Contents

• Purpose of baseline data
• External validity
  – Definition
  – Main limitations
  – How to assess it
• Internal validity
  – Definition
  – Main threats
  – Validating incentives design
  – Validating balance across study arms
• Ensure balanced randomization
• Ex-post power calculations
• Dissemination: Proposed outlines of baseline reports
Purpose of baseline data

Following baseline data collection: Validating IE design | After follow-up: Evaluating impact
---|---
• Describe initial conditions before intervention
• Evaluate how comparable the surveyed population is to national population or other groups of population (External validity)
• Assess if conditions differ across study arms (Internal Validity)
• Determine if competing interventions are already implemented (formal or informal): nature, scope, take-up, difference across study arms (Internal validity)
• Ensure balanced treatment assignment if randomization is done after collection of baseline
• Run ex-post Power calculations to confirm power of the study post-field work and feed discussion on exposure

• Use with follow-up data to assess impact (e.g. Difference in difference)

Contents

• Purpose of baseline data
• External validity
  – Definition
  – Main limitations
  – How to assess it
• Internal validity
  – Definition
  – Main threats
  – Validating incentives design
  – Validating balance across study arms
• Ensure balanced randomization
• Ex-post power calculations
• Dissemination: Proposed outlines of baseline reports
External validity

- To what extent can the causal inference, or impact of the intervention be generalized to a wider population?
  - To what type of population: e.g. all population, households with recent birth, mothers 15-49?
  - To what geographic level: e.g. Country, Province, District?

Limitations to External validity: Sampling

- Target population:
  - If sampled households are selected on specific criteria (e.g. most recent birth in village), results cannot necessarily be extended to general population.
  - If only certain types of health facilities (e.g. district hospitals) are selected, then results cannot be extended to all types of facilities.
  - Does not mean results will not inform effectiveness of the intervention, as the intervention is designed to impact a specific population (e.g. Maternal and child health intervention targets mothers and their children).
Limitations to External validity: Sampling

• Geographic area:
  – If selected geographic area is not randomly selected at national level, then results do not necessarily apply to the whole country.
  – If areas are selected based on criteria such as poverty level of the province/sector/district, results should be considered with respect to this selection.

External validity: How to assess it?

• Theoretically: refer to sampling strategy and determine who the results should be generalized to.
• Empirically: contrast basic socio-demographic characteristics from the survey to comparable surveys, or nationally representative surveys (e.g. DHS).
• Use both theoretical frame and empirical results to evaluate external validity.
Example: Rwanda Community PBF baseline

<table>
<thead>
<tr>
<th>Age</th>
<th>HRITF IE 2010</th>
<th>DHS 2007-08</th>
<th>Age</th>
<th>HRITF IE 2010</th>
<th>DHS 2007-08</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5</td>
<td>32.6</td>
<td>17.2</td>
<td>45-49</td>
<td>1.4</td>
<td>3.4</td>
</tr>
<tr>
<td>5-9</td>
<td>12.1</td>
<td>14.9</td>
<td>50-54</td>
<td>1.0</td>
<td>3.2</td>
</tr>
<tr>
<td>10-14</td>
<td>7.4</td>
<td>14.2</td>
<td>55-59</td>
<td>0.5</td>
<td>2.1</td>
</tr>
<tr>
<td>15-19</td>
<td>4.9</td>
<td>9.4</td>
<td>60-64</td>
<td>0.4</td>
<td>1.6</td>
</tr>
<tr>
<td>20-24</td>
<td>10.5</td>
<td>9.1</td>
<td>65-69</td>
<td>0.3</td>
<td>1.1</td>
</tr>
<tr>
<td>25-29</td>
<td>13.0</td>
<td>8.1</td>
<td>70-74</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>30-34</td>
<td>8.2</td>
<td>5.5</td>
<td>75-79</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>35-39</td>
<td>4.5</td>
<td>4.4</td>
<td>80+</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>40-44</td>
<td>3.0</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Why?**
- Sample = Households with birth in last 4 months: more children under 5, more adults of childbearing age

**Contents**

- Purpose of baseline data
- External validity
  - Definition
  - Main limitations
  - How to assess it
- Internal validity
  - Definition
  - Main threats
  - Validating incentives design
  - Validating balance across study arms
- Ensure balanced randomization
- Ex-post power calculations
- Dissemination: Proposed outlines of baseline reports
**Internal validity**

- Can we prove causal inference, i.e. the impact of the intervention, given our study design?
  - Randomized Impact Evaluation: the goal is to make sure the impact can be identified separately from all potential confounders
  - In practice: does the baseline data indicate we randomized properly?
    i.e. are our study arms comparable at baseline: no difference in observed characteristics across the groups that receive treatment and those that don’t?
- In summary: was the randomized assignment well implemented?

