null were true) following Hodges and Lehmann (1986)
- The expected W in this case is:

$$E(W) = \frac{P(P+1)}{4}$$

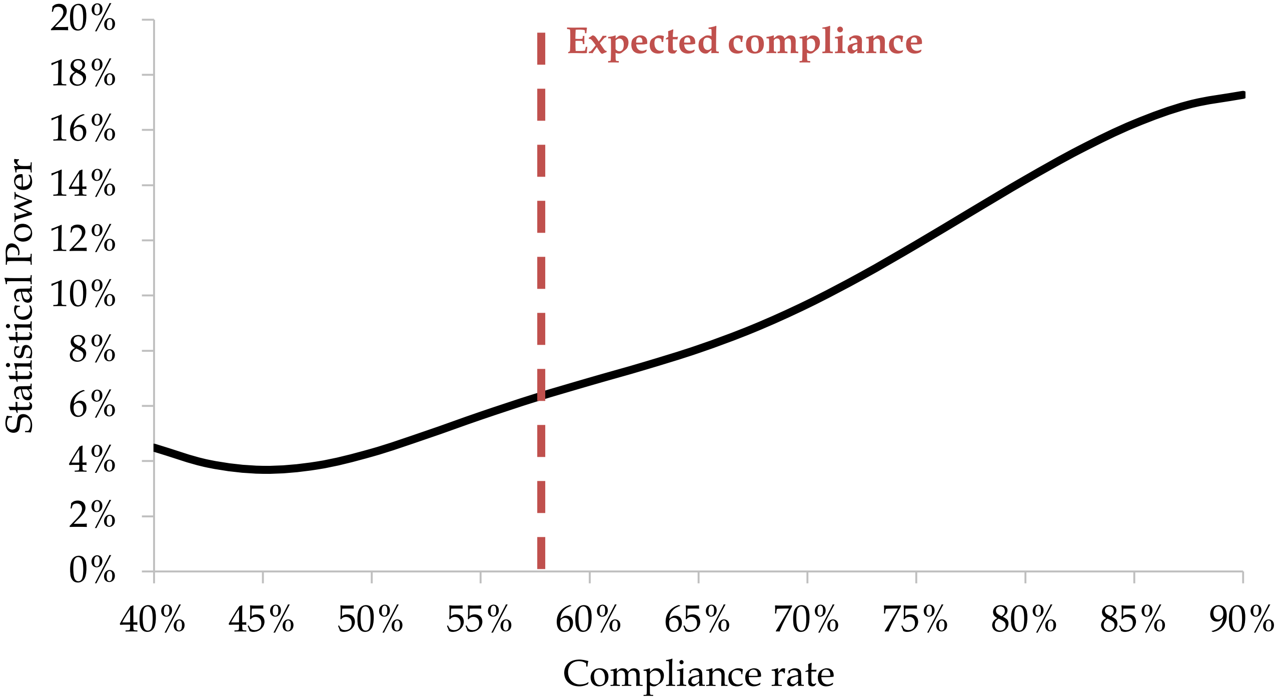
- The τ for which $W_\tau \approx$ E(W) is our estimate of the Intention-to-Treat effect

# Monte Carlo simulations (excluding efficiency gains) suggest that the power to detect an impact on yields will be very (!) low using Test 1

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

Test 1: (Yields) Estimated power to detect an effect with 11 pairs of Kebeles, based on compliance rate

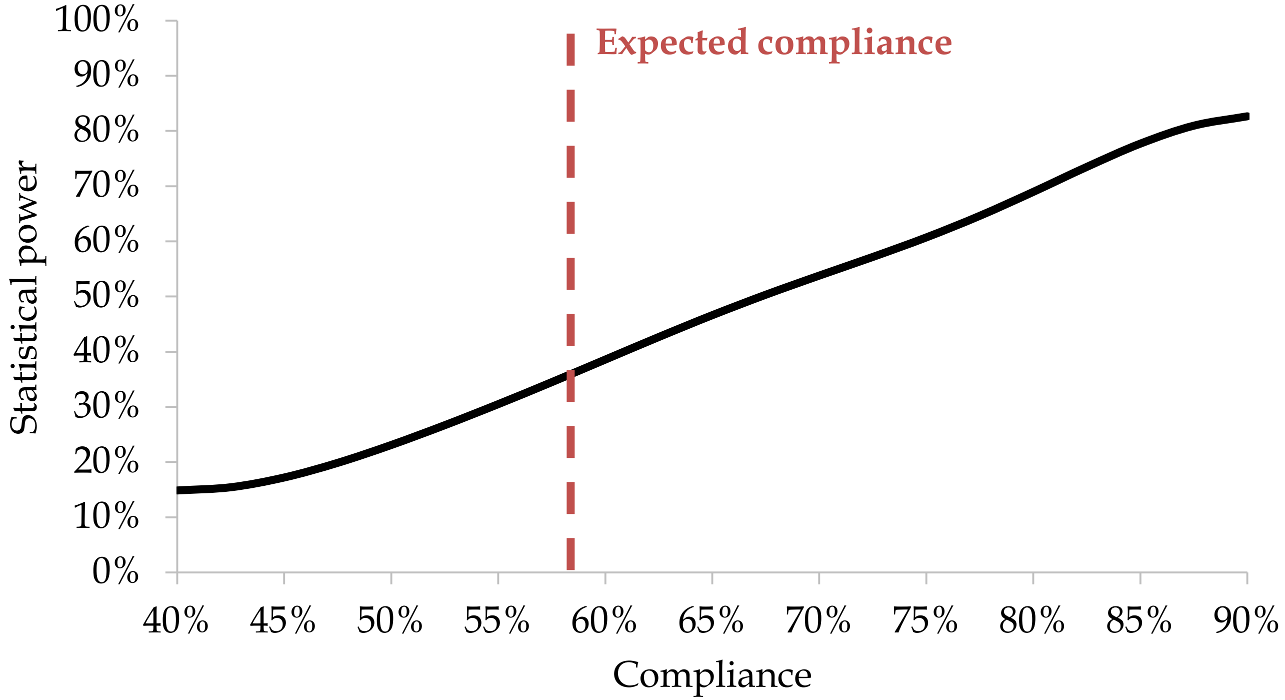


# Monte Carlo simulations (excluding efficiency gains) suggest that the power to detect an impact on yields will be low using Test 1

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11

Test 1: (BPs) Estimated power to detect an effect on BPs, based on compliance rate



- We speculate that with a compliance rate of about 90% this test will have sufficient power (80%) to detect an effect of 2.89 BPs, which is a large minimum detectable effect.

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (1)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## Why?

This test of the Intention-to-Treat effect is an enhanced version of Test 1, except that we work with individual level data and group level randomized. The advantage is greater power (especially when data non-normal) through individual controls. However it is slightly more complicated to implement than the cluster-level version of the test. It also requires that TNS collect data on exactly the same individuals at the baseline and endline, a requirement that is not needed in Test 1. Note that the same test could be conducted with endline data only (with equal validity), albeit with lower power.

The power of this test is superior to that of Test 1, but is highly sensitive to intra-cluster-correlation (see Small et al, 2008). Small proposes changes to the statistic, leading to power improvements, but we do not implement them here. Small adjusts for the effects on intra-cluster correlation using weights. We would recommend getting an expert statistician to implement this enhancement in the test statistic.

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (2)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## How?

For detailed explanation, see Rosenbaum (2010) or Small et al (2008). The intuition: test2 follows the same logic as test1, but adapted to individual level data, with adjusted test statistics. These statistics need to be computed manually.

Step 1: Calculate "observed response" for each individual

- Calculate the increase in the yield or BP level for each individual between endline and baseline (we call this $Y_{pki}$), where *p* refers to the pair number, *k* to the Kebele number within each pair (k=1 or 2), and *i* to the individual in each pair. We call k=1 the Treated Kebele, and k=0 the Control Kebele.

Step 2: Calculate "adjusted response" for each individual under τ

- The adjusted response is the "observed response" minus τ (this is our null hypothesis about the impact of the program)
- Do this using the following formula: $A_{pki}$=$Y_{pki}$- τ$Z_k$ where p is the pair number, and $Z_k$ takes the value 1 if the Kebele was included in the treatment, and 0 otherwise

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (3)

**Test parameters**

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## How? (continued …)

Step 3: Calculate individual residuals assuming impact τ, after controlling for covariates of interest in regression

- Select individual and group level covariates of interest (X)
- Estimate following regression: $A_{pki}$=$\beta K_{pki}$ + $\varepsilon_{pki}$, where X are all the covariates of interest (there should be no constant term)
- We are interested in capturing the individual error term $\varepsilon_{pki}$(i.e. the residuals), which will form the basis of the inference
- In practice, download KRLS package for stata (ssc install krls) and run the following command: krls a x1 x2 x3
- Follow-up with the command: kpredict a_pred (this predicts the adjusted A). Then create $\varepsilon$=a-$\beta$ ×a_pred (where $\beta$ is the coefficient of the regression)
- This can be done using other types of regressions as well (we like KRLS because it's easy to use and is non-parametric).

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (4)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## How? (continued …)

Step 4: Calculate the test value (W)

- This test statistic (see Rosenbaum, 2010) is essentially a weighted average of the W-statistic introduced in test 1, but will need to be programmed into Stata (we have not found a specific Stata command that can run this)
- The test value is as follows (assuming pairs are arranged such that the first Kebele in the pair is always the treated one!):

$$W = \sum_{p=1}^{P} \frac{1}{n_{p1} + 1} \sum_{i=1}^{n_{p1}} \sum_{j=1}^{n_{p2}} u_{p1i2j}$$

where $u_{p1ikj}$ = 1 if $\varepsilon_{p1i}$>$\varepsilon_{plj}$; $u_{p1ikj}$ = -1 if $\varepsilon_{p1i} < \varepsilon_{plj}$; and $u_{p1ikj}$ = 0 if $\varepsilon_{p1i} = \varepsilon_{pkj}$. $n_{pk}$ is the sample size of Kebele k, in pair p

- Actually implementing this is bit complicated and computationally heavy, but doable… we have provided an example of how to do this in the simulations do-file for test2

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (5)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## How? (continued …)

Step 5: Calculate the null distribution and estimate test statistic

- To calculate the null distribution, we have to calculate W for all the possible permutations of treatment and control, of which there are essentially $2^p$ possibilities
- So this basically involves repeating step 4 $2^p$ times. This can be by enumeration (running a loop over all the possible $2^p$ allocation of Kebeles to treatment and control) or by generating random draws of Kebeles to Treatment and Control to estimate the null distribution
- Once completed, count how many of these permutations produce a W higher or equal to the one obtained for the real experiment (call this z)
- Divide z by $2^p$ and you obtain the p-value for the test
- If the p-value is <0.05 then we can reject the null hypothesis that the impact =τ, hence the alternative is more likely and impact > τ (this is a one-sided test)

# Test 2: Randomization inference, one-sided test of Intention-to-Treat effect τ, with covariance adjustment, and individual data (6)

## Test parameters

- Non-parametric test: Intention to Treat Effect = τ
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010), Small et al (2008)

## How? (continued …)

Step 6: Final step – calculate point estimate of impact

- To estimate the Intention-to-Treat effect, we equate the obtained test-value W to the expected value of W if impact was indeed τ (i.e. if the null were true), which here is 0: i.e. E(W)=0
- Our point estimate of τ is therefore the τ that brings the W of the experiment as close as possible to 0
- By varying τ from 0 up, we can: (i) test the hypothesis of no effect; (ii) estimate a confidence interval for the impact of the program; and (iii) obtain a point estimate for the Intention-to-Treat effect

Notes

- Small (2008) proposes an improved and more powerful statistic, that adjusts the weights of the W estimator in line with intra-cluster correlation estimates. This adjustment greatly improves the power of the proposed estimation technique when intra-cluster correlation is high. Would recommend using it, but with the help of an expert statistician.

# Monte Carlo simulations (excluding efficiency gains) suggest that power to detect an impact will be better than for test 1 for yields

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Yield follows ChiSq distribution with 2 degrees of freedom
- Delta (yield): 580 (50% increase)
- Compliance=58%
- Pairs:11

**Estimated power if effect true = +/-47.3% (one-sided)**

(This estimate is based on 300 Monte-Carlo runs … these are few runs to get a good estimate, but the method is computationally intensive. These random draws took about 1h to run on Stata)

- Note that for tests 2,3, and 4 we do not conduct extensive power calculations. Looping over the runs is computationally too cumbersome. Even with very few runs, power calculations take up to 2-3 hours to run.

# Monte Carlo simulations (excluding efficiency gains) suggest that for the BP sample as well, power using this test might be low

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11

**Estimated power if effect true = +/-59.6% (one-sided)**

(This estimate is based on 300 Monte-Carlo runs … these are few runs to get a good estimate, but the method is computationally intensive. These random draws took about 1h to run on Stata)

# A few more insights about why Test 2 (and subsequently 3 & 4) is likely to perform much better than alternatives that assume normality?

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Yield follows ChiSq distribution with 2 degrees of freedom
- Delta (yield): 580 (50% increase)
- Compliance=58%
- R01 (individual)=70%
- Pairs:11

- **First of all because Test 2 (like Test 1, 3, and 4) is non-parametric**. It doesn't assume that the data has any type of particular distribution. This is a very attractive feature of randomization inference, especially when one goes through the effort of conducting an expensive randomized trial. Why then put the validity of the experiment at risk by assuming the normality of the underlying data or other type of distributions?

- **Second, because Test 2 (tests 3,4 as well) is sensitive to the individual ranking of the effect, rather its mean and standard deviation**. When one assumes normality, what matters is the mean and standard deviation of the effect. Here what matters is how each individual in the treatment group compares to individuals in the control group – better or worse, regardless of how much better or worse they are. Because an "additive" impact will have the largest proportional effect on low-yield farmers, many low yield farmers in the treatment group will see an impact. The number of farmers shifting from low to higher yield levels in the Treatment group vs Control group gives this test its power; not the size of the effect vs the standard deviation.

# Test 3: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment, and individual data (1)

## Test parameters

- Non-parametric test: Effect of Treatment Proportional to Dose = $\beta \times dosage$
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## Why?

This is an extension of both tests 1 and 2. A small tweak to the model, which better reflects reality, enables us to also take into account the effect of non-compliance. This instrumental variable-based model, applied to group randomization but individual compliance, works with non-compliance in both the Treatment and Control groups (i.e. treatment group members that don't take-up the treatment, or control group members that do). The intuition behind the model, is that treatment is not anymore considered equal across individuals, but rather proportional do the "dosage" received. We could define "dosage" here as share of sessions attended by a farmer for example. Alternatively, we could simplify and say a farmer has complied if he/she attended more than 50% of sessions (in line with the TNS definition).

