REAL-WORLD IMPACT EVALUATION – APPLYING IE METHODS CREATIVELY

7 May 2017

Jo (Jyotsna) Puri – Head of Evaluation, GCF
Anna Henttinen – IE Specialist, WFP
Bidisha Barooah – Evaluation Specialist, 3ie
Q1: Have you worked on an IE before?

Q2: Do you think you might work on an IE in the future?
Welcome!

UNEG Impact Evaluation Workshop –
Real-World Impact Evaluation – Applying IE methods creatively
7 May 2018 – Draft Agenda

Workshop Purpose: To consider options for designing creative impact evaluations in difficult (i.e. real-world) contexts where data may not be available or the context may be shifting. The objective of the workshop is to introduce the audience to main impact evaluation techniques, and share how they have been applied creatively. The meeting of the workshop will be more traditional presentations and discussion, while during the afternoon, the participants will have an opportunity to design their own impact evaluation with support from the facilitators.

Workshop style: This workshop will be facilitated, and highly participatory, with presentations and discussion in the morning and an interactive impact evaluation exercise in the afternoon.

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Description</th>
<th>Activities</th>
<th>Intended Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:30</td>
<td>Introduction and</td>
<td>Go through the workshop expectations and agenda, purpose and outline of the</td>
<td>Clarify the purpose and scope of the day and objectives</td>
</tr>
<tr>
<td></td>
<td>overview to the day.</td>
<td>day, scope and outcomes.</td>
<td></td>
</tr>
<tr>
<td>09:30-10:30</td>
<td>What is Impact</td>
<td>Definition of impact evaluation, basic design frames for undertaking impact</td>
<td>Shared understanding of the methods and definition of impact</td>
</tr>
<tr>
<td></td>
<td>Evaluation and</td>
<td>evaluation; Experimental and Quasi-Experimental Impact Evaluation designs,</td>
<td>evaluation</td>
</tr>
<tr>
<td></td>
<td>what are</td>
<td>with a particular focus on quasi-experimental.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the common design</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>options</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00-11:20</td>
<td>Examples of Being</td>
<td>Examples from GIZ and UNEP on how impact evaluation techniques have been</td>
<td>Exploration of “real-world” scenario and application of</td>
</tr>
<tr>
<td></td>
<td>Creative with</td>
<td>applied creatively in the field.</td>
<td>quasi-experimental methods.</td>
</tr>
<tr>
<td></td>
<td>Impact Evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>designs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:30-11:45</td>
<td>Lunch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00-12:45</td>
<td>Introduction to the</td>
<td>Explain the afternoon session and the “Impact Evaluation Design Game.”</td>
<td>Clarify over the mode and agenda for the afternoon.</td>
</tr>
<tr>
<td></td>
<td>afternoon and the</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Impact Evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Design Game</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:45-14:00</td>
<td>Planning and Designing your Impact Evaluation</td>
<td>Audience divided into groups to design their impact evaluation on a specific topic, based on a menu of design choices.</td>
<td>Participants use their own knowledge and expertise, with the help of the facilitator to design impact evaluations.</td>
</tr>
<tr>
<td>14:00-14:15</td>
<td>Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:15-17:00</td>
<td>Presenting the</td>
<td>Feedback from groups on their design choices.</td>
<td>Participants share their work and design choices with the group.</td>
</tr>
<tr>
<td></td>
<td>different Impact</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaluation Designs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AGENDA
SESSION 1 - IMPACT EVALUATION – COMMON DESIGN OPTIONS
WHAT IS IMPACT EVALUATION?

• Answers **Cause-and-Effect** questions

• Can identify what happened and how it happened

• *Works at any point* of the results chain;

• Can identify who benefitted or if a programme is cost-effective.

• Can measure short, medium or long-term effects

• Can be retrospective or prospective
DEFINITION – WHY IT MATTERS

- Most international organisations, and Donors include the following words:
  - Counterfactual
  - Attribution
Use of a credible counterfactual to identify what would have happened in the absence of the intervention.
Why counterfactual matters

- What if outcomes and impact worsen during operations?

**Literacy Rate of girls**

- Baseline before programme: 50%
- 3 years after programme: 30%

Diagram showing literacy rate decrease from 50% to 30% over 3 years.
WHAT IF WE DIDN’T DO THEM
What do we need to measure impact?