---

**Main threats to internal validity**

- Confounding factors: outcome can be explained by other factor not accounted for in the analysis (study arms not comparable at baseline)
  E.g. Treated sectors are wealthier. Better outcomes at follow-up in treatment group due to better financial access to healthcare, not necessarily or not only to incentive.
- Selection bias: beneficiaries self-select themselves into the intervention
  E.g. Demand-side incentives are taken on by women more concerned about getting antenatal care (ANC) in the first place. Better outcomes in treatment group are due to beneficiaries taking better care of themselves during pregnancy than in comparison group, not necessarily or not only to incentive.
- Contagion: Comparison imitates Treatment
  E.g. Comparison sectors start implementing demand-side incentives to get women to deliver in facility. Comparison group does benefit from an intervention in the end.
Validating Incentives Design

- Is incentive already in place at baseline?
  - If No: Perfect.
  - If Yes: ?

- Does not necessarily invalidate design
- Check expected effect and magnitude of incentives: could it outweigh incentives from the intervention?

- Yes and incentive is implemented similarly in Treatment and Comparison (same magnitude and implemented everywhere)
- Yes and incentive is implemented differently in Treatment and Comparison (different magnitude and/or not implemented everywhere)
- Does invalidate design.

Example: Rwanda Community PBF baseline

- Variables of interest (demand-side):
  - Gift at first ANC visit
  - Gift at delivery
  - Gift at first PNC visit
  - Value of gifts

  Frequency

  Magnitude
Example: Rwanda Community PBF baseline

• Idea: Making sure that if incentives are already being implemented at baseline, they are implemented in all study arms in a similar way (frequency and magnitude)

• Practically with the data:
  – Calculate means of each study arm
  – Compare those means:
    • Run F-test to detect significant difference in means across study arms overall
    • Run T-tests to detect significant difference in means across each combination of study arms
    • Pvalues ≤ 10%: Significant difference across study arms

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>C</th>
<th>All</th>
<th>T1-C</th>
<th>T2-C</th>
<th>T3-C</th>
<th>T1-T2</th>
<th>T1-T3</th>
<th>T2-T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>RWF</td>
<td>USD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GiY</td>
<td>ANC</td>
<td>5.6%</td>
<td>3.6%</td>
<td>4.2%</td>
<td>4.1%</td>
<td>0.778</td>
<td>0.452</td>
<td>0.718</td>
<td>0.952</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td></td>
<td>749</td>
<td>905</td>
<td>815</td>
<td>853</td>
<td>0.863</td>
<td>0.682</td>
<td>0.86</td>
<td>0.885</td>
<td>0.437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.25</td>
<td>1.51</td>
<td>1.36</td>
<td>1.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GiY</td>
<td>Delivery</td>
<td>2.9%</td>
<td>3.6%</td>
<td>2.0%</td>
<td>1.8%</td>
<td>0.437</td>
<td>0.247</td>
<td>0.186</td>
<td>0.816</td>
<td>0.611</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1384</td>
<td>1974</td>
<td>637</td>
<td>791</td>
<td>0.367</td>
<td>0.327</td>
<td>0.308</td>
<td>0.638</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.31</td>
<td>3.29</td>
<td>1.06</td>
<td>1.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GiY</td>
<td>PNC</td>
<td>2.2%</td>
<td>0.7%</td>
<td>1.9%</td>
<td>2.0%</td>
<td>0.35</td>
<td>0.947</td>
<td>0.223</td>
<td>0.951</td>
<td>0.418</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
There are incentives given to women at baseline already. But:

- No statistically significant difference across study arms in the proportion of beneficiaries or in the magnitude of existing incentives.
- Magnitude of existing incentives is smaller than in intervention.

All clear.

Example: Rwanda Community PBF baseline

Validating balance across study arms

- Idea: We want households/health facilities/communities to be similar across study arms at baseline, so the only difference between treatment and comparison occurring between baseline and follow-up is the intervention.
  - Socio-demographic characteristics: Covariates
  - Behavior of interest: Outcome Indicators, e.g. ANC visit (1+, 4+), assisted delivery, etc.
- Practically with the data: Same technique
• Visually, this is the comparison we do:

Example: Rwanda Community PBF baseline

Antenatal care indicators by treatment group
Women 15-49 pregnant or had pregnancy since Jan 08

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>F-test Pvalue</th>
<th>T-test Pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
</tr>
<tr>
<td>ANC 1+</td>
<td>0.984</td>
<td>0.997</td>
<td>0.98</td>
</tr>
<tr>
<td>Timely ANC (&lt;4m)</td>
<td>0.585</td>
<td>0.675</td>
<td>0.654</td>
</tr>
<tr>
<td>ANC 4+</td>
<td>0.344</td>
<td>0.408</td>
<td>0.398</td>
</tr>
<tr>
<td>Skilled delivery</td>
<td>0.865</td>
<td>0.897</td>
<td>0.905</td>
</tr>
</tbody>
</table>
Validating balance across study arms: Results summary

