Empirical Methods for Policy Evaluation

Second Part

Matteo Bobba

Toulouse School of Economics (TSE)

TSE PhD Program (MRes)
Fall 2024

Outline and Readings for this Section (3 Classes)

1. Statistical analysis of field experiments (RCTs)
   - SUTVA, assignment mechanisms and randomization designs (Imbens-Rubin, Ch 1,3,4)
   - Randomization inference (IR, Ch 5)
   - Stratified RCTs and Clustered RCTs (IR, Ch 9; Athey-Imbens, Sections 7-8)
2. RCTs and Risk Sharing Models
   - **Meghir, Mobarack, Mommaerts, and Morten (ReStud, 2022)**

SUTVA, Assignment Mechanisms and Randomization Designs

Causal Inference as a Missing Data Problem

- Population of units, indexed by $i = 1, ..., N$
- Treatment indicator W_i taking values 0 and 1
- For each unit $i \in \{1, ..., N\}$ there is one realized (and possibly observed) outcome and one missing potential outcome

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) \text{ if } W_i = 0 \\ Y_i(1) \text{ if } W_i = 1 \end{cases}$$

$$Y_i^{\text{miss}} = Y_i(1 - W_i) = \begin{cases} Y_i(1) \text{ if } W_i = 0 \\ Y_i(0) \text{ if } W_i = 1 \end{cases}$$

- Unit-level causal effect $Y_i(1) - Y_i(0)$ is unobserved

The Stable Unit Treatment Value Assumption (SUTVA)

- To estimate the causal effect for any particular unit, we will generally need to predict, or impute, the missing potential outcome
- To do so, we need the following assumption:

  The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

- Denote $W_{-i} = (W_j)_{j \neq i}$ as the treatment status of all other observations in the sample or the population except i
- SUTVA requires that

$$(Y_i(1), Y_i(0)) \amalg W_{-i}$$

Two Parts of SUTVA

1. **No interference**. Example of possible violations include:
   - Fertilizer in one plot may affect yields in contiguous plots
   - Wages after job training may be affected by the number of people trained
   - Immunization efficacy may depend on the number of people immunized
2. **No hidden variations of treatments**. Example of possible violations include:
   - Different efficacies of treatments
   - Differences in the method of administering the treatment

Assignment Mechanism

- There is a population of N units, and a set $\mathbb{W} = \{0,1\}^N$ of N-vectors with all elements equal to 0 or 1
- The assignment mechanism is a function $P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1)) \in [0,1]$ such that

$$\sum_{\mathbf{W} \in \{0,1\}^N} P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1)) = 1$$

- $P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1))$ is the probability that a particular value for the joint assignment will occur (out of 2^N possible assignment vectors)
- The unit-level assignment probability is

$$p_i(\mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{W}: W_i = 1} P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1))$$

Restrictions on the Assignment Mechanism

1. **Individualistic**: requires the dependence of the treatment assignment for unit i to exclusively depend on the outcomes and assignment of that unit

$$p_i(\mathbf{Y}(0), \mathbf{Y}(1)) = q(Y_i(0), Y_i(1)),\ q(\cdot) \in [0, 1]$$

2. **Probabilistic**: requires every unit to have positive probability of being assigned to treatment level 0 and to treatment level 1

$$0 < p_i(\mathbf{Y}(0), \mathbf{Y}(1)) < 1$$

3. **Unconfounded**: requires that it does not depend on potential outcomes

$$P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{W}|\mathbf{Y}'(0), \mathbf{Y}'(1)) = P(\mathbf{W})$$

Restrictions on the Assignment Mechanism

- The combination of individualistic and unconfounded assignment implies that the assignment mechanism can be re-written as:

$$P(\mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q^{W_i} (1-q)^{1-W_i}$$

- The constant c ensures that the probabilities add to unity
- An assignment mechanism that satisfies the three restrictions is called **regular assignment mechanism**

