Impact Evaluation in Practice: HISP & HISP+

Design, Estimation, and Power—Concepts and Interpretations

Justin S. Eloriaga

September 25, 2025

What This Evaluation Is About

*Note: This was directly taken from Impact Evaluation in Practice by the World Bank.
Context. Health Insurance Subsidy Program (HISP) lowers out-of-pocket (OOP) expenditures for poor rural households. HISP+ expands coverage to include hospitalization.

Policy Questions.

- Did HISP reduce OOP health spending among eligible households?
- Does HISP+ (with hospitalization coverage) further reduce OOP spending and hospitalization burdens?

Design Opportunity. Staged rollout $\Rightarrow$ randomized assignment at the community level (treatment vs. control) among *eligible* households.

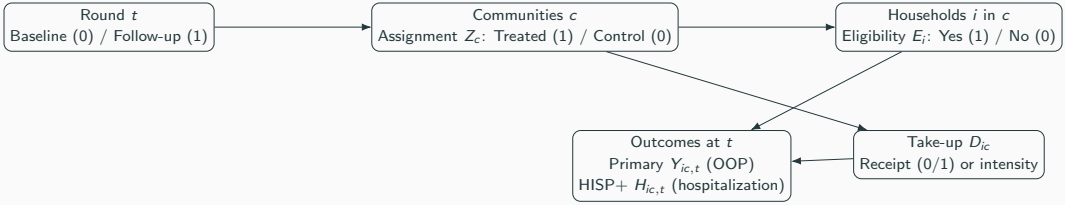
Roadmap

1. Causal questions & outcomes
2. Identification via randomized assignment
3. What to compare at baseline and follow-up
4. Estimands & estimators (ITT, TOT)
5. Power and sample size (MDEs, ICC, clustering)
6. Threats & mitigation, analysis plan

Key Variables & Coding (HISP / HISP+)

Name	Symb.	Type	Level	Coding	What it's for
Round (time)	t	binary/cat.	Time	0 = baseline; 1 = follow-up	Defines pre/post; ANCOVA uses baseline $Y_{t=0}$
Treatment locality (assignment)	Z_c	binary	Community c	1 = treated community; 0 = control	**Randomized**—drives ITT; cluster SEs at c
Eligibility	E_i	binary	HH i	1 = eligible poor rural HH; 0 = ineligible	Primary analysis sample is E_i=1
Take-up / receipt	D_{ic}	binary/cont.	HH in c	1 = received benefit (or intensity)	For TOT via IV: TOT = ITT/($\mathbb{E}[D\|Z{=}1] - \mathbb{E}[D\|Z{=}0]$)
Outcome (primary)	$Y_{ic,t}$	continuous	HH–time	OOP spend (level or log)	Policy target: reduction among eligibles at t=1
Outcome (HISP+)	$H_{ic,t}$	bin./cont.	HH–time	Hosp. incidence or spending	Incremental effect from

Design Schematic: Who, Where, When



Analysis set: E_i=1 (eligible). **Causal contrast:** ITT compares Z_c=1 vs. Z_c=0 at t=1 with cluster SEs.
Efficiency: ANCOVA (post on Z_c + baseline $Y_{t=0}$). **TOT:** IV with Z_c instrumenting D_{ic}.

Units, Eligibility, and Randomization

Units. Households nested in communities.

Eligibility. The program targets poor rural households (*eligible* $= 1$). Ineligible households are not part of the primary target group.

Treatment. Communities randomized to HISP (pilot) vs. control; later, HISP+ pilots hospitalization coverage in a randomized expansion.

Implication. Randomization at the *community* level $\Rightarrow$ cluster-RCT; analyses must account for clustering.

Outcomes & Hypotheses

Primary Outcomes.

- OOP health expenditures (level or log)
- Hospitalization incidence and/or spending (for HISP+)

Secondary Outcomes.

- Preventive care utilization; catastrophic health spending; financial distress

Hypotheses (directional).

- HISP reduces OOP spending for eligible households.
- HISP+ yields *additional* reductions linked to hospitalization risks/costs.

What To Compare (Conceptually)

Baseline (pre): Check that treatment and control groups are similar on observables (balance). No causal claim here—just credibility of randomization.

Follow-up (post): Compare means between treatment and control among *eligible* households (*post-only difference in means*). Randomization ensures unbiased ITT.

Gain scores vs. ANCOVA.

- *Gain scores* (post − pre) remove individual fixed differences but add measurement noise from two waves.
- *ANCOVA* (regress post on treatment + baseline) is typically more efficient when baseline and post are correlated.

Estimands & Estimators

Intent-to-Treat (ITT). Average effect of being assigned to a treated community:

$$\text{ITT} \;=\; \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0],$$

where Z is assignment. Estimated by difference in means (or ANCOVA).

Treatment-on-the-Treated (TOT). Effect of actually receiving the benefit:

$$\text{TOT} \;=\; \frac{\text{ITT}}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]},$$

with D = take-up. Needs first-stage (compliance). In practice: report ITT as policy-relevant; TOT as mechanism/complier effect.

Standard Errors. Cluster at the randomization unit (community).