This is an instrumental variables approach, because we are in essence using the random assignment of Kebeles to the Treatment and Control groups as a (random) instrument that only affects outcomes through its impact on dosage levels.

# Test 3: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment, and individual data (2)

**Test parameters**

- Non-parametric test: Effect of Treatment Proportional to Dose = $\beta \times dosage$
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## How?

For a detailed explanation, see Rosenbaum (2010). The intuition: we replace the assumption that impact = $\tau$ by a formula that captures the fact that impact is proportional to the dosage received. We formalize this as follows: $impact = \beta(d_{T,i}\text{-}d_{C,i})$ where $d_{T,i}$ is the dosage received by individual *i* if in Treatment, $d_{C,i}$ the dosage received in Control, and $\beta$ a measure of impact that is proportional to the "dosage received", i.e. an individual's level of compliance. Of course only one of the two states (treatment or control) is observed in reality.

Step 1: Calculate "observed response" for each individual

- Calculate the increase in the yield or BP levels for each individual between endline and baseline (we call this $Y_{pki}$), where *p* refers to the pair number, *k* to the Kebele number within each pair (k=1 or 2), and *i* to the individual in each pair.

# Test 3: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment, and individual data (3)

## Test parameters

- Non-parametric test: Effect of Treatment Proportional to Dose = $\beta \times dosage$
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## How? (continued)

Step 2: Calculate the "adjusted response" for each individual

- Under the hypothesis of non-compliance, the calculation of the adjusted response (A) is different than in Tests 1 and 2. Here we calculate A as follows:

$$A_{pki}=Y_{pki}- \beta d_{pki}$$

where p is the pair number, $k$ the Kebele within the pair and $d_{pki}$ is the dosage received by individual $i$

- We can define dosage as either as the share of sessions attended or as compliance, where $d_{pki}$ = 1 if the individual/household has attended 50% of sessions or more, or 0 otherwise.

- The remaining steps are similar to test 2 (see steps 3-5 of test 2).

# Test 3: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment, and individual data (4)

**Test parameters**

- Non-parametric test: Effect of Treatment Proportional to Dose = $\beta \times dosage$
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## How? (continued)

Notes

- This test is more likely to detect an effect $\beta$ than test 2 because the sharp null hypothesis only applies to compliers (i.e. the adjusted response will only affect compliers).
- This model allows for non-compliance in the control group as well. Control group members that end up-receiving the treatment will also have an "adjusted outcome" that is calculated in the same way as for compliers in the treatment group. We assume that there is no non-compliance in the control group, but this might not end up being the case in reality.
- Note as well that this model will run the test assuming that there are no spill-over effects for non-compliers in the treatment group. This is not a realistic scenario, as we would expect non-treated farmers to benefit as well. One idea to overcome this issue might involve having a minimum level of dosage for all treatment group members that captures spill-over effects.

# Test 3: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment, and individual data (5)

## Test parameters

- Non-parametric test: Effect of Treatment Proportional to Dose = $\beta \times dosage$
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## How? (continued)

Notes (continued …)

- We can find evidence of the existence of spill-over effect by comparing non-compliers in the Treatment group to control group farmers. The size of spill-over effects can be estimated by matching non-compliers in the Treatment Group to Control group farmers in the matching pair, based on individual level characteristics at the baseline.
- If TNS defines dosage as the number of sessions attended, then one good modelling assumption might be to introduce decreasing returns to the dosage received. The logic is that the more classes you attend, the lower the marginal effect of that extra class. Or classes might end up showing increasing returns to attendance. Either way, this might be something TNS can consider. An expert statistician could help refine the model to take decreasing or increasing returns to dosage/attendance into account.

# Test 4: Randomization inference, one-sided test of Effect Proportional to Dose $\beta$, with covariance adjustment and quantile effects

## Test parameters

- Non-parametric test: Quantile Effect of Treatment
- Exact inference at group level, but with individual data, one-sided test
- Specifically adapted to the case of matched-pair cluster randomization
- References: Rosenbaum (2010)

## Why?

This model can be further refined to include quantile effects; i.e. not just taking non-compliance into account, but also differential effects by baseline level of yields/BP adoption. While we do not expect to find any differences in the treatment effect between quantiles (we assume the constant treatment effect is actually reflective of reality), we do believe that comparing results at the quantile level will increase the power of the experiments for both the yield and BP cases. This is because of the expected inverse association between yield and BP outcomes, and baseline yield and BP levels.

This test is substantially more complicated than tests 1-3 and we do not develop it here. We would recommend that TNS get an expert statistician to conduct this test. From a modelling perspective this is a sensible approach that relaxes one of the main constraints imposed by randomization inference: the assumption of an "additive" or "proportional" equally sized impact, even though we do believe that the "additive"/"proportional" effects assumptions apply to the case of both the yield and BP samples.

# Inference options and guidelines

## *Neyman based randomization inference, applied to pair-matched cluster randomization (PMCRT)*

# What makes the following 2 Neyman-based "randomization inference" techniques attractive for this evaluation?

- ❑ First of all these techniques (basedon Neyman statistics) have been specially adapted to the case of Matched-Pair Cluster Randomization (PMCRT) by Imai, King and Nall (2009). The authors are the first to explain exactly how to calculate unbiased average treatment effect estimators and variance estimators for PMCRTs (Fisher based randomization inference – see previous slides - are based on the sharp hypothesis of a constant effect across all individuals, with is different from an average treatment effect).

- ❑ These methods are non-parametric and hence do not assume any underlying distribution of the data or the error terms. This is a useful feature for both the BP and yields samples, for which outcomes are a function of baseline values.

- ❑ Given that inference is based on the random allocation of Kebeles to treatment and control within pairs, and are not based on asymptotics, the standard errors do not suffer from the cluster or the few cluster problem.

- ❑ **Do not get discouraged by the low estimated power calculations for these tests!** They are due to the fact that pair-wise matching in our simulations was done at random. In reality the matches are likely to be much better. The better the match, the better the power of these tests. Moreover, if the first round of matching was not successful, an option at the endline is to conduct pair-wise matching ex-post … using more complete data and information.

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (1)

## Test parameters

- Non-parametric test: Cluster Average Treatment Effect
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

## Why?

Both Neyman and Fisher, around 1920-1930, derived inference methods based on randomization. The big difference between Neyman-based inference and Fisher based inference, is that Neyman tests for an average treatment effect (some effects on clusters can be positive, others negative), whereas Fisher tests for the sharp null hypothesis that all the individuals in the Treatment Group experienced a positive impact. Neyman's approach was adapted to matched-pair cluster randomization by Imai, King and Nall (2009), who show that: (i) it is possible to determine accurate estimators for the pair-wise cluster matching case (with known biases that can be bounded); (ii) to determine a generally unbiased (albeit conservative) variance estimator.

This is again a non-parametric method of inference, that does not assume the outcome indicator follows any type of particular distribution. It does assume however, that with a sufficient number of observations within each cluster (following the Central Limit Theorem), outcomes at the pair level will be normally distributed.

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (2)

**Test parameters**

- Non-parametric test: Cluster Average Treatment Effect (CATE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

## How?

For a detailed explanation, see Imai , King and Nall (2010). The intuition: the selected impact estimator (CATE) is a weighted average of within pair differences. Confidence intervals for the estimator are calculated using a variance estimator for CATE, with critical values drawn from the t-statistic with P-1 degrees of freedom, where P is the number of pairs. This is because impact at the pair-level is expected to follow a normal distribution (provided the number of individuals per cluster is sufficient – minimum of about 30 per cluster). Random draws from a normal distribution follow a t-statistic. This experiment is relatively straightforward to implement and can be done in-house by TechnoServe.

Step 1: Calculate "observed response" for each individual

- Calculate the increase in the yield or BP levels for each individual between endline and baseline (we call this $Y_{pki}$), where $p$ refers to the pair number, $k$ to the Kebele number within each pair (k=1 or 2), and $i$ to the individual in each pair.

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (3)

## Test parameters

- Non-parametric test: Cluster Average Treatment Effect
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

## How? (continued …)

Step 2: Calculate the difference-in-difference at the pair level

- Calculate $D_p$, which we define as the difference in the means of the outcome variable ($Y_{pki}$) between the Treatment and Control Kebeles within each pair $p$ (we order the Kebeles such that k=1 corresponds to the treatment Kebele, and k=2 corresponds to the control Kebele). This is in a way a difference-in-difference estimate of the impact of the program at the pair level.

- Formally we can write:

$$D_p = \frac{\sum_{i=1}^{n_{p1}} Y_{p1i}}{n_{p1}} - \frac{\sum_{j=1}^{n_{p2}} Y_{p2j}}{n_{p2}}$$

where $n_{p1}$ and $n_{p2}$ are the number of individuals in the sample in the Treatement (1) and Control (2) Kebeles.

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (4)

## Test parameters

- Non-parametric test: Cluster Average Treatment Effect

- Test statistic based on approach proposed by Imai, King, and Nall

- Specifically adapted to the case of matched-pair cluster randomization

- References: Imai, King, Nall (2009)

## How? (continued …)

Step 3: Calculate the CATE estimator

- To calculate the impact estimator we first need to define $N_{pk}$ which we define as the known population of Kebele k in pair p. The difference with the "small n" is that "small n" corresponds to the sample size (i.e. different from population).

- Next we can calculate CATE, the estimator of interest, which is essentially a weighted sum of the cluster level differences. CATE is defined as follows:

$$CATE = \frac{1}{\sum_{p=1}^{P}(N_{p1} + N_{p2})} \sum_{p=1}^{P} (N_{p1} + N_{p2})\, D_p$$

- Note that this estimator is unbiased if Kebele population sizes are identical within each pair (the smaller the difference, the smaller the bias) or if matching is extremely effective. It is possible to determine the upper bound of this bias.

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (5)

**Test parameters**

- Non-parametric test: Cluster Average Treatment Effect
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

## How? (continued …)

Step 4: Calculate a "conservative" estimate of the variance of CATE

- First calculate the "normalized" weight of each pair (p), defined as:

$$w_p = n \times \frac{(N_{p1}+N_{p2})}{\sum_{p=1}^{P}(N_{p1}+N_{p2})}$$

where n is the total sample size and p is the pair number.

- This number corresponds to the share of the total target population that comes from pair p, times the sample size
- Next calculate the mean impact observed, using these normalized weights. Call this M, such that:

$$M = \frac{1}{n}\sum_{p=1}^{P} w_p\, D_p$$

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (6)

**Test parameters**

- Non-parametric test: Cluster Average Treatment Effect
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

**How? (continued …)**

Step 4: continued …

- Now calculate the following variable at the pair level, which is the square difference between the "observed" impact at the pair level and the mean impact. We call it $S_p$, such that:

$$S_p = (w_p D_p - \frac{n}{p} M)^2$$

- Finally, we can calculate the variance estimator, V, such that:

$$V = \frac{p}{(p-1)n^2} \sum_{p=1}^{P} S_p$$

Step 5: Construct a one-sided confidence interval for CATE

- Given that the distribution of impacts at the pair level is assumed to have a normal distribution (following the CLT), draws from this hypothetical distribution will follow a t-statistic with P-1 degrees of freedom, where P is the total number of pairs

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (7)

## Test parameters

- Non-parametric test: Cluster Average Treatment Effect

- Test statistic based on approach proposed by Imai, King, and Nall

- Specifically adapted to the case of matched-pair cluster randomization

- References: Imai, King, Nall (2009)

## How? (continued …)

Step 5: continued …

- We look up the critical value corresponding to a t-statistic with P-1 degrees of freedom and an $\alpha$ of 0.05 (where $\alpha$ is the level of statistical significance desired). You can obtain that very easily in Stata using the following commands: scalar t=invttail(P-1,0.05) and "display t" to see the result.

- We will construct a one-sided confidence interval of the CATE statistic using this critical value, such that:

$$[M-t\sqrt{V}\,;\ \infty\,[$$

where t is the critical value corresponding to a t-distribution with P-1 degrees of freedom with an $\alpha$ of 0.05.

- If $M-t\sqrt{V}>0$ then then the estimated Cluster Average Treatment Effect (CATE) is statistically significant (one-sided).

