Appendix A

Table (A1). Comparative Analysis of Impact Evaluation Methodologies.

Approach	Intervention group	Reference group	Fundamental hypothesis	Constraints
Difference-in-difference	Participants in the program (recipients of benefits)	Individuals not participating (non-recipients of benefits).	If the program were not in place, the outcomes for both the treatment and control groups would evolve similarly over time.	Testing this assumption necessitates several rounds of pre-intervention data collection, which may not always be feasible.
Propensity score matching	Participants in the program (recipients of benefits)	Non-participating individuals (non-beneficiaries); each member in the control group is matched with a treatment group member based on observable characteristics so that an equal likelihood of program participation can be predicted.	Program participation is influenced solely by the characteristics observed and used for matching.	Presumes that no unobserved differences exist between the treatment and control groups.
Regression discontinuity	Entities surpassing a certain threshold when ranked according to specific criteria, such as a poverty index	Entities near the threshold but not qualified to receive the intervention.	Entities both above and below the cutoff share identical statistical characteristics. The population near the cutoff (both above and below it) accurately represents the entire population.	Comparability "similarity" between entities on either side of the cutoff.
Randomized control trial	Random assignment to the treatment group before the intervention	Assigned to the control group through randomization before the intervention.	Random assignment creates two identical groups. Both the treatment and control groups share similar observable and unobservable characteristics.	An experiment's external validity is constrained when conducted in a particular context.

Reference: *Compiled from multiple sources, primarily [17] and [1] (p. 256).*

DRAFT

Appendix B

B1. Example of RCTs Using Completely Randomized Design (CRD).

Description

This example demonstrates how, in a CRD, treatments—including control treatments—are randomly assigned to experimental units, emphasizing that all units have an equal chance of receiving any treatment without blocking. Control of treatments is a fundamental aspect of experimental design; it works hand-in-hand with randomization to ensure that the results are due to the treatments applied and not to other extraneous factors. By including control treatments and randomizing their assignment, researchers can make meaningful comparisons and strengthen the validity of the study's conclusions.

Table (B1): Random Assignment of Treatments in CRD.

Experimental Unit	Treatment Assigned
Unit 1	Treatment A
Unit 2	Control
Unit 3	Treatment B
Unit 4	Treatment A
Unit 5	Treatment B
Unit 6	Control
Unit 7	Treatment A
Unit 8	Treatment B
Unit 9	Control
Unit 10	Treatment A

Explanation

In a CRD, treatments are assigned completely at random to each experimental unit, making it suitable for situations where the units are homogeneous [30]. As noted in the table above, Treatment A was assigned four times, while the Control and Treatment B were each assigned three times. This occurred because, in a CRD, treatments are assigned randomly to experimental units, and due to the nature of randomization, it's possible—though not ideal—for treatments to have unequal replication.

The main concern with unequal replication is that it can affect the balance of the experiment and reduce the statistical power of the analyses. Unequal sample sizes can lead to less precise estimates of treatment effects and may complicate statistical comparisons between treatments. To address this issue, researchers often aim for equal numbers of replicates per treatment to ensure balance and improve the reliability and validity of the results. Therefore, while pure randomization can sometimes lead to unequal assignments, it is common practice to design experiments with equal replication for each treatment to enhance the robustness of the findings.

B2. Example of RCTs Using Randomized Complete Block Design (RCBD)

Description

This example illustrates how an RCBD is employed to control variability in an RCTs by blocking experimental units based on a known source of variation—in this case, soil type. The experimental units (plots) are grouped into blocks where each block represents a different soil type (Soil Type A, B, and C). Within each block, treatments are randomly assigned to the experimental units [30]. This design controls for variability among blocks by ensuring that each treatment is tested under similar conditions within each block. By accounting for the known variability due to soil differences, the RCBD enhances the precision of the experiment and allows for a more accurate assessment of treatment effects.

Table (B2): Random Assignment of Treatments within Blocks in RCBD

Block 1: (Soil Type A):

Plot	Treatment Assigned
Plot 1A	Treatment A
Plot 1B	Control
Plot 1C	Treatment B

Block 2: (Soil Type B):

Plot	Treatment Assigned
Plot 2A	Treatment B
Plot 2B	Treatment A
Plot 2C	Control

Block 3: (Soil Type C):

Plot	Treatment Assigned
Plot 3A	Control
Plot 3B	Treatment B
Plot 3C	Treatment A

DRAFT

Explanation

The following important aspects are illustrated in this RCBD example:

Blocking Based on Soil Type: The experimental units are grouped into three blocks according to soil type, which is a significant factor that can influence the outcome of the treatments. By doing so, we control for the variability associated with different soil conditions.