Randomized Experiments Vs. Observational Studies

- A regular assignment mechanism in which the functional form of the assignment is known corresponds to a **(classical) randomized experiment**
- An assignment mechanism corresponds to an **observational study** if the functional form of the assignment is unknown
  - Regular: adjusting for differences in covariates between treated and control units is enough to draw valid causal inferences
  - Latently regular: assignment to treatment may differ for some units from the receipt of treatment. To conduct inference in such settings, it is often useful to invoke additional conditions (exclusion restrictions)

Taxonomy of Randomization Designs

- Randomization designs can be characterized by the restrictions on the assignment vectors $\mathbf{W}$ with positive probabilities, $\mathbb{W}^+$
  1. Completely randomized experiments
  2. Stratified randomized experiments
  3. Paired randomized experiments
  4. Clustered randomized experiments

A Prelude: Coin Tossing

- Unit-level probabilities are all equal to 0.5

$$P(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1)) = 0.5^N$$

- Here $\mathbb{W}^+ = \{0,1\}^N = \mathbb{W}$
- More generally, with probability of assignment to treatment $\neq 0.5$

$$P(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1)) = q^{N_t}(1-q)^{N_c}$$

- One disadvantage is that there is a positive probability (small, and essentially zero in large samples) that all units will receive the same treatment
- More generally, there is no way to ensure that there are "enough" treated and control units under each assignment

Completely Randomized Experiments

- A completely randomized experiment is a classical randomized experiment with an assignment mechanism satisfying

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i=1}^{N} W_i = N_t \right\}.$$

- Given a population of size N, we draw N_t units at random to receive the treatment, such that $1 \leq N_t \leq N-1$
- Each unit has probability $q = \frac{N_t}{N}$ of receive the treatment, and the number of possible assignment vectors is $\binom{N}{N_t}$

Stratified Randomized Experiments

- The population of units is first partitioned into blocks or strata $B_i = B(\mathbf{X}_i)$, where $\mathbf{X}_i$ are covariates thought to be predictive of potential outcomes
- Within each block, we conduct a completely randomized experiment, with assignments independent across blocks
- A stratified randomized experiment with J blocks is a classical randomized experiment with an assignment mechanism satisfying

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i:B_i=j}^{N} W_i = N_t(j) \right\}.$$

- Randomizing within the strata will lead to more precise inferences by eliminating the possibility that all or most units of a certain type, as defined by the blocks, are assigned to the same treatment status

Pairwise Randomized Experiments

- A paired randomized experiment is a stratified randomized experiment with $N(j) = 2$ and $N_t(j) = 1$ for $j = 1, ..., N/2$, so that

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i:B_i=j}^{N} W_i = 1 \right\}.$$

- In this design, each unit has probability 0.5 of being assigned to the treatment group
- It is an extreme version of the randomized block experiment in which there are exactly two units within each block

Number of Possible Values for the Assignment Vector

Type of Experiment and Design	Number of Possible Assignments Cardinality of $\mathbb{W}^+$	Number of Units (N) in Sample			
		4	8	16	32
Bernoulli trial	2^N	16	256	65,536	4.2×10^9
Completely randomized experiment	$\binom{N}{N/2}$	6	70	12,870	0.6×10^9
Stratified randomized experiment	$\binom{N/2}{N/4}^2$	4	36	4,900	0.2×10^9
Paired randomized experiment	$2^{N/2}$	4	16	256	65,536

Clustered Randomized Experiments

- $G_{ig} = G(\mathbf{X}_i)$ is an indicator for unit i belonging to group of units (cluster) g
- A clustered randomized experiment is a completely randomized experiment in which the assignment mechanisms concerns clusters rather than units

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{g=1}^{G} \overline{W}_g = G_t \right\}.$$

- $\overline{W}_g$ is the (average) value of W_i for units in cluster g
- This design may be motivated by concerns that there are (local) interactions between units

Randomization Inference

Two Approaches for Inference (Not Just in RCTs)