# Test 5: Neyman-based inference adapted to PMCRT, one-sided test of CATE estimator (8)

## Test parameters

- Non-parametric test: Cluster Average Treatment Effect
- Test statistic based on approach proposed by Imai, King, and Nall
- Specifically adapted to the case of matched-pair cluster randomization
- References: Imai, King, Nall (2009)

## How? (continued …)

Notes

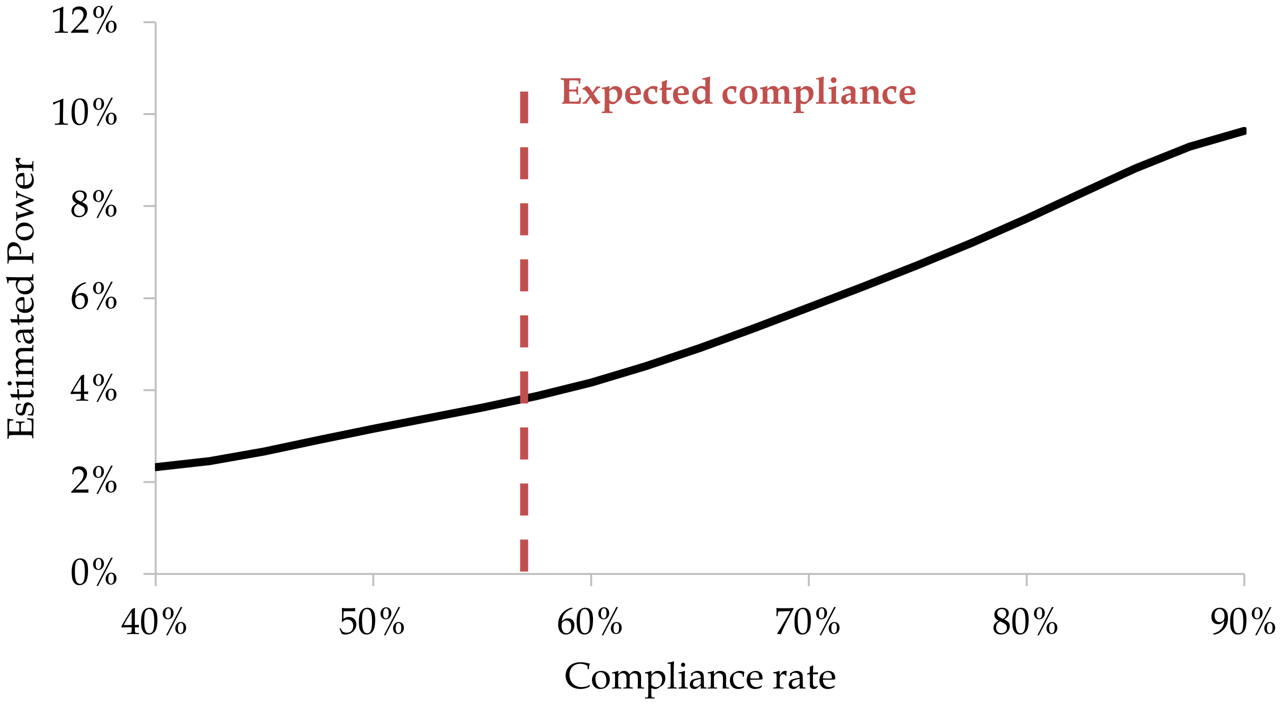
- Some of these calculations can be done automatically using the "experiment" package in R, created by the authors. See http://cran.r-project.org/web/packages/experiment/experiment.pdf for more information

# Simulations (excluding efficiency gains) suggest that power to detect an impact on yields will be extremely low using Test 5

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

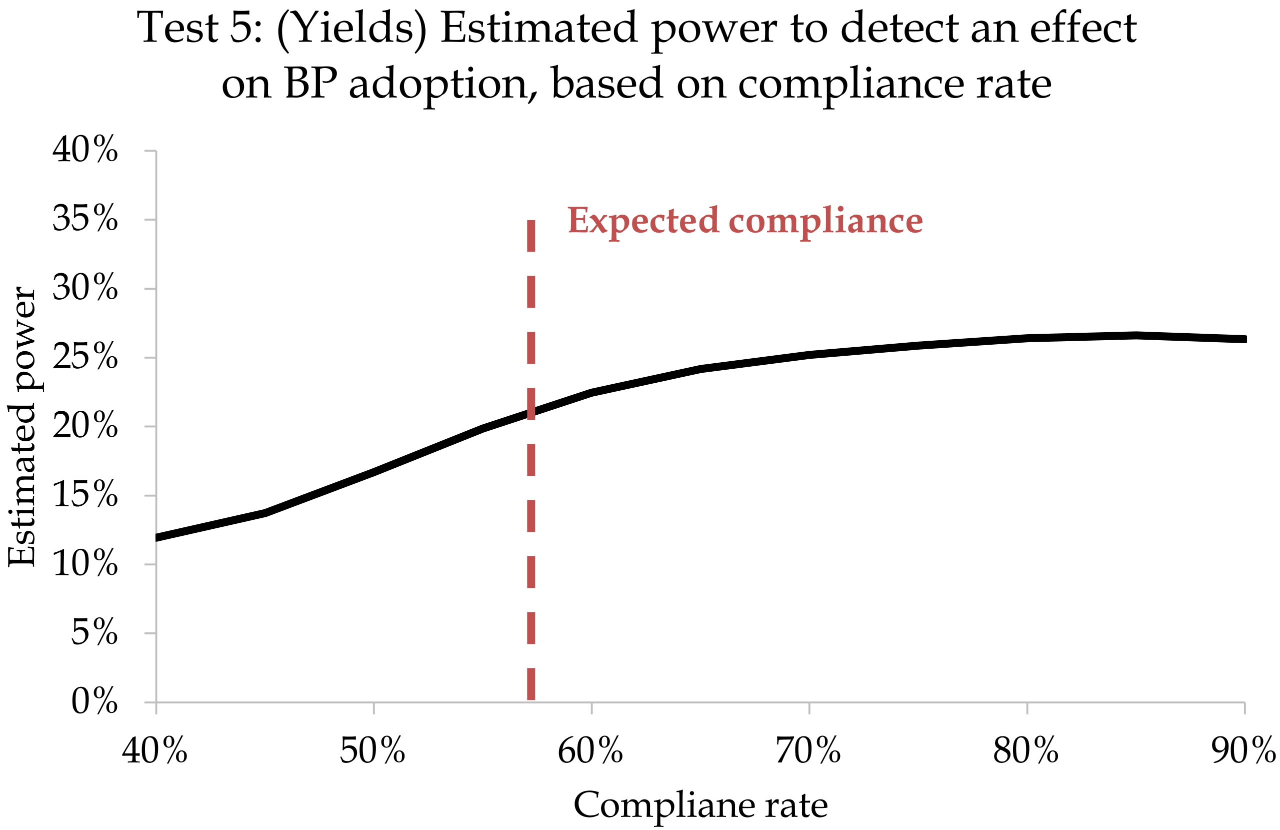
Test 5: (Yields) Estimated power to detect a CATE of 50%, based on compliance rate



# Simulations (excluding efficiency gains) suggest that power to detect an impact on BP is also very low using test 5

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11



# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (1)

**Test parameters**

- Non-parametric test: Complier Average Causal Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

**Why?**

Imai, King and Nall (2009), also propose an instrumental variables estimator to estimate the Complier Average Causal Effect (CACE), i.e. the effect of the program on the farmers/households who actually received the treatment. This is a measure that takes into account the effect of non-compliance. It is slightly more complicated to implement – we provide an example of how to do this in the do-file associated to test 6.

The method assumes that: (i) there are no spill-over effects between individuals (a scenario which is unlikely in this case, but it might still be useful to get an estimate of CACE regardless); (ii) compliance is only affected by the random assignment of Kebeles to treatment and control (this is called the exclusion criterion); and (iii) there are no "defiers" (i.e there are no farmers/households that would only take the program if they were not given the opportunity AND refuse to do so if they were given the opportunity). The method does allow for non-compliance in both the treatment and control groups.

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (2)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

For a detailed explanation, see Imai , King and Nall (2010). The intuition: the selected impact estimator (CACE) is a weighted average of within pair differences, divided by a weighted average of pair-wise differences in take-up between treatment and control. The following slides build on estimates and notations from Test 5.

Step 1: Calculate the difference in the cluster-averaged take-up of treatment by pair (assume cluster 1 in each pair is the treated cluster)

- Calculate this as follows:

$$DC_p = \left(\frac{\sum_{i=1}^{n_{p1}} R_{p1i}}{n_{p1}} - \frac{\sum_{j=1}^{n_{p2}} R_{p2j}}{n_{p2}}\right)$$

where $DC_p$ is the difference in cluster average take-up of treatment between treatment and control Kebeles in each pair, and $R_{p1i}$ is a dummy that stands for the receipt of treatment or not of individual $i$, in Kebele 1, in pair $p$. $DC_p$ is basically a measure of how much more of the treatment, the treatment group received.

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (3)

**Test parameters**

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 2: Calculate weighted average difference of take-up between treatment and control

- Call C, the weighted average difference in the take-up of the treatment across pairs. It can be calculated easily in Stata using the following formula:

$$C = \frac{1}{N_{p1}+N_{p2}} \sum_{p=1}^{P} ( N_{p1} + N_{p2})\ DC_p$$

Step 3: Calculate the CACE estimator

- Now we can calculate the Cluster Average Complier Effect (CACE), which is our main estimator of interest. The estimator is very simple – it's just a division of the Intention-to-Treat effect, estimated by CATE (see test 5) and the difference in take-up between treatment and control:

$$CACE = \frac{CATE}{C}$$

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (4)

**Test parameters**

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

The most challenging part of this test is estimating the variance of the CACE estimator. To simplify, we break this down into several steps. See the do-file for simulations on test 6 for more insights.

Step 4: Calculate the variance of the CATE estimator as in test 5

- Follow each of the distinct calculations described in step 4 of test 5. We call this variance estimator: V

Step 5: Calculate the variance of C, the difference in take-up between treatment and control

- As described in test 5, first calculate the "normalized" weight of each pair (p), defined as:

$$w_p = \text{n} \times \frac{(N_{p1}+N_{p2})}{\sum_{p=1}^{P}(N_{p1}+N_{p2})}$$

where n is the total sample size and p is the pair number. We use this measure later on as well.

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (5)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 5 (continued)

- Next calculate the mean difference in take-up between treatment and control, using these normalized weights. Call this MC, such that:

$$\text{MC}= \frac{1}{n}\sum_{p=1}^{P} w_p \, DC_p$$

- Now calculate the following variable at the pair level, which is the square difference between the "observed" difference in take-up at the pair level and the average difference in take-up. We call it $SC_p$, such that:

$$SC_p = (w_p DC_p - \frac{n}{p} MC)^2$$

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (6)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 5 (continued)

- Finally, we can calculate the variance estimator of C, call it VC, such that:

$$VC=\frac{p}{(p-1)n^2}\sum_{p=1}^{P} SC_p$$

Step 6: Now calculate the covariance of CATE and C … last step before we estimate the variance of the CACE estimator

- We need to break down the calculation of the co-variance estimator into some sub-steps again.
- First, estimate what we call term 1, such that:

$$Term1_p = w_p{}^{*}Dp\text{-}\frac{n{*}M}{p}$$

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (7)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 6 (continued)

- Next, estimate term 2, such that:

$$Term2_p = w_p{}^{*}DCp\text{-}\frac{n*MC}{p}$$

- Now we can calculate the covariance between CATE and C, which is:

$$cov(CATE, C) = \frac{p}{(p-1)n^2}\sum_{p=1}^{P}(Term1_p \times Term2_p)$$

- This looks more complicated than it is. When broken down into small pieces it is not very difficult to compute in Stata, it's just a bit tedious.

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (8)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 7: Finally we can calculate the variance estimator for our estimator of the Cluster Average Complier Effect (CACE)

- This can be estimated as follows:

$$V_{CACE} = \frac{1}{C^4} \times [C^2V + CATE^2VC - 2CATE \times C \times cov(CATE, C)]$$

Step 8: Construct a one-sided confidence interval for CAC E

- Given that the distribution of impacts at the pair level is assumed to have a normal distribution (following the CLT), draws from this hypothetical distribution will follow a t-statistic with P-1 degrees of freedom, where P is the total number of pairs

- We therefore look up the critical value corresponding to a t-statistic with P-1 degrees of freedom and an $\alpha$ of 0.05 (where $\alpha$ is the level of statistical significance desired). You can obtain that very easily in Stata using the following commands: scalar t=invttail(P-1,0.05) and "display t" to see the result.

# Test 6: Neyman-based inference adapted to PMCRT, one-sided test of CACE estimator (9)

## Test parameters

- Non-parametric test: Cluster Average Complier Effect (CACE)
- Test statistic based on approach proposed by Imai, King, and Nall
- Takes into account both individual and group level data,
- References: Imai, King, Nall (2009)

## How?

Step 8 (continued)

- We construct a one-sided confidence interval of the CACE statistic using the critical value corresponding to the t-statistic with P-1 degrees of freedom, such that:

$$[MC\text{-}t\sqrt{V_{CACE}}\ ;\ \infty\ [$$

where t is the critical value corresponding to a t-distribution with P-1 degrees of freedom with an $\alpha$ of 0.05.

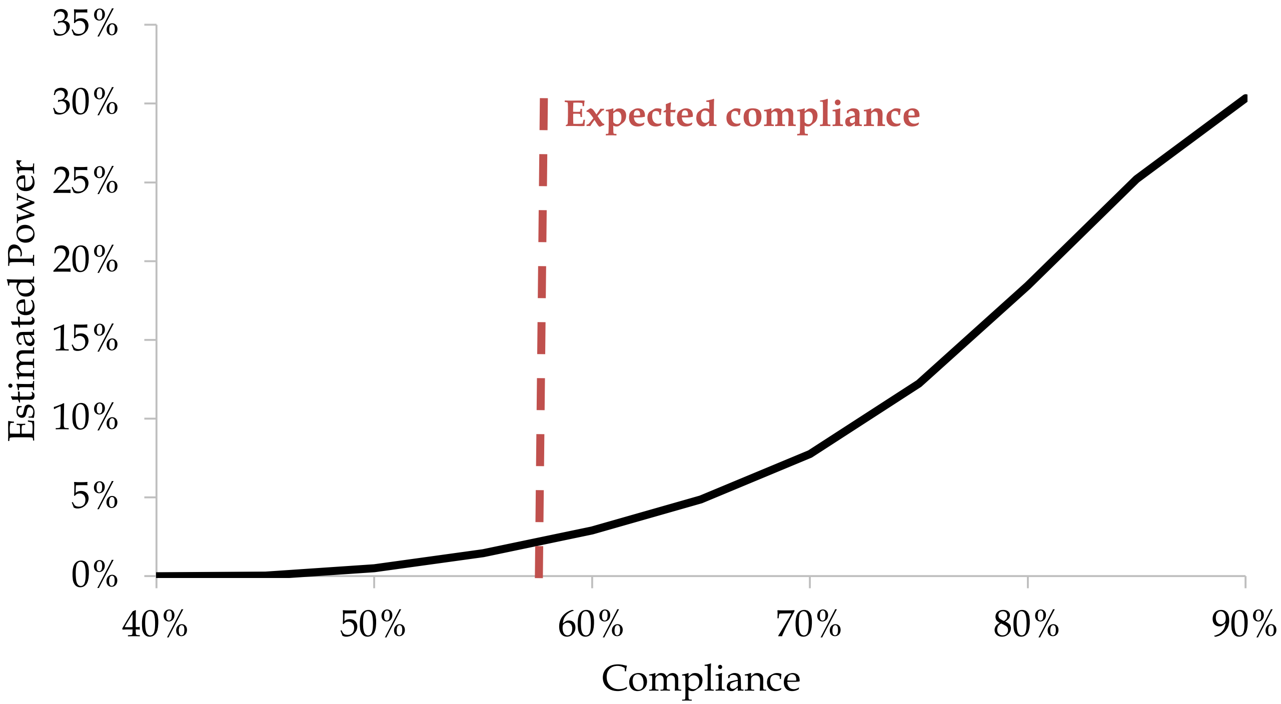
- If $MC$-t$\sqrt{V_{CACE}}$>0 then then the estimated Cluster Average Complier Effect (CACE) is statistically significant (one-sided).