PROVIDING CASH TRANSFERS TO THE DISADVANTAGED AND LOW INCOME GROUPS

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project (treatment)</td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>comparison</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The majority of evaluations have just this information … which means we can say absolutely nothing about impact.
BEFORE VERSUS AFTER SINGLE DIFFERENCE COMPARISON
BEFORE VERSUS AFTER = 92 – 40 = 52

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project (treatment)</td>
<td>40</td>
<td>92</td>
</tr>
</tbody>
</table>

"the cash transfer project has led to a higher incomes in a number of villages"

This ‘before versus after’ approach is outcome monitoring. Outcome monitoring has its place, but it is not impact evaluation
**POST-TREATMENT COMPARISON**

**COMPARISON**

**SINGLE DIFFERENCE = 92 − 84 = 8**

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project (treatment)</td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>comparison</td>
<td></td>
<td>84</td>
</tr>
</tbody>
</table>

But we don’t know if they were similar before…
DOUBLE DIFFERENCE =
(92-40)-(84-26) = 52-58 = -6

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project (treatment)</td>
<td>40</td>
<td>92</td>
</tr>
<tr>
<td>comparison</td>
<td>26</td>
<td>84</td>
</tr>
</tbody>
</table>

Conclusion: Longitudinal (panel) data, with a comparison group, allow for the strongest impact evaluation design (though still need matching).

SO WE NEED BASELINE DATA FROM PROJECT AND COMPARISON AREAS
What do we need to measure impact?

<table>
<thead>
<tr>
<th>Project</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="Before" /></td>
<td><img src="image2" alt="After" /></td>
</tr>
<tr>
<td>Comparison</td>
<td><img src="image3" alt="Before" /></td>
<td><img src="image4" alt="After" /></td>
</tr>
</tbody>
</table>
SO IN FACT

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>![Before Project]</td>
<td>![After Project]</td>
</tr>
<tr>
<td>Comparison</td>
<td>![Before Comparison]</td>
<td>![After Comparison]</td>
</tr>
</tbody>
</table>
EXERCISE: 10 MINUTES
EXERCISE: PART 1

Step 1: Think of an intervention you would like to assess the impact of.

Step 2: Define one main impact indicators for your intervention

Step 3: Using hypothetical outcome data for one indicator write down the before/after, comparison/treatment numbers in the table below

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 4: Write down the following numbers in the sheet you received:

- Ex-post single difference
- Before versus after (single difference)
- Double difference impact estimates
HOW DO YOU CREATE A COUNTERFACTUAL?
Most development impact evaluations today use different methods and mixed methods.

Some are ‘conventional’ RCTs ... but increasingly other more creative methods are used in more complex settings.

What follows is a light taster of a range of methods...
RANDOMIZATION

- Two levels of randomization
  - Individual randomization
  - Cluster randomization
- Threats of spillover and contamination
- Ethics
- Cluster randomization
  - Eg. Schools instead of students
- Sample size requirements may be bigger

Randomly assign
‘BUT I CANNOT RANDOMIZE EVERYONE IN MY PROGRAM…!’

- Pipeline design
  - Most development programs are implemented in phases. Assignment to phases is random
  - Measures duration of program
- Factorial design
  - All groups get a base treatment
- Lottery
  - Oversubscription to a program
- Encouragement design
  - Low sign-up to a program, encourage to increase participation
RCTs – two practical ways to include an RCT
Program placement is hardly ever random.

There is ‘selection’ in who benefits from nearly all interventions.

Need a comparison group which has the same characteristics as those selected for the intervention.
COMMON QUASI-EXPERIMENTS

- Propensity Score Matching
- Difference in Differences
- Regression Discontinuity Design
- Instrumental variable
Prennushi and Gupta (2014)

- Evaluate a program on woman’s empowerment where women are mobilized into self-help groups. Joining a group is voluntary.
- Compare participants to non-participants.
  - Not as simple as matching on means.
- Each observation gets a ‘score’ of its probability of being in the program based on its observable characteristics.
Figure 2: Estimated propensity scores (early joiners vs. never joiners, poor households)