1. The usual approach (aka asymptotic inference)
   - Relies on (semi-)parametric models for the conditional mean of observed outcomes
   - Treatment assignment is fixed and observed outcomes vary through random sampling from a population of units
   - Derives/approximates the distribution of the test statistic through large-sample properties
2. Fisher's exact p-values (aka randomization inference)
   - Fully nonparametric (no restrictions on the distribution of the potential outcomes)
   - Potential outcomes are fixed and the treatment assignments are the sole source of randomness
   - The assignment mechanism determines the distribution of the test statistic

A Simple Example

Unit	Potential Outcomes				
	Cough Frequency (cfa)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (cfp)	Y_i^{obs} (cfa)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

A Simple Example (Continued)

- The p-value for the **sharp** null hypothesis that the treatment had no effect on coughing outcomes is

$$H_0 \ : \ Y_i(0) = Y_i(1) \ \forall i = 1, ..., 6.$$

- The average null hypothesis is weaker than the sharp null hypothesis
- Under the null hypothesis, all the missing values in potential outcomes can be inferred from the observed outcomes
- The test statistics is

$$\begin{aligned} T(W, Y^{\text{obs}}) &= \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \\ &= (Y_1^{\text{obs}} + Y_2^{\text{obs}} + Y_3^{\text{obs}})/3 - (Y_4^{\text{obs}} + Y_5^{\text{obs}} + Y_6^{\text{obs}})/3 \\ &= 8/3 - 5/3 = 1.00 \end{aligned}$$

A Simple Example (Continued)

- Under the null hypothesis, we can calculate the value of the test statistic under each of the $\binom{6}{3} = 20$ permutations of the vector of treatment assignments, W
- E.g. instead of $W^{\text{obs}} = (1, 1, 1, 0, 0, 0)$ take $\bar{W} = (0, 1, 1, 0, 0, 1)$
- No change in observed outcomes since under the null

$$Y_i(0) = Y_i(1) = Y_i^{\text{obs}}$$

- The value of the test statistic may change

$$\begin{aligned} T(\bar{W}, Y^{\text{obs}}) &= (Y_2^{\text{obs}} + Y_3^{\text{obs}} + Y_6^{\text{obs}})/3 - (Y_1^{\text{obs}} + Y_4^{\text{obs}} + Y_5^{\text{obs}})/3 \\ &= 6/3 - 7/3 = -0.33 \end{aligned}$$

A Simple Example (Continued)

- For each vector of assignments, we calculate the corresponding value of the statistic
- Under random assignment, each assignment vector has prior probability $1/20$
- How unusual or extreme is $T(W, Y^{\text{obs}}) = 1.00$ assuming the null hypothesis is true?
- There are $16/20$ assignment vectors with at least a difference in absolute value of 1.00: p-value $= 0.80$
- Under the null hypothesis of absolutely no effect, the observed difference could, therefore, well be due to chance

Computation of p-values

- The p-value calculations of the previous example ($N = 6$) have been exact
  - Recall that the number of distinct values of the treatment vector is $\binom{N_c+N_t}{N_t}$
  - For instance, if $N = 100$ and $q = 0.5$ then $dim(\mathbb{W}^+) = e^{29}$
- We thus need to rely on numerical approximations to calculate the p-value
  - Draw an N-dimensional vector with N_c zeros and N_t ones from $\mathbb{W}^+$
  - Repeat this process $K - 1$ times and approximate the p-value by:

$$\widehat{p} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{T^{\text{dif},k} \geq T^{\text{dif,obs}}}$$

  - With $K > 1,000$ each assignment vector has a similar probability of being drawn with or without replacement

Software Implementation

- Hess (Stata Journal, 2017)
  - `ritest` package
- Bowers et al. (Cran R project, 2024)
  - `RItools` package

Stratified Randomized Experiments

What's the Point of Stratification?