# Simulations (excluding efficiency gains) suggest that the power to detect an impact on yields will be very low using Test 6

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

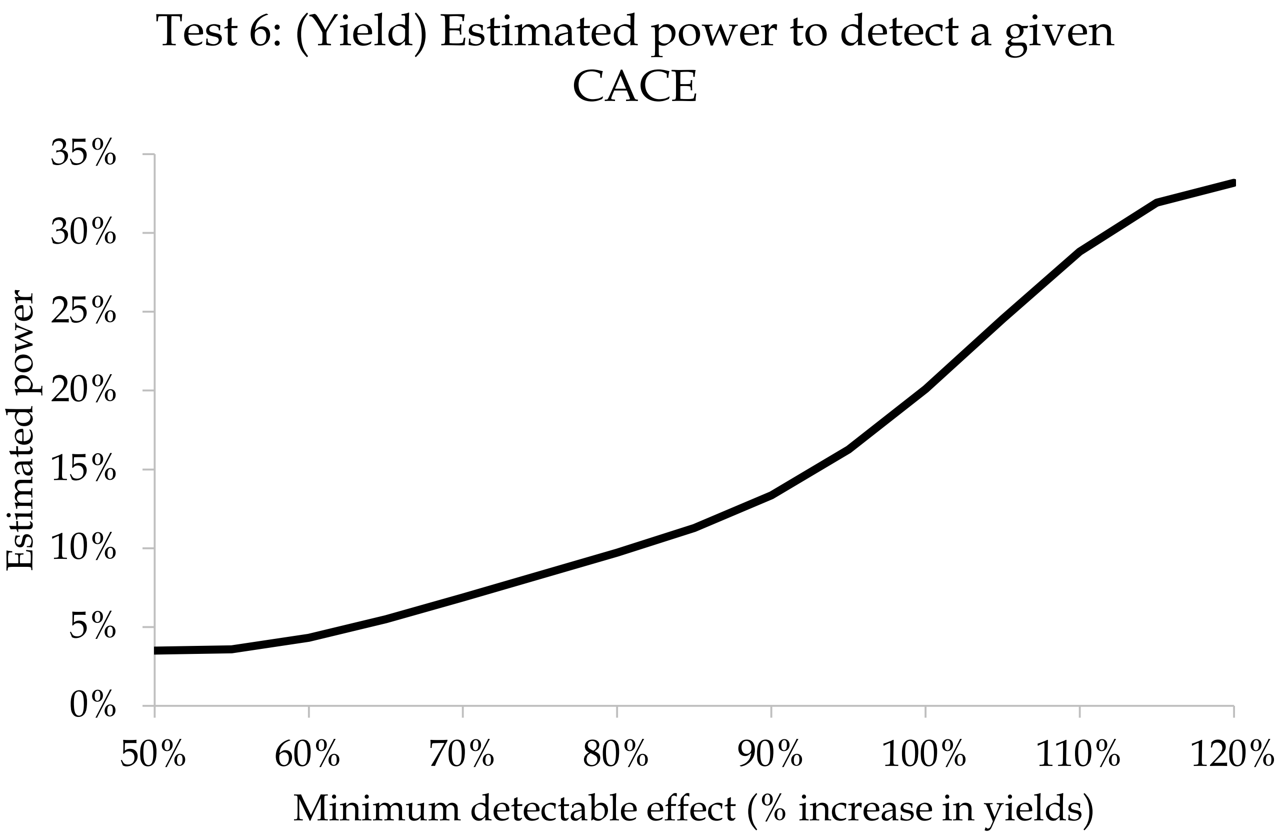
Test 6: Estimated power to detect an 50% CACE, based on compliance rate



# Simulations (excluding efficiency gains) suggest that even with much higher impact levels the power to detect an effect on yields is low

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): (varies)
- Compliance=58%
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

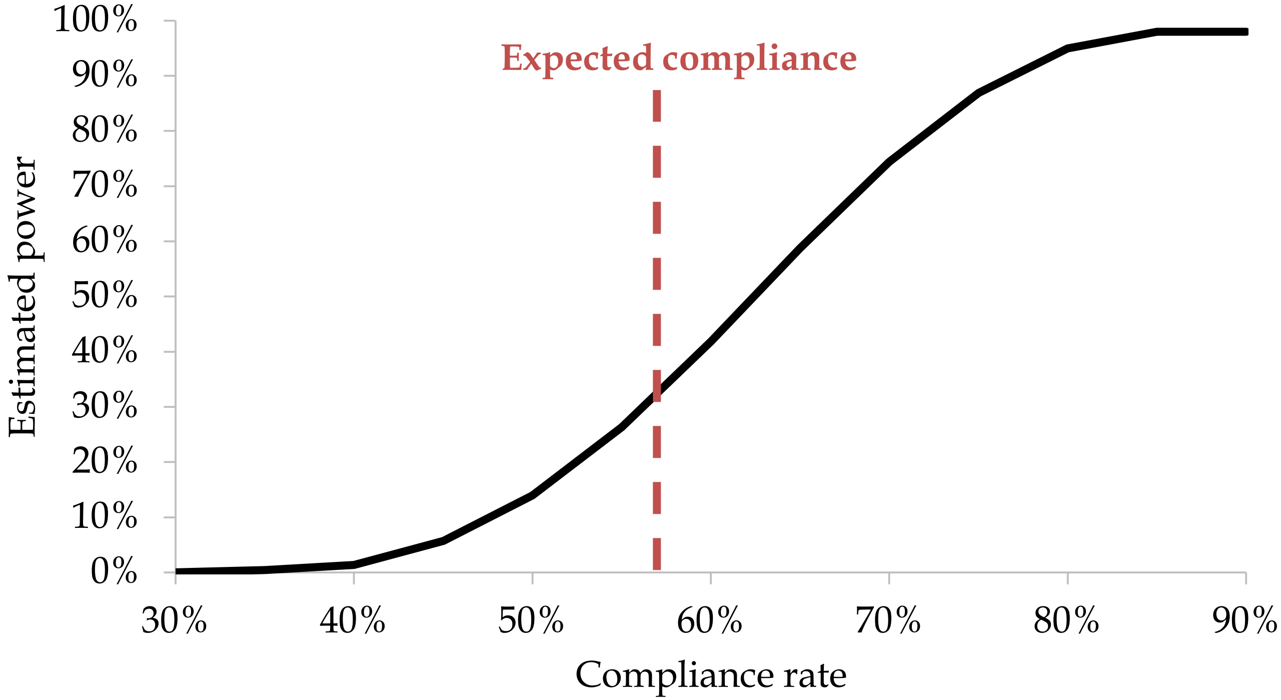


# Simulations (excluding efficiency gains) suggest that power to detect an impact on BP is also very low using test 5

**<u>Parameters for Monte Carlo Simulations of Statistical Power</u>**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=varies
- Pairs: 11

Test 6: (BP) Estimated power to detect a CACE effect, based on compliance rate



laterite TRANSFORMING RESEARCH IN AFRICA Rwanda Malawi Burundi

# Inference options and guidelines

## *Difference with a lagged dependent, using the wild cluster bootstrap*

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (1)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## Why?

In the case of few clusters, regressions - even when using cluster-robust standard errors - can lead to a large overestimation of the t-statistic (i.e. over-optimistic results about the statistical significance of the impact), and hence inaccurate inference about the impact of the program. One method that has proven quite successful in improving inference in the case of few clusters, is called **wild cluster-bootstrapping**. Adapted to cluster-standard errors by Cameron et al (2007), this method is based on the bootstrapping of the t-statistic (i.e. re-calculating the t-statistic over and over again), building on random changes to the error terms. This a bootstrapping method that provides "asymptotic refinement" … i.e. it reduces the error in the rejection probability of an asymptotic test.

To correct for the almost linear association between outcomes and baseline values – that lead the failure of the parallel trends assumption - we propose to use a lagged dependent variable assuming unconfoundedness given the lag (see Imbens et al, 2008). This relies on different assumptions than difference-in-difference.

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (2)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

For more details see Cameron et al (2007) as well as papers by MacKinnon and Webb between 2006-2014.

Step 1: Estimate the Difference estimator using cluster-robust standard errors and a lagged dependen
t

- The basic difference-in-difference model can be formally specified as follows:

$$Dif_{ikt} = \alpha + \beta_1 treat_k + \beta_2 Yield_{ik0} + \beta_3 X_{ikt} + u_{ikt}$$

where *i* stands for individual, *k* for the kebele, *t* for time, *yield* stands for yield levels, $Dif_{ikt}$ is the difference between baseline and endline values for individual *i*, *treatment* a dummy for whether an individual is in the treatment or control group, *X* are control variables (either at the individual or Kebele levels), $Yield_{ik0}$ is our lagged estimate of the dependent (assuming unconfoundedness) and $u_{ikt}$ is the error term, which will have a group component and an individual component.

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (3)

**Test parameters**

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Step 1: Continued …

- The coefficient $\beta_1$ of *treat* (the treatment indicator) is our estimator of the Average Treatment Effect.
- Calculating this, with cluster robust standard errors (i.e. the Cluster Robust Variance Estimator proposed by Liang and Zeger), can be done as follows in Stata:

$$reg\ dif\ treatment\ yield\_lag\ X, cluster(kebele)$$

where *kebele* is an identifier for each Kebele.

- Note that in the case of few clusters (which is the case here), the t-statistic obtained will tend to over-state the significance of the results obtained. This is because cluster robust standard errors are negatively biased.

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (4)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Step 2: Implement wild-cluster bootstrapping, by imposing the null hypothesis, and using one-sided tests

- We have provided a do-file that will run the wild cluster bootstrap method, with an imposed hypothesis, adapted to the the TNS case and including one-tailed testing– see the dofile for test 7 monte carlo simulations (this has been adapted from the wild cluster do-file provided by Douglas Miller in Cameron et al, 2008)
- The outcome of this method is not a revised t-statist or a confidence interval, but rather a revised p-value (one-tailed test) for the difference-in-difference estimator only
- This p-value basically tells us how likely the null hypothesis is. If based on the one-tailed test we find that $p<0.05$ then we can reject the null hypothesis that Dif=$\tau$
- We do not explain the wild cluster bootstrap in detail here, but the basic intuition, the basic steps, are as follows (see next slide):

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (5)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Step 2: (continued)

- Basic steps:
  1. Calculate potential outcomes and residuals if the null hypothesis is true, i.e. if Dif=$\tau$ (we impose the null)
  2. Then randomly multiply these residuals by 1 or -1, where randomization is done at the Kebele level (in some clusters individual residuals are multiplied by 1, in others by -1)
  3. Calculate a "wild outcome" indicator (or a "wild yield" measure), which is constructed using the residuals from the imposed regression
  4. Re-run the difference-in-difference equation using the "wild yield"measure, instead of the actual yield measure, and estimate and store the cluster-robust t-statistic
  5. Repeat 2-4 many times (recommended>500) – that's called bootstrapping. Let's call B the number of bootstraps.
  6. Count the number of times the t-statistic obtained is larger than the actual t-statsic. Divide that by B and you obtain the desired p-value for the test.

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (6)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Step 3: Obtain confidence interval, by looping over different values of $\tau$

- To obtain a confidence interval for Dif, one can loop over different values of $\tau$
- Note that this can be a very computationally intensive process (i.e. it will take quite a long time to run through all the loops … the monte carlo simulations we run take about 30-45 minutes to complete)
- For more details on how to construct confidence intervals using this method see MacKinnon (October, 2014)

Notes to keep in mind

- The proposed equation can be modified to include random and fixed effects at the Kebele/Pair level. Wild cluster bootstrapping does not work in this case. However, OLS regression preserves the coefficient estimates of other fixed/random effect models; the bootstrapping accounts for clustering (see Wild 2013)

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (7)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Notes to keep in mind (continued)

- Here we ignore the PMCRT design. This is not optimal, but is justifiable in the sense that MCPR was selected to obtain a greater balance between the treatment and control groups. The assignment to treatment and control still remains random.
- One way to account for pairs, would be to include pair-level controls in the regression, to include the same covariates in the regression that were used for pair matching (ie. number of coffee farmers, altitude, location), but also potential to run the wild bootstrapping at the pair level instead of the Kebele level
- Note that methods to apply cluster-robust bootstrapping methods to instrumental variable (IV) regressions in the case of few clusters are currently being developed (see Finlay discussion paper, 2014). We would recommend checking at the endline whether new tools have been developed for IV regressions. This would enable reliable testing of the significance of the Local Average Treatment Effect – i.e. the effect of the treatment on the treated.

# Test 7: Difference with a lagged dependent, using wild cluster-bootstrapping, imposing the null, one-tailed test (8)

## Test parameters

- Null: Difference = $\tau$
- Test statistic based on approach proposed by Cameron et al (2007)
- Takes into account both individual and group level data,
- References: Cameron et al (2007), Cameron (2014), MacKinnon 2007-2014), Webb (2013), Imbens and Wooldridge (2008)

## How?

Notes to keep in mind (continued)

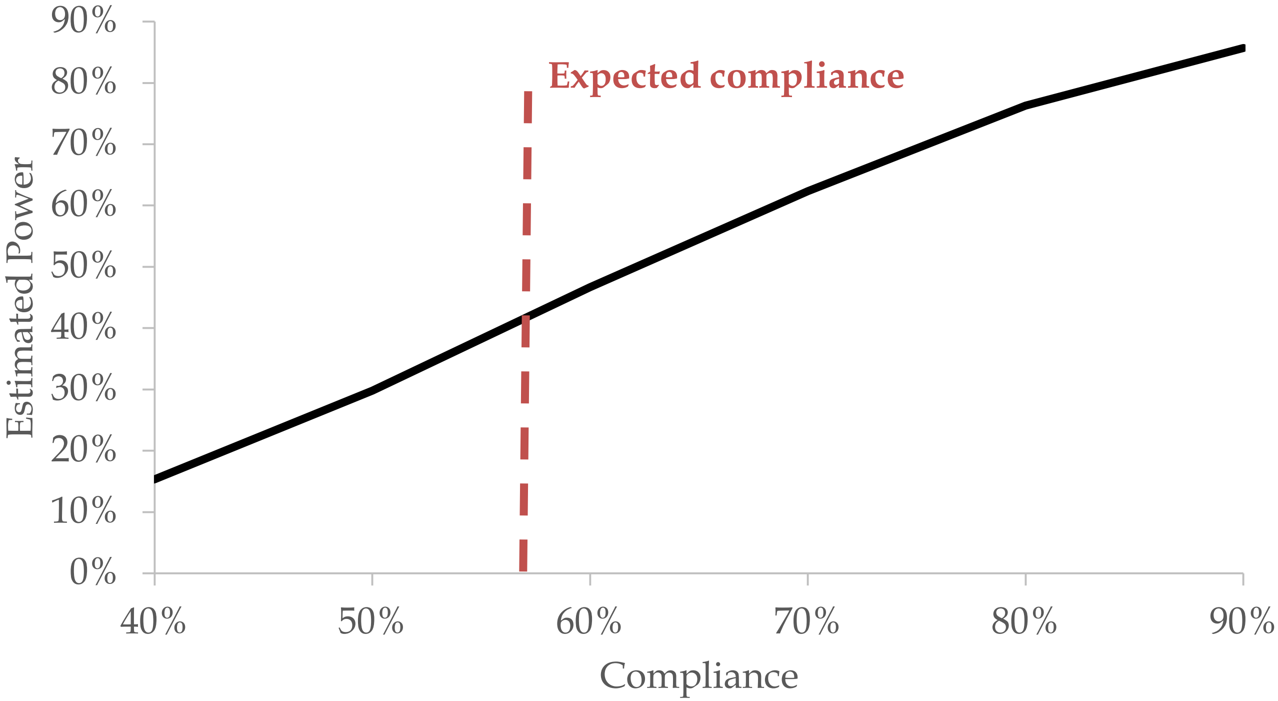
- Also note, that other techniques – e.g. Feasible GLS combined with standard error adjustments – have been proposed for the case of few clusters. We do not introduce them here as we do not think they will lead to significant improvements in inference. For more, see a summary of available options in the case of few clusters in Cameron (2015).
- **Lastly and this is important, how do we check that the assumption of unconfoundedness using the lag is valid??** One simple way of checking this (though it does not fully test the assumption) is to see whether you can detect a treatment effect using the lagged outcome as the dependent. We know for a fact that the lagged outcome variable was not affected by the treatment. If we cannot reject the null that the treatment effect is 0, then it is more likely that the unconfoundedness assumption holds. To learn about this, see: Imbens and Wooldridge (2008).