Common support
| Variable                                                      | Unmatched | Matched | Mean | Treated | Control | %bias | %reduce | |bias| |
|---------------------------------------------------------------|-----------|---------|------|---------|---------|-------|---------|------|------|
| Household size (zhhsiz)                                      | Unmatched | 4.3395  | 3.8141 | 32.3    |         |       |         |      |      |
|                                                               | Matched   | 4.3395  | 4.4917 | -9.4    | 71      |       |         |      |      |
| Female head (zfemalehead)                                    | Unmatched | 0.07792 | 0.09295| -5.4    |         |       |         |      |      |
|                                                               | Matched   | 0.07792 | 0.11503| -13.3   | -147    |       |         |      |      |
| Highest year of schooling in the family (zeducyears)         | Unmatched | 6.7328  | 4.9071 | 43.7    |         |       |         |      |      |
|                                                               | Matched   | 6.7328  | 6.5974 | 3.2     | 92.6    |       |         |      |      |
| No of members that can write in the household (zowriters)    | Unmatched | 1.833   | 1.1923 | 47.5    |         |       |         |      |      |
|                                                               | Matched   | 1.833   | 1.9221 | -6.6    | 86      |       |         |      |      |
| Total expenditure 2004 Rs. (ztotexpb)                        | Unmatched | 2094.3  | 1744.7 | 36.9    |         |       |         |      |      |
|                                                               | Matched   | 2094.3  | 2168.7 | -7.8    | 79      |       |         |      |      |
| Household owned any land in 2004? (zanylandowned)            | Unmatched | .50649  | .48718 | 3.9     |         |       |         |      |      |
|                                                               | Matched   | .50649  | .48980 | -1.3    | 91      |       |         |      |      |
| Household owned any livestock assets in 2004? (zanylivestock)| Unmatched | .30241  | .23718 | 14.7    |         |       |         |      |      |
|                                                               | Matched   | .30241  | .30798 | -4.2    | 71.6    |       |         |      |      |
| Household owned any farm assets in 2004? (zanyfarmassets)    | Unmatched | .83117  | .80769 | 6.1     |         |       |         |      |      |
|                                                               | Matched   | .83117  | .82931 | 0.5     | 92.1    |       |         |      |      |
| SC/ST (zcaste2_1)                                            | Unmatched | .45455  | .30769 | 30.6    |         |       |         |      |      |
|                                                               | Matched   | .45455  | .43785 | 3.5     | 89      |       |         |      |      |
| Other Castes (zcaste2_3)                                     | Unmatched | .11317  | .21795 | -28.4   |         |       |         |      |      |
|                                                               | Matched   | .11317  | .12987 | -4.5    | 84      |       |         |      |      |
Afridi, Barooah and Somanathan (in progress)

School meals were started in urban public schools of Delhi in 2003


**TABLE 2: AVERAGE ATTENDANCE LEVELS AND CHANGES.**

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Δ 2002</td>
<td>0.06</td>
<td>0.06</td>
<td>0</td>
</tr>
<tr>
<td>Mean attendance in April 2002</td>
<td>0.81 (0.073)</td>
<td>0.79 (0.087)</td>
<td></td>
</tr>
<tr>
<td>(B) Δ 2003</td>
<td>0.07 (0.008)</td>
<td>0.11 (0.009)</td>
<td>0.04*** (0.012)</td>
</tr>
<tr>
<td>Mean attendance in April 2003</td>
<td>0.80 (0.086)</td>
<td>0.78 (0.063)</td>
<td></td>
</tr>
<tr>
<td>Difference (B)-(A)</td>
<td>0.01 (0.011)</td>
<td>0.05*** (0.014)</td>
<td>0.04*** (0.018)</td>
</tr>
</tbody>
</table>

Note: The sample consists of 40 schools (20 control and 20 treatment).
REGRESSION DISCONTINUITY DESIGN

There is a programme allocation ‘threshold rule’ dividing participants and non-participants

<table>
<thead>
<tr>
<th>Variable</th>
<th>Threshold rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty index</td>
<td>Impact of development projects to households below a poverty incidence threshold (eg BPL cards)</td>
</tr>
<tr>
<td>Age</td>
<td>Impacts on subsidies for senior citizens (above 60 y.o.)</td>
</tr>
<tr>
<td>Date</td>
<td>Impact of introduction of a reform after a certain time</td>
</tr>
</tbody>
</table>
Before Intervention

After Intervention

Impact
SESSION 2:

HOW TO BE CREATIVE WITH IMPACT EVALUATIONS
BIASES IN IMPACT EVALUATIONS

- Spillover and Contamination
  - Threatens the validity of IEs

- What to do? Examples
  - Design
    - Unit of treatment of a village and not some villagers
  - Intervention
    - Non-transferrable vouchers
  - Monitoring

[Diagram showing examples of treatment and control groups]
Funnel of Attrition
OTHER BIASES

- **Hawthorne Effect**
  - Treatment group modifies behavior not because of the treatment but being observed

- **John Henry effect**
  - Control groups change behavior

- **What to do? Examples**
  - Sensitive survey and monitoring systems
Assessing the robustness of our methods

SAMPLE SIZE AND POWER
Sample Size Calculations

Larger sample → more likely that treatment and control are comparable
LESSONS LEARNT: GENERAL RULES

1. You need a minimum sample size to make good estimations.
2. You need a sample that is diverse enough to represent the population studied.
3. The larger the sample the better and more accurate are your estimations.
4. If you increase sample size, you are likely to increase power.