- Units are grouped together according to some pre-treatment characteristics into strata
- The stratification rules out substantial imbalances in the covariate distributions in the two treatment groups that could arise by chance in a completely randomized experiment
- Within each stratum, a completely randomized experiment is conducted
- The interest is not about hypotheses or treatment effects within a single stratum, but rather it is about hypotheses and treatment effects across all strata

The Benefits of Stratification

- Consider a case with one covariate $G_i \in \{f, m\}$, with $p(G_i = f) = p$
- Completely randomized design: $N_t = qN$ and $N_c = (1-q)N$:

$$\hat{\tau}^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$$

$$\mathbb{V}(\hat{\tau}^{\text{dif}}) = \frac{\sigma_t^2}{N_t} + \frac{\sigma_c^2}{N_c}$$

The Benefits of Stratification

- Stratified design, two subsamples:
  1. $N_t(f) = pqN$ and $N_c(f) = p(1-q)N$
  2. $N_t(m) = (1-p)qN$ and $N_c(m) = (1-p)(1-q)N$

$$\hat{\tau}^{\text{strat}} = p\hat{\tau}(f) + (1-p)\hat{\tau}(m)$$

$$\mathbb{V}(\hat{\tau}^{\text{strat}}) = \frac{p}{N}\left(\frac{\sigma_t^2(f)}{p} + \frac{\sigma_c^2(f)}{1-p}\right) + \frac{1-p}{N}\left(\frac{\sigma_t^2(m)}{p} + \frac{\sigma_c^2(m)}{1-p}\right)$$

- Hence, the difference in the two variances is

$$\mathbb{V}(\hat{\tau}^{\text{dif}}) - \mathbb{V}(\hat{\tau}^{\text{strat}}) = \frac{p(1-p)}{N}((\mu_c(f) - \mu_c(m))^2 + (\mu_t(f) - \mu_t(m))^2) \geq 0$$

An Alternative to Stratification: Re-randomization

- What if after the random draw some (important) covariates are unbalanced?
- Randomize many times and select the draw that achieves better balance
  - E.g. pick the draw with the minimum maximum t-stat
- Preferred over stratification when one needs to ensure balance among several variables
- Inference is tricky as not every combinations of allocation is ex-post equally probable
- p-values need to be adjusted for the re-randomization, although ignoring the adjustment simply leads to conservative p-values

Re-randomization: Example

- $N = 100$ individuals, with 50 women and 50 men
- Completely randomize 60 individuals to treatment, then reject and re-randomize many times until we get 30 men and 30 women assigned to treatment
- This is a stratified experiment
- With the important difference that to make correct inference we would need to know the entire sequence of assignment vectors that led to the final assignment

The Structure of Stratified Randomized Experiments

- Let J be the number of strata/blocks, and $N(j), N_c(j), N_t(j)$
- Let $G_i \in \{1, ..., J\}$ be the stratum for unit i
- Let $B_i(j) = \mathbf{1}_{G_i=j}$ be the stratum indicator for unit i
- Within stratum j there are $\binom{N(j)}{N_t(j)}$ possible assignments, so that the assignment mechanism is

$$P(\mathbf{W}|\mathbf{B}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \text{ for } \mathbf{W} \in \mathbb{W}^+$$

- where $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} \mid \sum_{i=1}^{N} B_i(j) \cdot W_i = N_t(j) \text{ for } j = 1, ..., J\}$

Example: Tennessee Project Star

School/ Stratum	No. of Classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	−0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, −0.202
3	5	−0.496, 0.225	0.341, 0.561, −0.059
4	4	−1.104, −0.956	−0.024, −0.450
5	4	−0.126, 0.106	−0.258, −0.083
6	4	−0.597, −0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	−0.934, −0.633	−0.870, −0.496, −0.444, 0.392
9	4	−0.891, −0.856	−0.568, −1.189
10	4	−0.473, −0.807	−0.727, −0.580
11	4	−0.383, 0.313	−0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	−0.434, −0.293	−0.545, 0.234
16	4	0.355, −0.130	−0.240, −0.150
Average		−0.13	0.09
(S.D.)		(0.56)	(0.61)