# Simulations (excluding efficiency gains) suggest that the power to detect an impact on yields will be low using Test 7

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

Test 7: (Yield) Estimated power to detect an effect, based on compliance levels



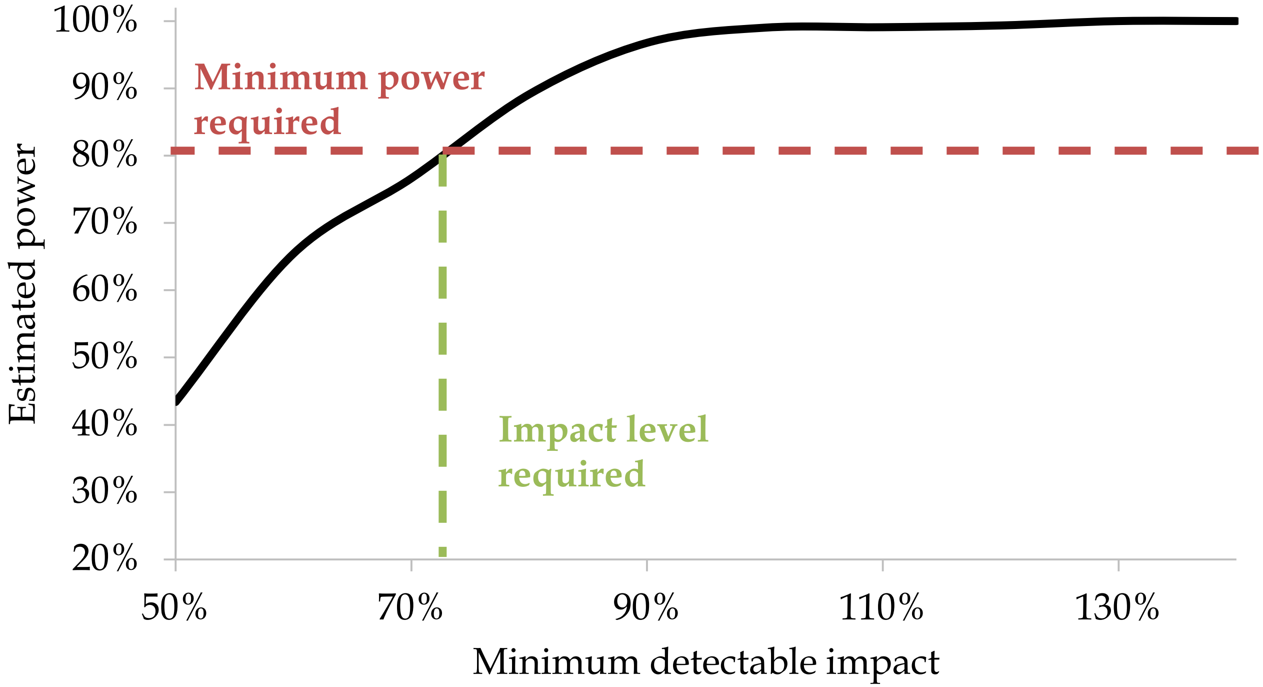
- We speculate that with a compliance rate of about 85% this test will have sufficient power (80%) to detect an effect on yields, if the expected increase is 50%.

# Simulations (excluding efficiency gains) suggest that Test 7 should have enough power to detect an impact on yield of minimum 73%

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

Test 7: (Yield) Estimated power to detect an effect, based on different impact levels

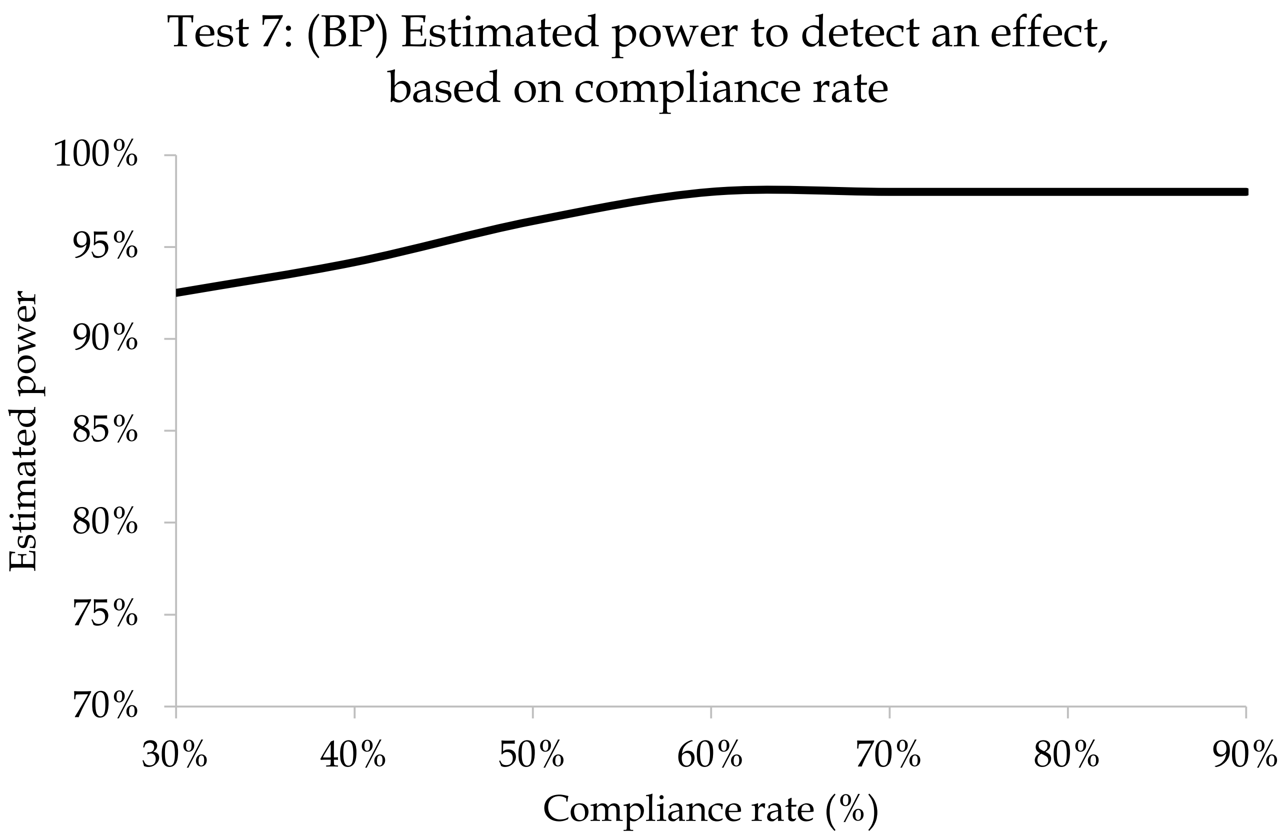


- We speculate that with an yield impact of about 73%, Test 7 will provide sufficient statistical power to detect and effect in the yield sample.

# Simulations (excluding efficiency gains) suggest that Test 7 should have enough power for the BP sample, regardless of compliance

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=varies
- Pairs: 11



- Regardless of the compliance level, we expect power in the BP sample to be very high with Test 7

# A simulation using real baseline and endline data from the 2013 Cohort in Ethiopia (see modelling slides)

**Parameters for Simulation**

- As predictors for the synthetic match, we select all BPs and total BP (the outcome of interest)
- In the modelling slides, we showed that a DiD estimator would have resulted in an average impact estimate of 2.3 BPs
- A revised approximation, using modelling, yielded an impact of 1.42 BPs

**Estimated impact = 1.54*****

The result obtained is statistically significant at the 1% level. By looping we estimate the one-sided lower bound of the effect at 1.3, and the two-sided confidence interval as [1.15-1.9]. Results are at the 95% confidence interval.

- We estimate that in Cohort 2013 treatment led to 1.54 increase in the average number of BPs adopted. This is quite close to our corrected impact estimate of 1.42 BPs, derived in the modelling section, and a significant improvement on the non-lagged estimate of 2.3.
- A test on the unconfoundedness assumption however fails in this case as we find a treatment effect on the lagged variables, which shouldn't be the case as there was no treatment effect on the lags. This means that randomization in this test failed and that the control and treatment groups are significantly different on baseline BP levels. This might change if additional covariates are included in the regression.

# Inference options and guidelines

## *An ex-post method: Synthetic controls combined with randomization inference*

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (1)

## Test parameters

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## Why?

Synthetic Control methods, first developed by Abadie and Gardeazabal (2003), offer an interesting alternative inference strategy for both the yield and BP samples. Intuitively, synthetic controls is based on the idea that it is possible to create a good counterfactual for a treatment region of interest using a weighted linear combination of control regions. Comparing one treated Kebele to its matched control Kebele (in the PMCRT case) might not be a very good idea. Despite being the closest possible match, there might still be very large differences between two Kebeles in a given pair. Synthetic control methods provide a fix for this, by focusing on a combination of multiple control regions that provide the best possible match at the baseline for a given treatment Kebele.

Here we propose to run synthetic controls methods on each treated Kebele to obtain Kebele-level average treatment estimates. We propose to conduct inference using randomization inference (see test 1) or by assuming normality. See do-file for Monte-Carlo simulations as an example of how to do this.

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (2)

**Test parameters**

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## Why synthetic controls would work well for the BP and yield samples?

A major issue in both samples is that we suspect the impact (as was the case in Rwanda) will be a function of baseline levels. In particular, the higher the baseline yield/BP, the lower the absolute increase post treatment. In Rwanda, farmers with the highest baseline yields/BPs in the treatment group even experienced a decrease in performance post-treatment. We interpret this outcome as being a combination of two dynamics: (i) the impact of the program, which is approximately constant for all farmers; and (ii) a negative relationship between baseline and endline outcomes, that holds regardless of whether a farmer was treated or not. In other words, for a given baseline, the observed treatment effect will be similar for all farmers (or Kebeles); for farmers (or Kebeles) with different baseline outcomes, the observed impact will be different. Synthetic controls, if successful, makes it possible to create a synthetic counter-factual with a very similar baseline to a given Kebele of interest (provided this Kebele is not an outlier). This should lead to better impact estimates at the Kebele level.

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (3)

## Test parameters

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

For more details on how to run synthetic controls, please consult Abadie, Hainmueller and Diamond's papers (2010 and 2014). The intuition: we create a synthetic control group by minimizing the difference between the treatment group and a linear combination of potential control regions on a set of covariates, using an algorithm. The resulting synthetic control region is the best possible match, given the selected parameters. It's easier than it sounds, thanks to the "synth" package developed by the authors. To download it simply type: "ssc install synth" in Stata.

Step 1: For each Kebele, run the synthetic control algorithm

- First organize the data into time series, by Kebele and time. To do this collapse yield/BP and covariate data by Kebele and time.
- Sort the data by time, treatment and Kebele and assign an id to each Kebele, such that Kebele no {1 .. P} are treatment Kebeles, and Kebeles {P+1 .. 2P} are control Kebeles, where P is the total number of pairs

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (4)

**Test parameters**

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

Step 1: (continued)

- Tell Stata that your data is organized as time series data, using: "*tset id time*", where id is the Kebele id and time is either baseline or endline (or other time periods in between).
- Select covariates/predictors on which the matching will be conducted. Limit them to 5-8 at most. Covariates/predictors should be important determinants of average yield or BP levels at the Kebele level. We recommend including population and altitude, as these were two key parameters used in the matching algorithm. Yield and BP levels will also have to be included.
- Now run the synthetic control algorithm on Kebele 1, using the following syntax:

*synth yield yield bp predictor1-predictor5, trunit(1) trperiod(1) resultsperiod(0(1)1) counit(a(1)b) figure*

where *predictor1-5* are the selected covariates for matching, *trunit* is the treated unit (Kebele 1),

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (5)

## Test parameters

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

Step 1: (continued)

*trperiod* the period in which the treatment occurs (here baseline=0 and endline=1), *resultsperiod(0(1)1)* tells Stata that we are interested in getting results for both the baseline and the endline, *counit(a(1)b)* tells stata that the control Kebele have ids that go from a to b.

- See the the monte-carlo simulation do-file or the help-file of the "synth" package for more information on how to do this.