• Calculate for Outcome indicators on the one hand, and for Covariates on the other hand:
  – Proportion of F-tests that detected a significant difference across study arms
  – Proportion of T-tests that detected a significant difference across study arms for each combination of study arms
• Do it for the whole survey, or focus by questionnaire section if most relevant (e.g. ANC/Delivery/PNC section includes most interesting indicators)
• Conclude on general assessment of the balance across study arms
• If unbalanced in certain dimensions, need to anticipate how results can be affected and possibly gather corollary evidence to support argument

Contents

• Purpose of baseline data
• External validity
  – Definition
  – Main limitations
  – How to assess it
• Internal validity
  – Definition
  – Main threats
  – Validating incentives design
  – Validating balance across study arms
• Ensure balanced randomization
• Ex-post power calculations
• Dissemination: Proposed outlines of baseline reports
Using the baseline to assign treatment status

• If baseline is collected before treatment/control assignment then can guarantee balance across study arms
  – Identify key characteristics that partially determine outcomes of interest
    • For community PBF: region, village size, distance to facility, etc.
  – Key characteristics define strata, assign treatment status within strata
  – Alternatively, pair-match units on basis of characteristics and randomize within match
    • Both methods will improve power of study

Contents

• Purpose of baseline data
• External validity
  – Definition
  – Main limitations
  – How to assess it
• Internal validity
  – Definition
  – Main threats
  – Validating incentives design
  – Validating balance across study arms
• Ensure balanced randomization
• Ex-post power calculations
• Dissemination: Proposed outlines of baseline reports
What does IE do?

- Example of community PBF and healthy skeptic
- IE tries to convince Mr. Critic whether community PBF works or not
- Mr. Critic’s opinion is that “the program does not work”
- We are trying to show whether the program DOES work
- So: we are trying to investigate whether Mr. Critic is wrong

**Statistical terminology**

- “null hypothesis” = zero impact
- “Alternative hypothesis” = positive impact
- Aim: try to reject the null hypothesis

What is power?

- Power is the likelihood that we will be able to prove Mr Critic is wrong when indeed he is wrong.
- = likelihood that we can prove that an effective program is effective, i.e. to statistically demonstrate there is an effect when there is indeed a true effect
- For a “good” evaluation, need at least 80% power!
What does power depend on?

- Size of the sample
- Variation in outcome of interest
- The size of correlation in outcomes within study clusters (here: sectors)
- Anticipated size of the impact

Example: Rwanda Community PBF baseline
Some variables of interest

- ANC coverage: 1+ visit
- Timely ANC (<=4m)
- ANC coverage: 4+ visits
- Tetanus Toxoid 2 coverage in pregnancy
- 90 day Iron supplementation in pregnancy
- Skilled delivery
- Delivery in formal health facility
- Low-birth-weight newborns
- Timely initiation of breastfeeding
- Exclusive breastfeeding 0-6months
- Timely PNC in formal health facility
- Postnatal supplementation with vitamin A
- Modern contraceptive prevalence
- Unmet need for Family planning
Example: Rwanda Community PBF baseline
Basic design and parameters

- Blocked 2-level cluster-randomized trial (sectors ranked by poverty ranking before treatment assignment) and binary outcome
- Outcome as defined at the household level
- Households nested within clusters
- Treatment defined at the cluster level
- Clusters are blocked
- \( n=12 \) (households), \( J=2 \) (clusters per block), \( K=50 \) (blocks)
- Two treatment conditions
- Binary outcome
- Size of plausible interval for treatment and comparison groups assumed to remain constant, when possible. When bounded by 0 or 1 the plausible interval for the treatment group is appropriately adjusted.
- Type 1 error rate is 0.05.

Example: Rwanda Community PBF baseline
Simulation 1

Minimum detectable effect
\[ = 0.75 - 0.64 = 0.11 \]

Mean = 0.64
Lowest level = 0.75
Example: Rwanda Community PBF baseline Simulation 2

Minimum detectable effect
= 0.96 – 0.89
= 0.07

Detectable effect at 80% power:

<table>
<thead>
<tr>
<th></th>
<th>ANC coverage: 1+ visit LB</th>
<th>Timely ANC (&lt;=4m) LB</th>
<th>ANC coverage: 4+ visits LB</th>
<th>Timely PNC in pregnancy LB</th>
<th>90d Iron suppl pregnancy LB</th>
<th>Skilled delivery</th>
<th>Deliv in formal HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest level</td>
<td>0.75</td>
<td>0.49</td>
<td>0.42</td>
<td>0.09</td>
<td>0.96</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.99</td>
<td>0.64</td>
<td>0.38</td>
<td>0.31</td>
<td>0.03</td>
<td>0.89</td>
<td>0.80</td>
</tr>
<tr>
<td>Detectable effect</td>
<td>--</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.06</td>
<td>0.07</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Low-birth-weight newborns</th>
<th>Timely initiation breastfeeding</th>
<th>Exclusive breastfeeding 0-6m</th>
<th>Timely PNC in formal HF</th>
<th>Postnatal suppl w/via</th>
<th>Modern contraceptive prev</th>
<th>Unmet need for FP - WHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest level</td>
<td>0.12</td>
<td>0.90</td>
<td>0.97</td>
<td>0.90</td>
<td>0.34</td>
<td>0.16</td>
<td>0.93</td>
</tr>
<tr>
<td>Mean</td>
<td>0.05</td>
<td>0.82</td>
<td>0.91</td>
<td>0.78</td>
<td>0.24</td>
<td>0.08</td>
<td>0.84</td>
</tr>
<tr>
<td>Detectable effect</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
<td>0.12</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Implications from ex-post power analysis

• Use of monitoring: How far are we from attaining those “smallest detectable effects”
  – How will we know?
• Length of exposure: How much time is it going to take to attain those “smallest detectable effects”?
  – How long should the comparison group be maintained?

Contents

• Purpose of baseline data
• External validity
  – Definition
  – Main limitations
  – How to assess it
• Internal validity
  – Definition
  – Main threats
  – Validating incentives design
  – Validating balance across study arms
• Ensure balanced randomization
• Ex-post power calculations
• Dissemination: Proposed outlines of baseline reports
Dissemination

- Report on:
  - Conditions at baseline
  - External validity incl. limitations/implications
  - Internal validity (Incentives design + Balance) incl. limitations/implications
  - Power of the experiment incl. recommendations to increase power if need be at follow up
- For each survey (household, health facility, community, community health worker, etc.)
- Proposed outlines:

<table>
<thead>
<tr>
<th>Outline 1</th>
<th>Outline 2 (Gertler et al 2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overview</td>
<td>1. Introduction</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>2. Description of the Intervention</td>
</tr>
<tr>
<td>1.2 Project Background</td>
<td>(Benefits, Eligibility Rules, etc.)</td>
</tr>
<tr>
<td>1.3 Project Components</td>
<td>3. Objectives of the Evaluation</td>
</tr>
<tr>
<td>1.4 Objectives of the Study</td>
<td>3.1 Hypotheses, theory of changes, results chain</td>
</tr>
<tr>
<td>2. Methodology</td>
<td>3.2 Policy questions</td>
</tr>
<tr>
<td>2.1 Randomization</td>
<td>3.3 Key outcome indicators</td>
</tr>
<tr>
<td>2.2 Study Design</td>
<td>4. Evaluation Design</td>
</tr>
<tr>
<td>2.3 Sample Size and Strategy</td>
<td>4.1 Original design</td>
</tr>
<tr>
<td>2.4 Variables for Data Analysis (Indicators)</td>
<td>4.2 Actual program participants and nonparticipants</td>
</tr>
<tr>
<td>2.5 Instruments for Data Collection and Data Quality Insurance</td>
<td>5. Sampling and Data</td>
</tr>
<tr>
<td>2.6 Data storage, Management &amp; Access Policy</td>
<td>5.1 Sampling strategy</td>
</tr>
<tr>
<td>3. Sample Representativeness and External Validity</td>
<td>5.2 Power calculations</td>
</tr>
<tr>
<td>3.1 Geographic Representativeness</td>
<td>5.3 Data collected</td>
</tr>
<tr>
<td>3.2 Comparison between Baseline Study and Country Population</td>
<td>6. Validation of Evaluation Design</td>
</tr>
<tr>
<td>4. Findings (organized by section of the questionnaire or relevant topics)</td>
<td>7. Comprehensive Descriptive Statistics</td>
</tr>
<tr>
<td>5. Internal Validity of the Study</td>
<td>8. Conclusion and Recommendations for Implementation</td>
</tr>
<tr>
<td>5.1 Threats to internal validity</td>
<td>5.2 Sample Balance: Summary of tests Results</td>
</tr>
<tr>
<td>5.2 Sample Balance: Summary of tests Results</td>
<td>5.3 Data collected</td>
</tr>
<tr>
<td>7. Recommendations for follow-up survey(s)</td>
<td>7. Comprehensive Descriptive Statistics</td>
</tr>
</tbody>
</table>
Thank you