Randomization Inference for Stratified Experiments

- Let us focus on the sharp null hypothesis that all treatment effects are zero:

$$H_0 \ : \ Y_i(1) = Y_i(0) \ \forall i = 1, 2, ..., N.$$

- Define average observed outcomes in stratum j as

$$\overline{Y}_t^{\text{obs}}(j) = \frac{1}{N_t(j)} \sum_{i:G_i=j} W_i Y_i^{\text{obs}}$$

$$\overline{Y}_c^{\text{obs}}(j) = \frac{1}{N_c(j)} \sum_{i:G_i=j} (1 - W_i) Y_i^{\text{obs}}$$

- Strata-level average assignment probability is

$$e(j) = \frac{N_t(j)}{N(j)}$$

Test Statistics

- Within-stratum test statistic

$$T^{\text{dif}}(j) = \left|\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)\right|$$

- Not very informative as we are interested in treatment effects across all strata
- Linear combination of the within-stratum statistics

$$T^{\text{dif},\lambda_{RSS}} = \left|\sum_{j=1}^{J} \frac{N_j}{N}\left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)\right)\right|$$

- Need $e(j) = N_t(j)/N(j)$ to substantially vary across strata j for the test to have power over the standard $T^{\text{dif}} = \left|\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right|$

Randomization Inference of the Tennessee Project Star

- $B_i(j), i = 1, ..., 68$ (class-level data)
- Total number of possible assignments of teachers to class type is a very large number
  - 13 Schools with two classes in each group: $\binom{4}{2} = 6$
  - 2 Schools with three small classes and two regular classes: $\binom{5}{2} = 10$
  - 1 School with four small classes and two regular classes: $\binom{6}{2} = 15$

$$H_0 \ : \ Y_i(1) = Y_i(0) \ \forall i = 1, 2, ..., 68.$$

- $T^{\text{dif}} = \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid = 0.224$, with $p = 0.034$
- $T^{\text{dif},\lambda_{RSS}} = \mid \sum_{j=1}^{J} \frac{N_j}{N} (\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)) \mid = 0.241$, with $p = 0.023$

Regression Analysis

$$Y_i^{\text{obs}} = \tau W_i + \sum_{j=1}^{J} \beta(j) B_i(j) + \epsilon_i$$

- Recall that $B_i(j) = \mathbf{1}_{G_i=j}$ is the stratum indicator for unit i
- In general $\widehat{\tau}^{\text{ols}}$ is not a consistent estimator of τ. It estimates a weighted average of the within-stratum average effects

$$\tau_\omega = \frac{\sum_{j=1}^{J} \omega(j)\tau(j)}{\sum_{j=1}^{J} \omega(j)}$$

- $\omega(j) = \frac{N_j}{N} \frac{N_t(j)}{N(j)} \frac{N(j)-N_t(j)}{N(j)} = q(j)e(j)(1-e(j))$
- $\tau(j) = \mathbb{E}[Y_i(1) - Y_i(0) \mid B_i(j) = 1]$

Asymptotic Inference

- The estimated variance of the weighted average treatment effect τ_ω is

$$\widehat{V}^{\text{strata}} = \frac{\sum_{i=1}^{N} \widehat{\epsilon}_i^2 \cdot \left(W_i - \sum_{j=1}^{J} q(j) B_i(j)\right)^2}{\left(\sum_{j=1}^{J} \omega(j)\right)^2}$$

- The weights $\omega(j)$ are proportional to the precision of the estimator of the within-stratum treatment effects

$$\widehat{\tau}^{\text{dif}}(j) = \overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)$$

- Sampling variance of $\widehat{\tau}^{\text{dif}}(j)$ is $(\sigma^2/N) \cdot (q(j)e(j)(1-e(j)))^{-1}$