Step 2: Obtain the Treatment Effect estimates for each Kebele

- We recommend using the following estimator, which is essentially the difference-in-difference between the endline and baseline results for the treated and synthetic control Kebele:

$$\tau=(Y_{T1}\text{-}Y_{C1})\text{-}(Y_{T0}\text{-}Y_{C0})$$

- This is straightforward to implement using the synth package, by looking at the post-estimation stored matrices (see do-file)

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (6)

**Test parameters**

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

Step 3: Conduct inference using randomization inference techniques

- Once impact estimates have been obtained for each Kebele, we can repeat Test 1 using randomization inference.
- Inference will therefore not be conducted at the individual Kebele level, but rather across Kebeles. What we'll have achieved using the synthetic controls method is to fine-tune our Kebele-level estimates. We're not just taking the difference within pairs, but rather trying o create the most optimal synthetic control group for each Kebele.
- To test for the scenario of no-effect, and without using further controls, simply use the Wilcoxon signed rank statistic. This can be estimated easily in stata using the "signtest" command:

$$signtest\ \tau = 0$$

- Look for the one-sided p-statistic associated to the test statistic, and reject the null hypothesis of no effect (or of a given effect $\tau$) if the p-value is smaller than 0.05.

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (7)

**Test parameters**

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

Step 4: Alternatively, assume estimates from the synthetic controls method follow a normal distribution, and conduct inference using a simple t-test

- Assuming the synthetic control method provides a very good match for each Kebele, and that the treatment effect on average is constant for a given level of yield at the baseline, then we could assume that the Kebele-level effects measured using synthetic controls follow a i.i.d normal distribution.
- To test whether the Synthetic Control method provides valid controls, one needs to look at how well the Treatment and Synthetic Control groups are balance on certain covariates. This is easy to check using the "synth" package (see the stored results)
- In this case inference can be conducted using a simple t-statistic with P-1 degrees of freedom (where P is the number of pairs):

$$\text{ttest } \tau = 0$$

# Test 8: Estimating the average treatment effect per Kebele using synthetic controls combined with randomization inference (8)

## Test parameters

- Average Treatment Effect (ATE) per Kebele and Overall Average Treatment Effect
- Tests are conducted using randomization inference
- Inference is conducted at the group level
- References: Abadie (2003), Abadie (2010), Abadie (2014)

## How?

Test 4 (continued …)

- Reject the null hypothesis of no effect if the corresponding one-sided p-value is smaller than 0.05.
- Assuming normality increases the power of this test substantially (if indeed the data is normally distributed)

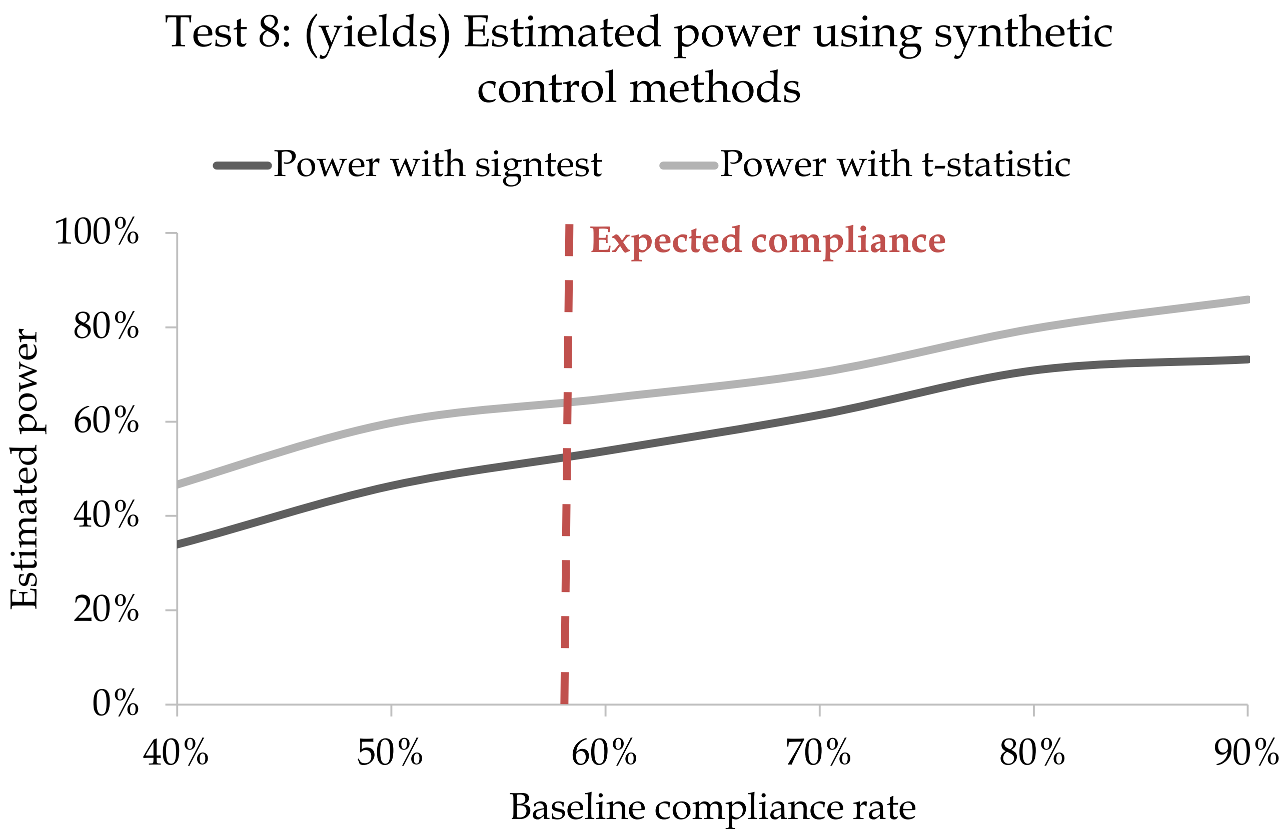
Notes

- Note that it is also possible to conduct inference at the Kebele level. Inference will be different and will involve conducting placebo synthetic control tests on all the control Kebeles to obtain a distribution. For relevant inference techniques see Abadie, Hainmueller and Diamond (2010 and 2014).

# Monte Carlo simulations (excluding efficiency gains) suggest that the power to detect an impact on yields will be relatively low using Test 8

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

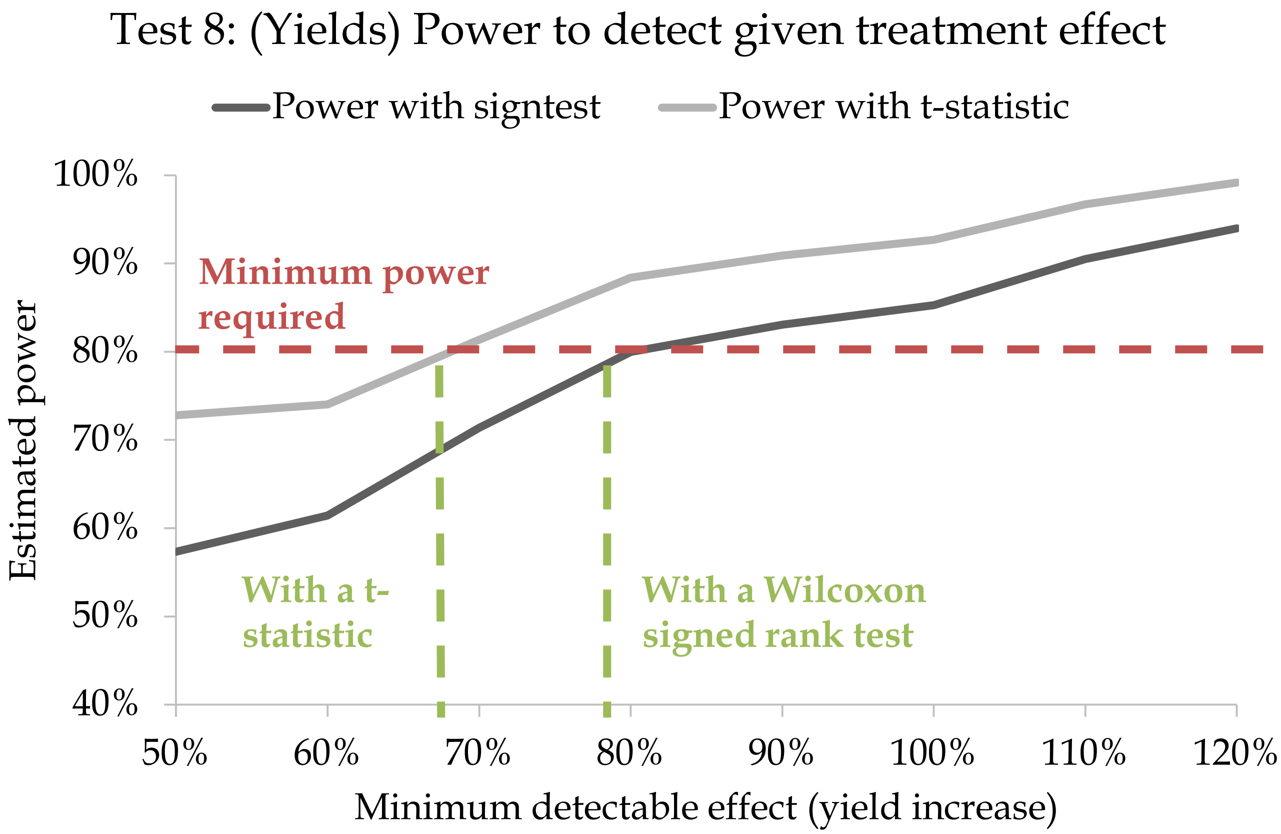


- We speculate that with a compliance rate of about 90% the t-test will have sufficient power (80%) to detect a 50% increase in yield levels, which is a large minimum detectable effect.

# Simulations (excluding efficiency gains) suggest that Test 8 should have enough power to detect a minimum effect of 67-80%

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Sample size: 650
- Mean=1160
- Sd (cluster)=587
- Delta (yield): 580 (50% increase)
- Compliance=varies
- Attrition: 10% (T) and 15%(C)
- Pairs: 11

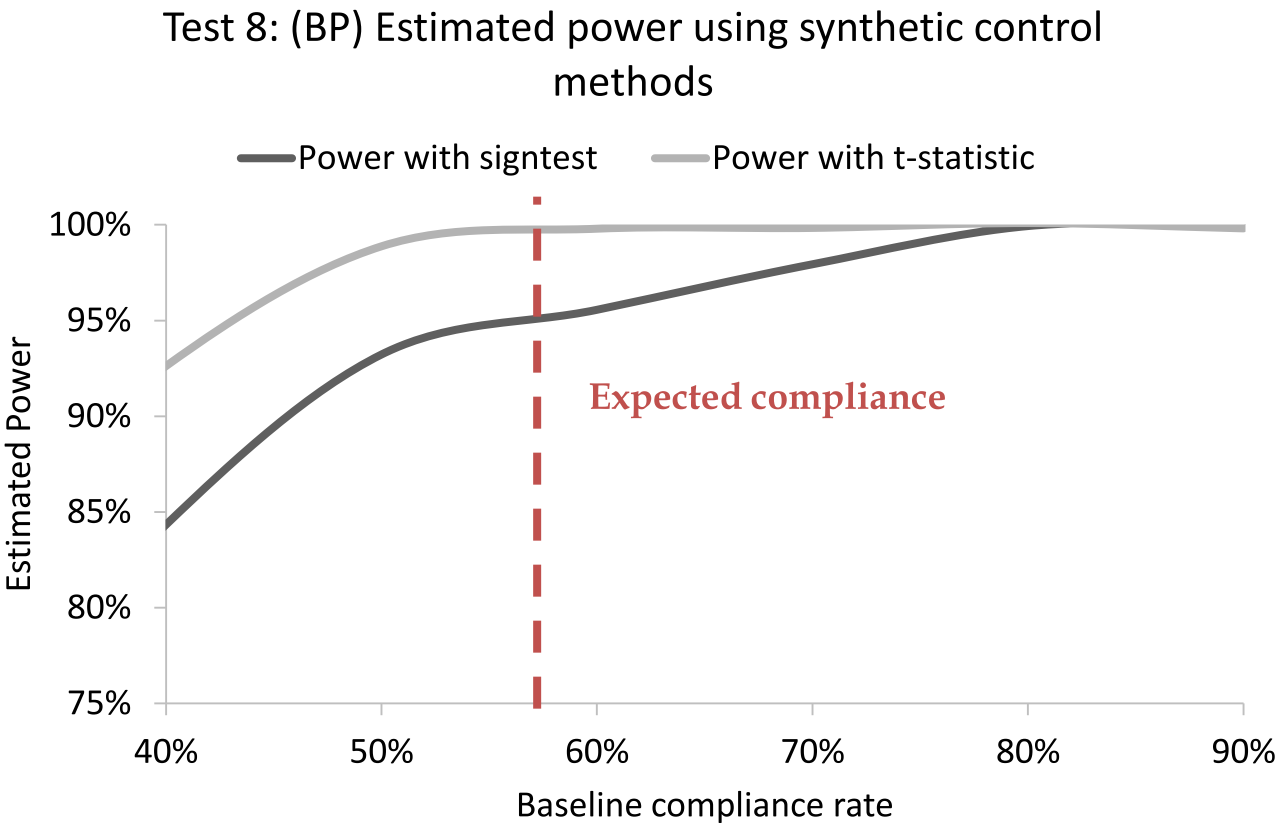


- We speculate that the synthetic control test should give us enough power to detect a minimum of 67-80% increase in yields, depending on the test and the assumptions.