Randomization Within Blocks: Within each block, the treatments (Control, Treatment A, and Treatment B) are randomly assigned to the plots. This randomization ensures that the assignment of treatments is unbiased and that each treatment has an equal chance of being assigned to any plot within the block.

Equal Representation of Treatments: Each treatment appears once in every block, providing balanced replication across the entire experiment. This means that all treatments are tested under each soil type, allowing for comparisons both within and across blocks.

Control of Variability: By keeping the soil type constant within blocks and randomizing treatments within those blocks, the RCBD effectively controls for the variability due to soil differences. This design isolates the treatment effects from the block effects, making it easier to detect true differences caused by the treatments.

B3. Comparative Table Highlighting Differences Between CRD and RCBD in RCTs:

In agricultural research, RCTs are fundamental experiments that eliminate bias by randomly assigning treatments to experimental units, ensuring that differences in outcomes are due to the treatments themselves. The structure of an RCTs can vary depending on the experimental design chosen, such as the CRD or the RCBD.

To clarify when to employ a CRD or an RCBD in RCTs, the following table presents a detailed comparison of their key features. It highlights how each design handles crucial factors like the homogeneity of experimental units, control of variability, methods of randomization, complexity, statistical analysis requirements, and suitability for various experimental conditions. Understanding these distinctions enables researchers to choose the design that best aligns with their study objectives and conditions, thereby enhancing the validity and reliability of their experimental outcomes.

Table (B3.1): Comparison of CRD and RCBD.

Criteria	CRD	RCBD
Homogeneity of Units	Requires experimental units to be homogeneous.	Accommodates heterogeneous units by grouping them into blocks where units within each block are similar [29] (p.29).
Control of Variability	Offers limited control over external variability. Not ideal for field experiments with significant variation among plots [29] (p.8).	Controls for known sources of variability through blocking. Effective in managing variability due to factors like soil type or farmer characteristics.
Randomization	Treatments assigned completely at random to experimental units. Useful when equal replication is difficult to achieve [29] (p.17).	Treatments are randomized within each block. Each block contains all treatments, ensuring fair comparisons [29] (p.20).
Complexity	Simpler design with straightforward analysis.	Slightly more complex due to blocking, requiring careful planning and statistical consideration.
Statistical Analysis	Analyzed using ANOVA without block effects included.	Analyzed using ANOVA with block effects included to account for variability among blocks.
Suitable For	Small-scale experiments with uniform conditions. Appropriate when experimental material makes equal replication challenging [29] (p.17).	Experiments where controlling variability is essential. Particularly suited for field experiments with a predictable productivity gradient and a manageable number of treatments [29] (p.20)

Source: *Compiled by the authors based on information from [29]; [80]; [30].*

Appendix C

Example C1. Sample Size Calculation for RCTs

Initial Calculation using STATA 15

Command Used: *Power two proportions 0.15 0.30, m1(10) m2(10) rho(0.11)*

In calculating the sample size for a RCT, we chose an intra-cluster correlation coefficient (ICC) of 0.11, informed by studies [59] and [81], which reported ICCs of 0.11 and 0.14, respectively. We opted for the lower value as it better aligns with the extension research context. Notably, ICC reporting is often insufficient, as highlighted by [82], who found that only 6 out of 352 trials reported ICCs. We set the average cluster sizes, M1 and M2, at 10, reflecting the practicalities of the extension interventions being evaluated.

Command Output

Performing iteration

Estimated numbers of clusters for a two-sample proportions test, cluster randomized design, Pearson's chi-squared test, Ho: $p_2 = p_1$ versus Ha: p_2 != p.