Fully-interacted Model

$$Y_i^{\text{obs}} = \tau W_i \frac{B_i(j)}{N(j)/N} + \sum_{j=1}^{J} \beta(j) B_i(j) + \sum_{j=1}^{J-1} \gamma(j) W_i \left(B_i(j) - B_i(J) \frac{N(j)}{N} \right) + \epsilon_i$$

- In this case OLS converges to the (population-)average treatment effect

$$\widehat{\tau}^{\text{ols,inter}} = \tau$$

- With estimated asymptotic variance equal to

$$\widehat{V}^{\text{strata,inter}} = \sum_{i=1}^{N} q(j)^2 \cdot \left(\frac{\sigma_c^2(j)}{q(j)(1-e(j))} + \frac{\sigma_t^2(j)}{q(j)e(j)} \right)$$

Regression Analysis of the Tennessee Project Star

- The point estimate of τ in the standard model is
  - $\widehat{\tau}^{\text{ols}} = 0.238$ $(\widehat{s.e.} = 0.103)$
- If there is variation in the effect of the class size across schools (i.e. $\tau(j) \neq \tau(j') \, \forall j \neq j'$), then this estimator is not consistent for the average effect of the treatment in the population
- The point estimate of τ in the fully-interacted model is
  - $\widehat{\tau}^{\text{ols,inter}} = 0.241$ $(\widehat{s.e.} = 0.095)$
- The two estimates for the average effect are close, consistent with limited heterogeneity in the treatment effects across strata

Clustered Randomized Experiments

What's the Point of Clustering?

- Instead of assigning treatments at the unit level, in this setting the population is first partitioned into a number of clusters
- Then all units in a cluster are assigned to the same treatment level
- Given a fixed sample size, this design is in general not as efficient as a completely randomized design or a stratified randomized design
- There may be interference between units at the unit-level violating SUTVA
- In many cases it is easier to sample units at the cluster level

Unit-level Vs. Cluster-level

- Cluster-level analysis is more transparent and more directly linked to the randomization framework
  - Inference at cluster-level is more precise when there are a few large clusters and many small clusters (e.g., clusters are geographical units, such as states or towns)
  - Inference at the unit-level is complicated in this case because many units will be in the same treatment group
- Unit-level is more flexible, as it allows to incorporate individual-level covariates and this may improve efficiency
  - When number of units per cluster is similar (e.g., in educational settings where the clusters are schools or classrooms)

The Structure of Clustered Experiments

- Let G_{ig} be a binary indicator that unit i belongs to cluster $g = 1, ..., G$
- The number of units in cluster g is $N_g = \sum_{i=1}^{N} G_{ig}$, so that N_g/N is the share of cluster g in the sample
- $\overline{W}_g \in \{0, 1\}$ is the (average) value of the treatment assignment for all units in cluster g
- G is the total number of clusters, with G_t the number of treated cluster and $G_c = G - G_t$ the number of control clusters
- The assignment mechanism is

$$P(\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \binom{G}{G_t}^{-1}$$

- where $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} \mid \sum_{g=1}^{G} \overline{W}_g = G_t\}$

Example: The *Progresa* Program

- Educational grants to mothers to encourage children's school attendance
- Clustered RCT during the roll-out of the program in rural areas
  - 506 villages among those eligible to receive the program
  - 320 early treatment and 186 late treatment (control)
- Rich data collected at the individual/HH level for both eligible and non-eligible HHs in each village
  - Approx. 30,000 program eligible children
  - About 50-100 HHs per village

Estimands

- The choice of estimand depends on the choice of the unit of analysis
- For analysis at the unit-level, a natural estimand is the population average treatment effect

$$\tau^{\text{pop}} = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

- For analysis at the cluster-level, we instead consider the (unweighted) average of the within-cluster average effect

$$\tau^{\text{C}} = \frac{1}{G} \sum_{g=1}^{G} \tau_g, \quad \text{where } \tau_g = \frac{1}{N_g} \sum_{i:G_{ig}=1}^{N} (Y_i(1) - Y_i(0))$$

Randomization Inference

- The usual statistic for unit-level analysis

$$T^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} = \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c}$$