# Monte Carlo simulations (excluding efficiency gains) suggest that the power to detect an impact on BPs will be very high using Test 8

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Total sample:650
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): 2.89 (50% adopt > 5 BPs)
- Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11



www.laterite-africa.com

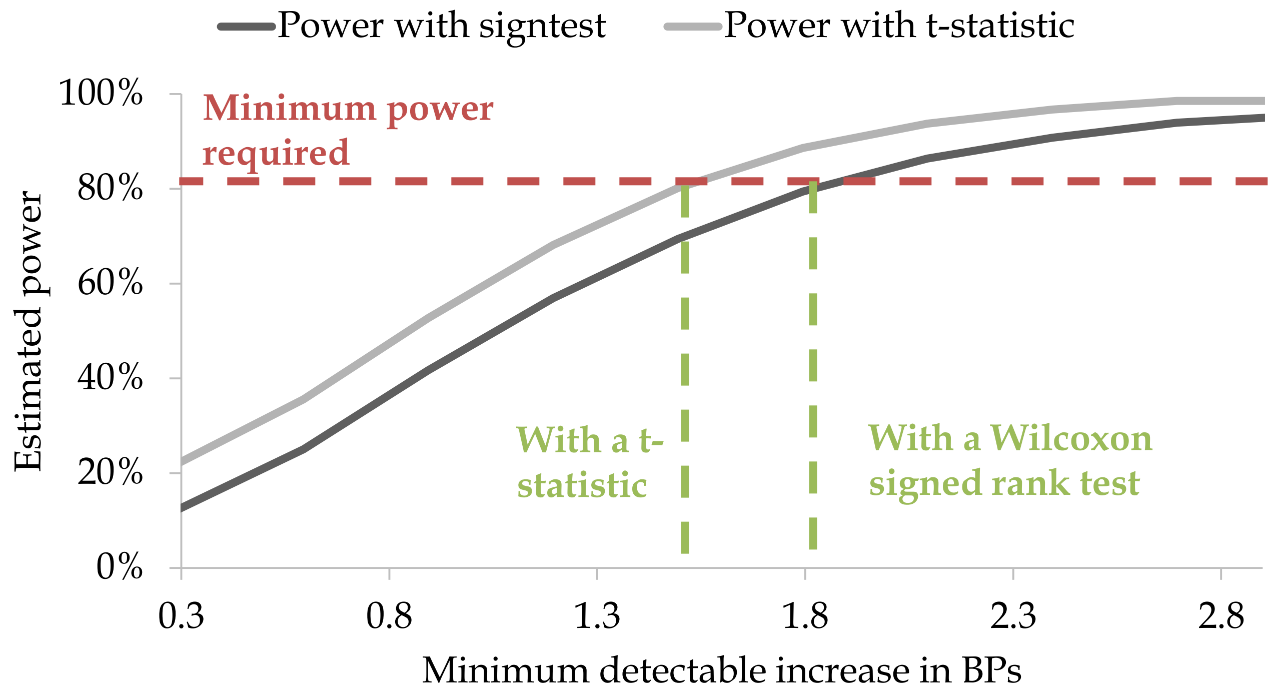
laterite TRANSFORMING RESEARCH IN AFRICA Rwanda | Malawi | Burundi

# Monte Carlo simulations (excluding efficiency gains) suggest that the minimum detectable increase using Test 8 will be about 1.5-1.9 BPs

**Parameters for Monte Carlo Simulations of Statistical Power**

- We assume pairing is done badly (as good as random)
- We assume no efficiency gains from control on covariates
- Total sample: 650
- Mean=1160
- Mean=3.11
- Sd (cluster)=0.79
- Delta (yield): varies Compliance=58%
- Attrition=10%(T) and 15% (C)
- Pairs: 11

Test 8: (BP) Minimum detectable increase in BPs and associated power



- We speculate that the synthetic control test should give us enough power to detect a 1.5 to 1.9 increase in BPs, depending on which test statistic we use and what assumptions we make

# A simulation using real baseline and endline data from the 2013 Cohort in Ethiopia (see modelling slides)

**Parameters for Simulation**

- As predictors for the synthetic match, we select all BPs and total BP (the outcome of interest)
- In the modelling slides, we showed that a DiD estimator would have resulted in an average impact estimate of 2.3 BPs
- A revised approximation, using modelling, yielded an impact of 1.42 BPs

**Estimated impact = 0.82*** (weighted = 1.46 BPs)**

The result obtained is statistically significant at the 1% level using both the Wilcoxon sign rank test and the t-test

- We estimate that in Cohort 2013 treatment led to a 0.82 BP increase on average (this is a simple average taken across Kebeles). Weighted, based on population values, this corresponds to a 1.46 BP increase in best practice adoption on average. This is remarkably close to our corrected impact estimate of 1.42 BPs, derived in the modelling section, and the 1.56 BPs estimate obtained using difference with lagged variables. Again this shows that the DiD model is far off.
- The similarity between the estimates based on synthetic controls and the corrected estimates based on our model, suggest that the model could be a good proxy for the behavior of the BP sample.

# Inference options and guidelines

## *Conclusions on the core analysis section*

# Summary of the estimated power of these tests to detect an effect in the yield and BP samples? Remember that no efficiency gains assumed!

**Projected power in proposed tests (note that these are based on different estimators!)**

| Method | Null hypothesis | Onus is placed on … | Yield sample (one-sided power) | BP sample (one sided power) |
|---|---|---|---|---|
| **Test 1 – Fisher group** | $\tau = 0$ | Group | 6% | 35% |
| **Test 2 – Fisher individual** | $\tau = 0$ | Individual | 47% | 59% |
| **Test 3 – Fisher non-compliance** | $\beta = 0$ | Individual | 47% | 59% |
| **Test 4 – Fisher quantile** | N/A | Individual | N/A | N/A |
| **Test 5 – Neyman CATE** | $CATE = 0$ | Group | 4% | 20% |
| **Test 6 – Neyman CACE** | $CACE = 0$ | Group | 3% | 35% |
| **Test 7 – Difference (lag)** | $Treatment = 0$ | Individual | 43% | 95% |
| **Test 8 – Synthetic** | $\tau = 0$ | Group | 50% | 93% |

# What conclusions can we draw from these tests? (1)

1. **These tests confirm that the BP sample systematically has higher power than the yield sample given the parameters we want to test**

2. **For the BP sample, the two most promising tests seem to be Test 7 (the difference with lags and wild cluster bootstrapping) and Test 8 (synthetic controls)**

- ❑ This provides TechnoServe with clear options: (i) one where individual data is taken into account in a regression (Test 7); (ii) the second where inference is conducted at the group level (Test 8)
- ❑ Both Tests 7 and 8 adapt well to the case of multiple time periods. Both take into account the lagged dependent variable, which seems to have a large impact on power.
- ❑ If TNS decides to implement more than 2 time periods for the BP evaluation, Feasible GLS estimators assuming auto-regressive processes might further increase power (see Hansen, 2007)

3. **For the yield sample, the highest chances of success come with tests 2-4, 7 and 8, which record a power of 40-50%, assuming no efficiency gains.**

- ❑ While low, substantial efficiency gains are possible through pair-matched clusters (if the matching ultimately proves to be successful) and covariate controls
- ❑ If the "additive effect" of the Treatment on the yield sample leads to an increase in yields of about 70-80%, then even without efficiency gains these tests should provide sufficient statistical power to detect an effect

# What conclusions can we draw from these tests? (3)

4. **In general, including the lagged dependent variable as a covariate increases power substantially**

❑ The lag is by design taken into account in the Synthetic Controls method (Test 8), but can also be implement in the Fisher tests, in particular 2-4. We expect substantial power gains from controlling on the lagged variable. This is because of the structure of the data and the negative association between endline and baseline results, regardless of whether a household/individual received the treatment or not.

5. **In general, and not surprisingly, tests that make more use of individual data (e.g. Tests 2-4, Test 7) perform better than tests that bring everything back to the group level**

❑ The exception to the rule, is the synthetic controls test which gains power by creating the best possible synthetic match to a given Treatment Kebele

6. **Tests using BP data from Cohort 2013 in Ethiopia also confirm that DiD estimators can lead to a very large over (or under) – estimations of impact**

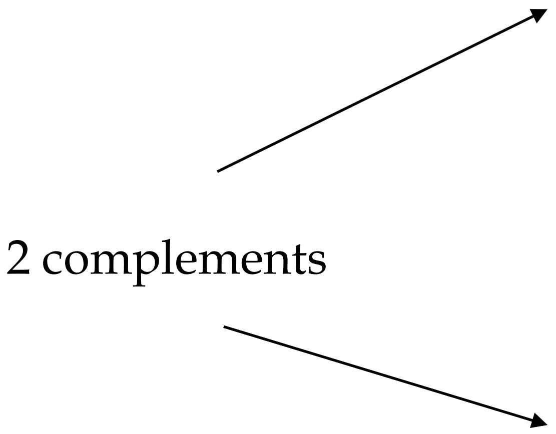
❑ DiD estimators are unreliable with coffee data (based on data from both Rwanda and Ethiopia), because the assumption of parallel trends does not hold

# Inference options and guidelines

*What alternative analysis options does TNS have if all of this fails?*

# We propose two sets of alternatives – or complements – to the core analysis: (i) instrumental variables; and (ii) ideas to build a good story

**In addition to the core analysis we recommend that TNS explore the following analytic options:**



2 complements

**Quasi-experimental methods using instruments**. We recommend one instrument in particular: "distance from the optimal population", which takes into account the difference between potential demand for training (the total number of coffee farmers in a Kebele) and the potential offer (number of Farmer Trainers in a Kebele times how many trainees they can take on). This difference, will impact both compliance rates and group sizes at the Kebele level, thereby creating exogenous variation in the outcome variable.

**Building a good story**. Regardless of how well the evaluation goes, the key to convincing your audience of the program's impact is to build a convincing story, combining the monitoring data TNS regularly collects (e.g. attendance) and the data collected during the evaluation phase. We propose some hypotheses than can be tested to form the basis of this story.

# Instrumental variables: Instrumenting or controlling for distance from the optimal population (1)

**What do we mean by distance from the optimal population?**

Each farmer trainer can train up to 12 groups of farmers, with a maximum of about 30 households per group. That means each farmer can train up to 360 coffee households. If one farmer trainer is assigned to a Kebele where there are 300 households only, then all households are given the opportunity to participate in the training. Average class sizes will be low, which we have shown in the Rwanda case is associated with better outcomes (in particular higher attendance rates). While this is great for the farmers, TNS would be operating below capacity in that Kebele (83%). If on the other hand there are 450 farmers in that Kebele, then not all farmers would be able to participate in the program and class sizes would have been larger. 450 farmers would also not warrant sending a second Farmer Trainer, because he/she would be operating at 25% capacity. In this Kebele, compliance will be lower and the quality of the training potentially lower because of larger class sizes. A positive distance from the optimal population (360-300)=60 is good for compliance and performance, a negative distance (360-450=-90), will lead to lower compliance rates and potentially also performance. We assume that where a Kebele stands in terms of optimal distance only affects the outcome variables of interest through it's effect on compliance levels and class sizes. It is therefore a good candidate for an instrument.

Formally this is how we define distance from the optimal population at the Kebele level:

$$\boldsymbol{Optimal\ distance = \#Trainers * 360 - \#Coffee\ households}$$

# Instrumental variables: IV in the case of few clusters

❑ We would recommend running this instrument alongside the random allocation to the Treatment or Control groups at the Kebele level. So we would have two instruments – belonging to the treatment group or not, and distance from the optimal population. Both impact the outcome variable by influencing the probability that an individual farmer ends-up receiving the treatment. The optimal population instrument also impacts the outcome variable through its effect average farmer group sizes.

❑ We would recommend running this IV approach in combination with a form of the wild-cluster bootstrap, adapted to the case of instrumental variables, or other methods that correct standard errors in the case of few clusters. Methods for doing this are still in the process of being developed, see for example Finlay's discussion paper, 2014.

# Building a convincing story … hypothesis that can be tested (1)

**What would we expect if the program did have an impact on yields and BP adoption?**

**Hypothesis 1**: We would expect to find significant differences between compliers and non-compliers in the Treatment group

- Impact is detected by comparing average values of the Treatment and Control groups
- However within Treatment Kebeles, we should also observe a difference between "treated" farmers and "non-treated" farmers … in other words compliers and non-compliers

**Hypothesis 2**: We would expect to find evidence of spill-over effects, within treated Kebeles

- Evidence of spill-over effects can be found by comparing "non-compliers" in the Treatment group to farmers in the Control group
- The best option, would to do this comparison using matching techniques

**Hypothesis 3**: We would expect to find an association between individual attendance rates and the outcome variables (BP adoption and yield)

- We can test this be simply running regressions using yield or BP as the dependant/outcome variable of interest and attendance as an explanatory variable for farmers in the Treatment group
- This statement should be true, albeit with diminishing returns

# Building a convincing story … hypothesis that can be tested (2)

**What would we expect if the program did have an impact on yields and BP adoption?**

**Hypothesis 4**: We would expect to find a clear association between BP levels and yield levels

- If BP adoption leads to higher yields, then endline results should show a clear association between the two
- For BP and yield data to be comparable, data on BP and yields needs to be collected from the very same plot

**Hypothesis 5**: We would expect that farmers that attended a specific session to perform better on the corresponding best practice, than farmers that did not attend

- This can be tested topic by topic, comparing "treated" farmers that attended the corresponding session and farmers that didn't
- Requires that data be recorded not only attendane, but also the corresponding session and BP taught

# Inference options and guidelines

*Risks to the validity of this experiment & ethical considerations*

# The main threat to the internal validity of the evaluation

**Potential threats to the internal validity:**

- ❑ ***The hulling station program***. The main threat to the internal validity of the experiment is the parallel program run by TechnoServe focusing on the hulling stations, which is implemented in all hulling stations in a given Woreda. Following discussions with the TechnoServe team, we believe that the risk of spill-overs from the hulling station program through to yield and BP levels is very low:
  - ✓ The hulling station program focuses on the sustainable management of hulling stations and is unlikely to have a direct effect on coffee prices, yields or best practice adoption
  - ✓ Farmers are blind to which hulling station their coffee is taken to … coffee is collected at the Kebele level by "Akarabis", traders who buy the coffee from farmers, and/or gathered in "collection centers", before the coffee is forwarded on to hulling stations for processing
  - ✓ In all likelihood, if there are spill-over effects, these effects should affect the treatment and control groups in a similar way

  We nevertheless recommend controlling for the distance from a farmer's plot, to the closest hulling station and collection center.

- ❑ There are also external risks that we cannot control: e.g. a weather shock on one Woreda in particular, or alternative programs implemented in one of the treatment Woredas or Kebeles. To control for potential external events it is important to collect Woreda and Kebele level information.

# Ethical concerns and considerations

- ❑ ***Random assignment to the treatment and control groups excludes a certain number of Kebeles from receiving the treatment for a period of 2-4 years.*** Within each pair, by randomly assigning one Kebele to the Treatment group and one to the Control, we are in effect preventing one of the two Kebeles from getting the Treatment for a period of several years, the time that the evaluation finishes. By having a control Kebele for each treatment Kebele, we would be preventing an estimated 14,000 farmers from receiving the treatment over the next few years, which is a very large number.

- ❑ ***This would have been the case regardless, given TNS targets.*** Regardless of how assignment to the Treatment group would have been conducted, TNS only has sufficient resources to reach a total of about 14,000 farmers in Lekempti over the next four years out of a total of about 150,000 potential coffee farmers across target Woredas in Lekempti, including Nejo, Lalo Asabi, Boji Dermaji, Haru, and Gulisso. The random assignment therefore doesn't prevent additional farmers from receiving the treatment.