You need a large sample size!
SOME SAMPLING BASICS

**Population mean**: the true value of a parameter, i.e. the average weight for age of all children aged under in the region of interest.

**Sample mean**: the average weight for age in a sample drawn from the population.

*The larger the sample the more likely it is that the sample mean is close to the population mean (provided our sample is a random sample)*
What do impact evaluators do?

DISTRIBUTION OF NEWBORN WEIGHT IN THE TREATMENT AND CONTROL POPULATIONS BEFORE TREATMENT

2.7

Treatment group

Control group
What do impact evaluators do?

On average weight of newborn is larger in treatment group than in control group.

AND AFTER TREATMENT

Control group 2.7 3.7

Treatment group
Power Calculation and sample size

- **Power (or statistical power)** of an impact evaluation is the likelihood that it will detect a difference between the treatment and comparison groups, when in fact one exists.

  - Power calculation indicate the smallest sample size required for an evaluation to detect a meaningful difference in outcomes between the treatment and comparison groups.
SAMPLE SIZE AND STANDARD ERROR

A larger sample

Control group  2.7  3.7
Treatment group
WHY POWER CALCULATION?

1. Not acceptable to conduct a study that would not be stringent enough to detect a real effect due to a lack of statistical power.

2. Not acceptable to conduct a study by recruiting 1000s of participants when sufficient data could be obtained with 100s of participants instead.

3. Avoid misleading policy recommendations
So how large a sample do we need?
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

- Less variability in the outcome variable

Control group: -0.6, -0.2

Treatment group
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

Less variability in the outcome variable

So we need to know that for our power calculation, but we can’t affect it (though we can change outcome variable)

Control group

Treatment group

-0.6

-0.2
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

A larger sample

Control group

Treatment group

-0.6

-0.2
WHAT MAKES IT EASIER TO DETECT PROGRAMME IMPACT?

A larger sample
More formally
EQUAL TREATMENT AND CONTROL SAMPLES

\[ MDE = \left( t_\alpha + t_{1-\beta} \right) \sigma_y \sqrt{\frac{1}{P(1-P)n}} \]

\[ MDE = f[1/P(1-P)] \]

And obviously increasing n helps

\[ \delta(MDE)/\delta P = (1-P) - P = 1 - 2P = 0 \rightarrow P = \frac{1}{2} \]

\[ \delta^2(MDE)/\delta P^2 = -2 \text{ so maximize MDE} \]
Understand the context and the program, ground realities

Develop the program theory of change

Set out research questions - what can the IE address and not?

Design the impact evaluation
1. Sample size
2. Data requirements
3. Costs
4. Methodology
5. Biases
6. Monitoring of implementation
7. Plan data collection

Formative/Process Evaluation
EVALUATION OF THE NATIONAL RURAL LIVELIHOODS MISSION IN INDIA

- Large scale program on group-based livelihoods support
- The government had conducted baseline surveys in 13 states of India before the roll-out of the program
- There were matched treatment and control areas
Our initial scope of work

- Design an endline survey and report on findings

BUT

- The baseline data was not usable
- The program was rolled out in control areas
What we did

- Examined program records and MIS
- Intensive ground work
- Conversations with field teams
- Proposed a Difference-in-Difference strategy
USING IMPACT EVALUATION: TO ESTIMATE THE IMPACT OF PROTECTED AREAS AND ROADS
Most protected areas and forest reserves in Thailand are in the north.
Areas with high elevation and slopes and bad soils are where protected areas are located.