Interpreting Baseline Differences

If groups look balanced at baseline: supports randomization credibility and internal validity.

If small differences appear: likely due to chance. Use ANCOVA to recoup precision; do not "control your way to balance." The causal claim hinges on *assignment*, not covariate equality.

Interpreting Follow-up Differences

Post-only comparisons (eligible households).

- A negative treatment-control gap in OOP spending $\Rightarrow$ HISP reduces OOP burden.
- For HISP+, additional reductions in hospitalization-related costs indicate the incremental benefit of expanded coverage.

Effect Sizes. Report absolute and percent changes; contextualize against baseline mean and policy costs.

Power & Sample Size: Why It Matters

Goal. Ensure the study can detect a *policy-relevant* minimum detectable effect (MDE) with high probability (power) at a chosen significance level.

Key Inputs.

- Variance of outcome (from baseline or prior data)
- Desired power (e.g., 0.8 or 0.9) and α (e.g., 0.05)
- Allocation ratio (often 1:1)
- Cluster design: average cluster size m and intra-cluster correlation (ICC, ρ)
- Anticipated take-up/noncompliance and attrition

Clustering & Design Effect

Design Effect (DE) for cluster-RCT:

$$\text{DE} \;=\; 1 + (m-1)\rho,$$

where m is average cluster size and ρ the ICC.

Interpretation. Clustering inflates variance; you effectively need DE times as many observations as a simple random sample to maintain precision.

Precision Gains. Blocking/stratifying randomization and using ANCOVA reduce residual variance and can lower the required sample size.

Minimum Detectable Effect (MDE) Intuition

For equal allocation and large-sample z-approximation, a stylized MDE is:

$$\text{MDE} \approx (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sqrt{\frac{2\,\sigma^2 \cdot \text{DE}}{N}},$$

where σ^2 is residual variance (after covariates/blocks), N total sample size, and DE accounts for clustering.

Reading the formula.

- Lower variance, stronger blocking, or larger $N \Rightarrow$ smaller MDE (more sensitive study).
- Higher ICC or larger clusters (same N) $\Rightarrow$ larger MDE (less sensitive).

Noncompliance & Attrition (Power Implications)

Noncompliance. Weakens the first stage; ITT remains unbiased for assignment, but the TOT scales up sampling error. Plan for take-up boosters (information, nudges).

Attrition. Reduces effective N and can bias if differential. Mitigate via tracking, small incentives, short surveys. Pre-specify handling (e.g., inverse probability weights, bounds).

Multiple Outcomes. Prioritize a primary outcome; control the family-wise error rate for secondary outcomes (e.g., Holm–Bonferroni, sharpened q-values) to avoid power dilution.

From HISP to HISP+: What Changes Conceptually

New Margins of Impact.

- Hospitalization coverage introduces a tail-risk channel: fewer catastrophic expenses, smoother consumption.
- Expect larger effects on hospitalization-related spending and potentially on utilization patterns (substitution from informal to formal care).

Design Tweaks.

- Power around rarer outcomes (hospitalization incidence) often requires more clusters or longer follow-up.
- Consider binary vs. continuous specifications and corresponding variance assumptions.

Threats to Validity & Mitigation

Spillovers/SUTVA. Cross-community interactions can dilute contrasts. Use geographic buffers or measure exposure.

Implementation Fidelity. Monitor rollout, document deviations, and analyze ITT + (if credible) TOT.

Selective Attrition. Track aggressively; compare attrition patterns; pre-register handling.

Measurement. Standardize survey modules; pre-define winsorization or transformations (e.g., log OOP).

Analysis Plan (Conceptual)

Primary.

- ITT using ANCOVA for OOP spending (eligible households), clustered SEs
- Pre-registered primary outcome & subgroups (e.g., poorest tercile)

Secondary.

- Hospitalization outcomes (incidence, spending) for HISP+
- Heterogeneity by baseline risk (elderly, chronic conditions)

Diagnostics.

- Balance tables; compliance rates; missingness patterns; robustness checks

How to Read the Results (Interpretations)

Magnitude. Express effects in currency units *and* percent of baseline mean to aid policy interpretation.

Precision. Pair point estimates with 95% CIs; discuss whether CIs exclude policy-relevant thresholds.

External Validity. Clarify the sample frame (poor rural households), delivery model, and key context that condition scalability.

What You'll Hand to the Minister

- **Clear ITT estimates** on OOP and hospitalization outcomes
- **Power-informed recommendations** for scaling HISP+ (clusters, households per cluster)
- **Risks & safeguards** (spillovers, take-up, attrition) with concrete monitoring steps
- **Cost–effectiveness framing** (if feasible): effect per $ of subsidy

Appendix: Quick Formula Shelf

Design Effect (clusters). $\mathrm{DE} = 1 + (m-1)\rho$

Stylized MDE (equal allocation).

$$\mathrm{MDE} \approx (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sqrt{\frac{2\,\sigma^2 \cdot \mathrm{DE}}{N}}$$

ANCOVA efficiency gain (intuition). Gains grow with correlation between baseline and follow-up.

TOT from ITT with compliance p_c. $\mathrm{TOT} \approx \mathrm{ITT}/p_c$