Input Parameters

Alpha (α): 0.0500

Power: 0.8000

Difference (Δ): 0.1500

Proportion 1 (p_1): 0.1500

Proportion 2 (p_2): 0.30

Cluster Design:

Average Cluster Size (M_1): 10

Average Cluster Size (M_2): 10

Intra-Cluster Correlation Coefficient (rho): 0.11

Estimated Numbers:

Number of Clusters for Group 1 (K_1): 24

Number of Clusters for Group 2 (K_2): 24

Total Sample Size for Group 1 (N_1): 240

Total Sample Size for Group 2 (N2): 240

Calculation of Design Effect: (Deff)

After the sample size is calculated, the next step is to determine the design effect due to clustering and then inflate the sample size accordingly. The value of the design effect determined by using the formula mentioned below that is derived from the different studies:

Calculating the Design Effect (abbreviated as Deff) is important when dealing with clustered sample designs. It provides a numerical measure of how much the clustering method increases the variance compared to simple random sampling. The formula for computing the design effect is as follows:

$$\textbf{Deff} = 1 + (m - 1) \times \textbf{ICC} \qquad (6)$$

where:

- *Deff* stands for the design effect, which shows how much the variance increases because of the cluster sampling method.
- *'m'* in the formula represents the average cluster size, an important factor in the calculation that considers the size of each cluster in the sampled population.
- *ICC* denotes the intra cluster correlation coefficient, that represents the similarity index of clusters.

To calculate the design effect following equation can be used:

$$\textbf{Deff} = 1 + (10 - 1) \times 0.11 = 1.1 \qquad (7)$$

Above calculation, Deff value = 1.1 suggests that there is moderate increase in variance because of clustering. However, it is important to understand, that the extent of the design effect can vary significantly across studies. Previous study represented in [45] and [81], has identified a Deff value of approximately 2.3, that actually suggest that clustering has a substantial impact on the variability of the resulted measure. This study emphasizes the vital role of the design effect in the analysis of clustered data.

Although Deff values holds much importance there is much gap and very few reports are available. These gaps highlight the challenges associated with estimating the design effect without specific experimental data from analogous research studies. Hence, the variance in presented Deff values highlights the need for experiential estimations that are specific to the unique characteristics of the design and context of study. While Equation 6 provides a basic method for calculating the design effect, researchers are highly recommended to consider the specificities of their study designs and the available experiential data to get more accurate and practical estimates.

DRAFT

Adjusted sample size calculation

When working with clustered sampling methods, it's vital to consider the clustering effect in calculating the adjusted sample size. The formula mentioned below can be used to evaluate the adjusted sample size:

$$n_{adjusted} = (n) \times Deff \quad (8)$$

In this formula, $n_{adjusted}$ gives the value of the sample size adjusted to account for the clustering effect. The term *n* in the formula represents the aggregate sample size derived from random sampling. To sum up the two individual group we use the (n1 + n2). After applying this formula, we get the value of $n_{adjusted}$ as mentioned below:

$$n_{adjusted} = 480 \times 2.3 = 1104 \quad (9)$$

The sample size is adjusted to compensate for the clustering effect, which we have denoted as 2.3 in this case. This adjustment maintains the sample's strength and representativeness, despite the possible decrease in precision because of clustering. This method and the above-mentioned formula for adjusting the sample size have been calculated by [83, 84], and other researchers. It highlights its credibility and extensive acceptance in research methodologies that involves clustered sampling designs.

Factoring in dropout rate

It is really important to consider the attrition rate when figuring out how many people we need for a study. Prior investigations have revealed varying attrition rates, including 4.14% [44], 1.4% [59], 7% [39], 9.24% [61], 6% [65], 22% [13], and 16% [23]. Such practice emphasizes the importance of allowing for dropout percentage when considering different sample sizes.

In this plan, we calculated the total sample size (N) by adding up the sub-samples (N1 + N2). Following, we applied a Deff of 1. 1, which finally resulted in the estimated number of 528 participants. In order to take into consideration a 20% attrition rate we added extra 105 participants, bringing the total sample size to 643 participants. This adjustment was made to preserve the statistical reliability of study in spite of the fact that some participants might drop out.

Alternative calculation method

Another method that is STATA 15 offers researchers the advanced tools for calculation of power and sample size and such a detailed and accurate analysis is provided. The software features a user-friendly interface, letting the researchers key in essential parameters such as the research design. These parameters include effect size, alpha (significance level) and power threshold and many others. Those inputs are used by the software to identify a minimal sample size that is required for achieving a given level of statistical power.

The following figures present the outcomes derived from the execution of the aforementioned STATA command are presented.