So do protected areas and forestry programmes really help?
The econometric model that we estimate is thus given by

- \( Z_i \): Plot attributes (Slope, Elevation, Impedance weighted travel time, Soil Dummy, Population density)

- \( Y_{1i}^* \): Net profit from clearing

- \( Y_{2i}^* \): Net utility from protecting a plot

\[
Y_{1i}^* = Z_i B_1 + \gamma e_{1i} + \epsilon_{1i} \\
Y_{2i}^* = Z_i B_2 + \gamma e_{2i} + \epsilon_{2i}
\]

\[
Y_{1i} = \begin{cases} 
1 \text{ if } Y_{1i}^* > 0; = 0 \text{ otherwise}
\end{cases}
\]

\[
Y_{2i} = \begin{cases} 
1 \text{ if } Y_{2i}^* > 0; = 0 \text{ otherwise}
\end{cases}
\]

THE ECONOMETRIC MODEL
THAILAND PROTECTED AREAS

NORTH THAILAND

Protected areas

Elevation

Forests

Roads

Elevat.shp

METHOD: INSTRUMENTAL VARIABLES

Probability of land getting cleared = determined by soil fertility, slope, elevation, distance to the market, administrative factors, population pressure etc.

Probability of land being protected = determined by some of the same factors AND closeness to a watershed area.
<table>
<thead>
<tr>
<th></th>
<th>Cleared Land (Y1 = 1)</th>
<th>T- Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope (degrees)</td>
<td>-0.088</td>
<td>-10.652</td>
</tr>
<tr>
<td>Elevation (m.)</td>
<td>-0.001</td>
<td>-8.095</td>
</tr>
<tr>
<td>Population density1990 (people/km²)</td>
<td>0.003</td>
<td>4.532</td>
</tr>
<tr>
<td>Log(cost) (1982)**</td>
<td>-0.191</td>
<td>-9.729</td>
</tr>
<tr>
<td>Soil and Province Dummies</td>
<td>Not Shown</td>
<td></td>
</tr>
<tr>
<td>Protected Area dummy (1986)</td>
<td>-6.28</td>
<td>-10.332</td>
</tr>
<tr>
<td>Constant</td>
<td>1.295</td>
<td>8.870</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>T- Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of observations</td>
<td>4946</td>
</tr>
<tr>
<td>Cleared Land (Y1 = 1)</td>
<td>T- Stats</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Slope (degrees)</td>
<td>-0.088</td>
</tr>
<tr>
<td>Elevation (ms.)</td>
<td>-0.001</td>
</tr>
<tr>
<td>Population density1990 (people/km²)</td>
<td>0.003</td>
</tr>
<tr>
<td>Log(cost) (1982)**</td>
<td>-0.191</td>
</tr>
<tr>
<td>Soil and Province Dummies</td>
<td>Not Shown</td>
</tr>
<tr>
<td>Protected Area dummy (1986)</td>
<td>-0.077</td>
</tr>
<tr>
<td>Constant</td>
<td>1.295</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protected Area (Y2 = 1)</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope (Degrees)</td>
<td>0.034</td>
</tr>
<tr>
<td>Elevation (ms.)</td>
<td>0.001</td>
</tr>
<tr>
<td>Population density1990 (people/km²)</td>
<td>0.001</td>
</tr>
<tr>
<td>Log(cost) (1982)</td>
<td>0.192</td>
</tr>
<tr>
<td>Soil and Province Dummies</td>
<td>Not Shown</td>
</tr>
<tr>
<td>Watershed dummy</td>
<td>0.188</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.098</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-3714.7</td>
</tr>
<tr>
<td>No. of observations</td>
<td>4946</td>
</tr>
</tbody>
</table>
Naïve model: Protection has a large effect on preventing deforestation.

After you account for selection bias, in the static model, there is no effect. Protected lands would not have been cleared even if they had not been protected.
LUNCH!

- All examine aspects of WFP’s food security and moderate acute malnutrition (MAM) prevention and treatment programmes, and their impact on nutrition and food security outcomes.
- Commissioned by WFP’s OEV and managed by the International Initiative for Impact Evaluation’s (3ie).
- All 4 Impact Evaluations implemented by different teams.
Background – Where?

Chad, Mali, Niger and Sudan

• Selection criteria: number of beneficiaries, countries with both prevention and treatment interventions, malnutrition figures, mix of programme categories and geographic representation.

• Short-list refined by feasibility for Country Office engagement and timeliness.
Overview - What is given to whom and why?

**Blanket Supplementary Feeding**
- Nutritional supplements and transferring children to treatment

**Targeted Supplementary Feeding**
- Treatment programme

- In-kind and cash transfers in all
- School feeding in **Mali**
- Behaviour Change Communications in **Sudan**
- Food for Assets in **Niger**
Things to note: Questions

- IEs appear to ask similar questions on similar outcomes. The detail underneath is wildly different – different things measured.