- ❑ ***Should TNS get increased resources to expand the program in Lekempti over the next few years, this equation would change.*** TNS would have to make a call on whether or not to keep most Kebeles in the control group, in particular for the estimation of the treatment effect on yields. This is because: a) the yield effect can only be detected after 4 years (as farmers first have to rejuvenate their trees first) as opposed to BP adoption effect which can be estimated at the end of the 2 year intervention; and b) because statistical power in the yield sample is low.

# How to improve our chances of success

# How to improve our chances of success

*Some initial thoughts*

# There are a number ways to positively influence the power of these experiments and the evaluation overall (1)

Option 1: Select treatment farmers in "saturated" villages within Kebele's, rather than at Kebele level

- Given that registration into the program is done village-by-village within each Kebele, one idea to increase compliance would be to sample farmers from villages that we know are "saturated", where the vast majority of farmers have registered.
- The alternative, sampling at the Kebele, level might lead to significantly lower power

Option 2: Nudge farmers in the sample to attend

- Inform and invite the farmers in the sample personally to attend the registration meeting
- Strongly encourage farmers in sample to register
- Send reminders on a regular basis for farmers in the sample to attend training

Option 3: Collect baseline and endline covariates at the Kebele and Woreda levels

- Including predictors in the analysis (e.g. altitude, socio-economic indicators) at the Kebele and Woreda levels will increase the effective power of the sample
- This is because intra-cluster correlation conditioned on good predictors is much smaller

# There are a number ways to positively influence the power of these experiments and the evaluation overall (2) …

Option 4: Ensuring loss due to bad Yield and Best Practice data minimal

- Missing variables can significantly reduce power
- Currently a lot of yield data is lost due to bad quality
- In terms of yield data, one option is to collect yield data by plot rather than across plots

Option 5: Limit to the maximum attrition through consent, good identification data and regular visits

- Obtaining consent and explaining the data collection process and use of data is key!
- Especially in the control group, bad identification data could unnecessarily increase the attrition rate
- Contact farmers on an annual basis in order to increase chances of finding them at endline

Option 6: Increase the number observations in time … i.e. conduct another round of data collection

- If the above fail, and we will know this at the endline, then we can add one more round of data collection
- The more periods in time, the smaller the sample size requirements

# There are a number ways to positively influence the power of these experiments and the evaluation overall(3) …

Option 7: Screen farmers before selecting them into the Treatment and Control groups

- Sample size calculations are very sensitive to the compliance rate
- We can increase compliance in our sample by only including farmers that would be likely to attend the training if given the opportunity (would mean sample representative of motivated farmers only)

# How to improve our chances of success

*Change how BP and yield data is collected*

# Too much data seems to be lost in the yield sample, based on baseline data from Jimma and Illubabor

**Data lost in Illubabor Baseline**

- Total number of farmers in sample: 531
- Total lost due to missing values: 59
- Total lost because of yield levels above, often way above, the 4200kgs/ha threshold: 44
- Total lost: 103 observations

**Data lost in Illubabor = 19.4% of sample**

**Data lost in Jimma Baseline**

- Total sample 771
- Total lost due to missing variables: 3
- Total with 0 yield: 16
- Total with yields above 4200kgs/ha: 96
- Total lost: 115

**Data lost in Illubabor = 14.9% of sample**

**The problem**

- This loss of data greatly affects sampling power
- It also introduces bias, given that it is not random – there is a reason why one farmer registers his data correctly and the other not
- If issues affect the treatment/control groups differently, then internal validity is reduced

# What can TechnoServe do to improve the way yield data is collected?

1. **Collect yield data by plot, rather than across plots (best option)**

- The TNS team believes that one of the main reasons yield levels in some cases were excessively high is because farmers were unwilling to take the TNS team to all their plots, yet collected production data from each plot
- This could also be related to the misreporting of yields or data entry errors, but it does suggest that errors would be reduced significantly if data was collected plot by plot (this is best practice in agricultural surveys)

2. **Collect yield data on only one plot (second best option)**

- Another option, that would reduce costs but not lead to a full picture of how the program has impact total production/yields for a farmer, is to collect data on only one plot per farmer
- A clear allocation mechanism is required. Options include: (i) randomly selecting from list of all the coffee plots a farmer owns; (ii) selecting the plot that is closest to the house; and (iii) selecting the largest plot a farmer owns

3. **Further simplify the data collection form / improve training on data collection**

# Data in the BP sample should be collected from the same plots as data from the yield sample

**Why is this important?**

- If Best Practice adoption really leads to higher yields, then we should observe a clear correlation between best practice adoption and yields
- If yield and Best Practice data are collected from different plots, then we the likelihood that we will obtain a correlation between observed Best Practice adoption and yield levels is significantly reduced
- Creating a link between Best Practice adoption and yield data is a core component of building a convincing story about the impact of the program
- Moreover, if Best Practice data is strongly linked to Yield data, then the predictive value of these covariates will greatly improve the statistical power of the experiment

**What does that imply?**

- Either Best Practice data needs to be collected by plot (best solution)
- Or Best Pratice data should be collected on one specific plot on which yield data (for that specific plot) is available (second best option)

# How to improve our chances of success

*Collect baseline and monitoring data on Woredas, Kebeles and households*

# Examples of baseline and monitoring data that TNS should consider collecting on the sample (1)

**Woreda-level data**

- Population and density
- Number of coffee farmers and number of coffee cooperatives
- Number of hulling stations
- Annual coffee production
- Local price of coffee
- Average yearly rainfall
- Keep log of Woreda level interventions by other organizations, companies or government

**Kebele-level data**

- Population and number of coffee farmers
- Share of farmers in cooperative
- Road connectivity (metric on quality of roads)
- Number of businesses operating in Kebele (e.g number of shops)
- Main crops produced in Kebele
- Distance to closest hulling stations, closest town
- Average Altitude
- Soil quality data
- Average yearly rainfall
- Keep log of Kebele-level interventions by other organizations, companies or government

# Examples of baseline and monitoring data that TNS should consider collecting on the sample (2)

**Household (hh) data**

**Household characteristics**

- Household roster (all members, age, education)
- Total number of plots owned and related crops
- Main income sources
- Assets owned (livestock and basic assets – e.g. radio, bycycle, etc)
- Characteristics of the main house (number of rooms, type of roof, floors, walls)
- The health of household members

**Coffee characteristics**

- How important coffee is for the household
- Whether hh coffee production increased/decreased in past years
- Whether optimistic or not about future coffee production
- Whether good year or low year (phase of the coffee cycle)
- How many years of experience with coffee
- Who in the household works on the coffee plots

**Social characteristics**

- Regular participation or not in village events
- Participation or not in cooperatives other groups
- Whether participates in borrowing groups

# Examples of baseline and monitoring data that TNS should consider collecting on the sample (3)

**Household (hh) data**

- Whether household helps out other farmers on their plots
- Whether other farmers support household to farm its plots
- Whether pools resources with other farmers (e.g. to buy seeds, fertilizer, etc)

**Location data**

- GPS coordinates of household and selected plot
- Time/distance to nearest school, market, collection point, road, etc

# How to improve our chances of success

*One idea to increase power - that will somewhat disrupt validity - is to screen farmers before they are selected into the sample*

# Screening farmers before selecting them into the sample, based on a set of clearly defined criteria, might increase compliance and power

**Why screen farmers at baseline?**

- We expect the compliance rate (i.e. share of coffee-growing households with at least one "treated" member) to be around 60% of total farmers in a Kebele
- This has a very negative effect on statistical power and increases the number of Kebeles required to achieve statistical significance
- Screening farmers to determine how likely they are to attend sessions will help us increase compliance levels in the sample and increase the efficiency of the sample

- We propose that TNS use a very simple screening mechanism, by simply only selecting farmers in the sample that a) are confirmed coffee farmers (rather than randomly sampling at the Kebele level); and b) say they would be likely to take-up the treatment if given the opportunity

- Another option would be to use a more sophisticated screening mechanism,by developing a "motivation score" – and then randomly select farmers into the sample depending on their performance on this metric, which we assume is predictive of compliance.

- One such screening mechanism is described in the next section, but we would recommend that TNS only implement this if compliance rates are expected to be well below 50%.

# What are the three main issues with introducing a screening mechanism?

1. **First of all, screening would mean we're not anymore estimating the ITT or CACE of the program, but rather the treatment effect on a sub-group of people.** This would create a problem of external validity within the evaluation itself – can we extrapolate the results of one sub-group of people, to the remaining farmers in the program? If TNS achieves its objectives for this sub-group, could we still say the program has achieved its objectives?

2. **The effectiveness of any screening mechanism is difficult to measure ahead of time, so we might end-up with a tool that is ineffective.** If the questionnaire is tested on farmers that have already been treated, and for which we have attendance data, then we won't know what is driving the hypothetical association between the results of the screening test and people's attendance rates: is it their motivation levels or the program effect? The same would be true if we compared the results of compliers to non-compliers for example.

3. **If scores on these motivation tests are very different across Kebeles, which they are bound to be, we would be introducing bias into our estimates by selecting individuals based on this score.** Imagine for example if a given score is representative of 80% of the population one Kebele, but just 20% of the population of the other. This would create unnecessary imbalances in the sample.

# The objective of this screening mechanism is to distinguish between farmers that are likely/unlikely to sign-up and attend the training

- We propose a screening mechanism that is based on five categories of interest:
  - ✓ **Attitudes** about coffee farming and training;
  - ✓ **Intentions** in terms of signing-up for training;
  - ✓ **Perceived behavioral control** in the context of a training program
  - ✓ **Social connectedness** applied to coffee farming
  - ✓ **The importance of coffee farming** for the household

- Attitudes, intentions and perceived control-related tests have been proven to be predictive of attendance rates in different settings (no evidence could be found for agriculture programs)

- **For these screening questions, TNS would probably get much more nuanced and better designed questions by conducting a couple of focus groups in the field and letting the field team come up with some of the questions.** The categories we propose can help channel and structure some of these questions.

- Please consider what follows as some examples of what might be interesting questions.

# Attitude related questions

| No | Question or statement | Options and score |
|---|---|---|
| 0 | Are you a coffee farmer? | If No, then excluded from sample automatically |
| 1 | I am a coffee farmer because I had not other choice, not because I am good at coffee farming | • Agree = 0<br>• Disagree =1 |
| 2 | In your opinion what determines your crop yield the most: luck or knowledge of coffee farming? | • Luck = 0<br>• Knowledge = 1 |
| 3 | Do you think a training program on coffee farming would be worth your time? | • Yes = 1<br>• No = 0 |
| | **Total points for attitude** | 3 points |

- Assumption: attitudes towards coffee farming, knowledge and training will play an important role in determining whether a farmer decides to register for the training

# Questions related to intention, willingness to join to learn more about coffee and join a potential training program

| No | Question or statement (agree/disagree) | Options and score |
|---|---|---|
| 4 | I feel like there is a lot more I need to learn about coffee farming | • Agree = 1<br>• Disagree =0 |
| 5 | Would you be willing to pay X (reasonable but high price) to attend a coffee training program? | • Yes = 2 (go to next section)<br>• No =0 |
| 6 | If not, would you be willing to pay Y (low price) to attend a coffee training program? | • Yes = 1 (go to next section)<br>• No =0 |
| 7 | If not, would you attend a coffee training program if it were free? | • Yes = 0<br>• No = excluded from sample |
|  | **Total points for attitude** | 3 points |

- Assumption: testing the willingness to learn and to pay for a training program is a good filter to identify farmers that are very unlikely to join the program if given the opportunity

# Questions related to perceived behavioral control

| No | Question or statement (agree/disagree) | Options and score |
|---|---|---|
| 8 | Sometimes I feel to lazy to go out and work in the field, especially on my coffee plots | • Agree = 0<br>• Disagree = 1 |
| 9 | I am generally very organized as a person and rarely procrastinate | • Agree=1<br>• Disagree=0 |
| 10 | You usually have little control over how your day plays-out? Unexpected issues often come-up | • Agree = 0<br>• Disagree = 1 |
|  | **Total points for attitude** | 3 points |

- Assumption: behavioral control will make the difference between registering and being a high attendance farmer

# Questions related to social connectedness in the coffee context

| No | Question or statement (agree/disagree) | Options and score |
|---|---|---|
| 11 | I make decisions on coffee farming alone and rarely consult with other farmers | • Agree = 0<br>• Disagree =1 |
| 12 | Coffee farmers in my Kebele, don't help each other as much as they should | • Agree = 0<br>• Disagree = 1 |
| 13 | I often help out on other people's plots, even if I have a lot of work on my own | • Agree = 1<br>• Disagree = 0 |
| | **Total points for attitude** | 3 points |

- Assumption: the farmer college program is very much influenced by group dynamics, as farmers are asked to form groups. Our working assumption is that the tighter and more connected group, the more likely farmers are to attend sessions regularly and feel peer pressure to do so

# Questions related to the importance of coffee for a given farmer

| No | Question or statement (agree/disagree) | Options and score |
|---|---|---|
| 14 | Coffee is not the most important source of revenue for our household | • Agree = 0<br>• Disagree =1 |
| 15 | My household has increased its coffee production over the past 2 years | • Agree = 1<br>• Disagree = 0 |
| 16 | I am worried that my coffee revenues will go down in the next few years | • Agree = 0<br>• Disagree = 1 |
| | **Total points for attitude** | 3 points |

- Assumption: from the Rwanda program, one thing that came out clearly was that the more important coffee was for a farmer, the more likely he was to attend. Questions such as – my household has increased coffee production over the past few years – were highly predictive of whether that farmer would have signed up for the treatment or not

# How to deliver the questionnaire

- Enumerators should be **“blind”** as to which clusters are in the Treatment group or not. The risk is that they might otherwise influence responses to these questions.

- This questionnaire should take no-longer than 10 minutes to administer – it will be paper based, papers should be kept and submitted to TNS for electronic data entry. Questionnaires should include relevant farmer identification and location information, so that the data can later be matched to the TNS id, yield, BP and attendance datasets

- Assuming enumerators can conduct 2 interviews per hour (taking the time to move between farms into account), we assume that an enumerator can conduct about 14-15 such interviews per day.

- 40 coffee farmers will randomly be selected from the list of all coffee farmers in a given Kebele. It will take a team of about 3 enumerators per Kebele to complete this task in a day. At the end of the data, enumerators will score each entry (score out of 15), rank the farmers on that score, and select the best 15 farmers to be included in the treatment or control group + 5 replacement farmers

- The next day, the can reach out to the selected farmers and provide them with scales

# Many thanks ….

*www.laterite-africa.com*