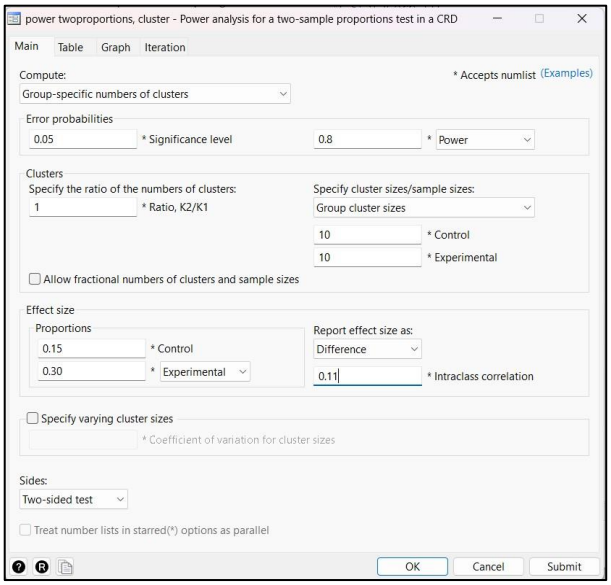

Figure (C.1): Power Analysis and Sample size calculation.

```
Performing iteration ...

Estimated numbers of clusters for a two-sample proportions test
Cluster randomized design, Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.1500  (difference)
           p1 =    0.1500
           p2 =    0.3000

Cluster design:

           M1 =        10
           M2 =        10
          rho =    0.1100

Estimated numbers of clusters and sample sizes:

           K1 =        24
           K2 =        24
           N1 =       240
           N2 =       240
```

Figure (C.2): Power Analysis and Sample size calculation.

Box (C1): Example of a Randomization Balance Test

Imagine you're evaluating the impact of a new farming technique. You randomly assign 200 farmers into two groups: 100 receive the new technique (treatment group) and 100 follow traditional methods (control group).

Before implementing the technique, you gather data on farm size, crop yield, and income. A balance test would compare these factors between the two groups. Similar average farm sizes, yields, and incomes between groups suggest effective randomization.

However, if you notice a trend (e.g., the treatment group has consistently higher crop yields) before the technique is applied, this could indicate a randomization issue.

To investigate the issue further, you would conduct a joint orthogonality test. You would use a regression model with the treatment status (new technique vs. traditional) as the dependent variable and farm characteristics as independent variables. If the coefficients for these characteristics are not significantly different from zero, it would confirm the groups' balance, indicating that the treatment status is not influenced by these characteristics. Thus, the randomization was successful.

Note: *The preceding calculations and discussions are intended to provide a further understanding of the determination of sample size.*

DRAFT

Appendix D7

D1. Available Software for Sample Size Calculations in RCTs:

The following table provides a concise overview of software tools available for performing sample size calculations in RCTs. Each software is briefly described along with its key features, reference, and download link to aid researchers in selecting the most suitable tool for their specific study requirements.

Table (D1): Available Software for Sample Size Calculations

Software	Description	Key features	Reference	Download Link
G*Power	A free, user-friendly tool widely used for basic RCTs power and sample size calculations. Suitable for simpler RCTs designs.	Supports t-tests, ANOVA, regression, and power analysis for straightforward RCTs designs.	[85].	G*Power
PASS	A comprehensive software for advanced RCTs power and sample size analysis, particularly useful for agricultural trials.	Handles complex study designs, including cluster randomization and stratified designs.	[86]	PASS
STATA	A versatile statistical software with built-in commands for complex RCTs designs, suitable for advanced agricultural trials.	Supports multi-level clustered trials and complex RCTs designs like CRD and RCBD.	[87]	STATA
SPSS Sample Power	An SPSS extension focused on sample size calculations for a variety of RCTs designs, ideal for researchers already using SPSS.	Integrates seamlessly with SPSS for conducting more advanced statistical analyses in RCTs.	[88]	SPSS SamplePower
OpenEpi	A free, open-source tool suited for basic sample size calculations in RCTs, with a focus on epidemiology and agricultural research.	Simple interface with tools for calculating sample sizes for various RCTs designs.	[89]	OPENEPI
Statulator	Web-based tool that allows for fast and easy sample size calculations for RCTs, suitable for quick statistical computations.	User-friendly tool for quick calculations of proportions, means, and survival rates in RCTs.	[90]	STATULATOR

Note: Compiled by the authors from various literature sources and references provided above.

DRAFT