- Questions were tailored, due to local contextual and data quality issues.

<table>
<thead>
<tr>
<th>Country</th>
<th>Primary Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niger</td>
<td>What is the impact of receiving different combinations of the components within WFP’s Protracted Relief and Recovery Operation?</td>
</tr>
<tr>
<td>Sudan</td>
<td>What is the impact of different MAM treatment and prevention interventions on the incidence and prevalence of MAM and SAM?</td>
</tr>
<tr>
<td>Chad</td>
<td>What is the difference in impact of MAM prevention on the incidence and prevalence of MAM, when access to MAM treatment is good or poor?</td>
</tr>
<tr>
<td>Mali</td>
<td>What is the impact of conflict and food assistance on child malnutrition and other developmental outcomes?</td>
</tr>
</tbody>
</table>
## Challenges? Of course not – it was easy!

<table>
<thead>
<tr>
<th>Country</th>
<th>Issues</th>
</tr>
</thead>
</table>
| **Niger** | • Baseline not designed for follow up + security risks. Result - high attrition (75%).  
• Everybody in baseline received something – no control.  
• Too small a sample to answer the initial study questions. |
| **Sudan** | • No baseline  
• One intervention did not reach the beneficiaries |
| **Chad** | • No maps – identification plan had to change  
• Targeting agreed so comparison groups from different areas  
• If a malnourished child was identified – she/he should be referred |
| **Mali** |  |
**Things to note: Creative Methodology**

- All use **different** methods and **mixed** methods
- None are ‘conventional’ RCTs  
  - but all have a way to identify the impact  
  - all methods are ‘complicated’

<table>
<thead>
<tr>
<th>Country</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niger</td>
<td></td>
</tr>
</tbody>
</table>
  - Difference-in-differences  
  - Instrumental Variables  
  - Qualitative analysis  
  - Selection correction models |
| Sudan   |  
  - Stepped wedge cluster controlled trial design  
  - Qualitative analysis |
| Chad    |  
  - Analysis of covariates and propensity score matching  
  - Use of qualitative data to inform and interpret results |
| Mali    |  
  - Qual. and Quant. data used to characterise exposure to conflict and humanitarian aid.  
  - Natural experiment, Difference-in-differences and propensity score matching |
Findings?

Niger
- Food for Assets with Prevention or Treatment has an impact on child nutrition and Food for Assets programme is well targeted
- Prevention and treatment programmes less well targeted

Sudan
- No impact on the prevalence, but impact on children-at-risk.
- No change in feeding behaviours and practices as a result of the behavioural intervention.

Chad
- Impact on caloric intake and zinc consumption, and increase in vitamin A availability
- Households receiving two forms of assistance had improved nutrition outcomes.

Mali
Technical difficulties that were resolved creatively

- Difficult to identify a counterfactual? Can often be done creatively.
- High level of attrition – complicates things but can be corrected for.
- Low sample sizes – change in design?
- No baselines – several techniques exist to constructing it either ex-post or artificially.
Lessons from creative IEs: 1. Evaluation Management

Robust management always important but with complex methods even more so:

- Regular comms between evaluation team
  - Changes in evaluation questions
  - Changes in evaluability
  - Unforeseen challenges
Lessons from Creative IEs: 2. Balance of skills

Need a range of skills:

- Rigorous impact evaluation
- Understanding of context and programmes
- Presenting the results and communicating
- Working to timelines
Lessons from creative IEs: 3. Define quality carefully

Agree a common understanding and expectation of “quality”

- High quality methodology
- Integration of gender dimension
- Ethical approvals and management of ethics
- High quality report drafting
- Bespoke communication products
Key takeaways

Creativity is a must!

IEs work in ‘real-life’ and complex settings

Quasi-experiments are a friend of complexity

Ethics is important but not an obstacle

Planning with programme/implementers is crucial

Extra focus on comms is key
AFTERNOON SESSION

DESIGN YOUR OWN IMPACT EVALUATION!
TWO TASKS:

1.) WHAT ARE YOUR EVALUATION QUESTIONS?

2.) WHAT IS YOUR IMPACT EVALUATION DESIGN?
- Jyotsna Puri: jpuri@gcfund.org
- Anna Hentinnen: anna.hentinnen@wfp.org
- Bidisha Barooah: bbarooah@3ieimpact.org