- The equivalent statistic for cluster-level analysis

$$T^{\text{dif,C}} = \frac{1}{G_t} \sum_{g:\overline{W}_g=1} \overline{Y}_g^{\text{obs}} - \frac{1}{G_c} \sum_{g:\overline{W}_g=0} \overline{Y}_g^{\text{obs}}$$

- As usual, consider all permutations (or a random subset) of the vector $\overline{W}_g$ and compute associated statistics and p-values accordingly

Randomization Inference of *Progresa*

- Children-level analysis on school enrollment (pre-program year 1997)

$$T^{\text{dif}} = 0.0075, \quad p\text{-value} = 0.400$$

- Children-level analysis on school enrollment (program year 1998)

$$T^{\text{dif}} = 0.0388, \quad p\text{-value} < 0.001$$

- Village-level analysis on school enrollment (program year 1998)

$$T^{\text{dif,C}} = 0.0234, \quad p\text{-value} = 0.0120$$

Regression Analysis: Unit-Level

- In unit-level analysis, we estimate the following regression

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \epsilon_i$$

- Let $\widehat{\epsilon}_i = Y_i^{\text{obs}} - \widehat{\alpha} - \widehat{\tau} W_i$ be the residual, then the estimator of the variance of τ^{ols} is:

$$\widehat{V}^{\text{clust}} = \left\{ \sum_{i=1}^{N} \begin{pmatrix} 1 & W_i \\ W_i & W_i \end{pmatrix} \right\}^{-1} \left\{ \sum_{g=1}^{G} \sum_{i:G_{ig}=1} \begin{pmatrix} \hat{\epsilon}_i \\ W_i \hat{\epsilon}_i \end{pmatrix} \sum_{i:G_{ig}=1} \begin{pmatrix} \hat{\epsilon}_i \\ W_i \hat{\epsilon}_i \end{pmatrix}' \right\} \left\{ \sum_{i=1}^{N} \begin{pmatrix} 1 & W_i \\ W_i & W_i \end{pmatrix} \right\}^{-1}$$

Regression Analysis: Cluster-Level

- In cluster-level analysis, consider the following regression

$$\overline{Y}_g^{\text{obs}} = \alpha + \tau \overline{W}_g + \eta_g$$

- The estimator of the variance of τ^{ols} is the usual one:

$$\widehat{V} = \frac{\sum_{g=1}^{G} \widehat{\eta}_g^2}{\sum_{g=1}^{G} \left(\overline{W}_g - \overline{W}\right)^2} = \widehat{\sigma}^2 \left\{ \frac{1}{G_t} + \frac{1}{G_c} \right\}$$

Regression Analysis of *Progresa*

- Children-level analysis on school enrollment (pre-program year 1997)

$$\widehat{\tau}^{\text{ols}} = 0.0075 \ (\widehat{s.e.} = 0.0091)$$

- Children-level analysis on school enrollment (program year 1998)

$$\widehat{\tau}^{\text{ols}} = 0.0388 \ (\widehat{s.e.} = 0.0104)$$

- Village-level analysis on school enrollment (program year 1998)

$$\widehat{\tau}^{\text{ols}} = 0.0234 \ (\widehat{s.e.} = 0.0092)$$

Meghir, Mobarack, Mommaerts, and Morten (Restud, 2022)

Migration and Informal Insurance: Evidence from a Randomized Controlled Trial and a Structural Model

- Migration subsidies in bad times crowd in informal insurance
- A joint model of migration and informal risk sharing explains why
- Quantify the welfare effect of the migration subsidies
- Conduct counterfactual experiments to evaluate different policy levers

The Migration Experiment (Bryan et al, ECMA 2014)

- The migration subsidy treatment was randomized at the village level
  - 68 treated villages and 32 control villages
  - Random sampling of 19 eligible (poor) HHs in each village
- Rich HH-level data collected before, during, and after the intervention
  - Annual income (home, city, and total)
  - Consumption (food, and non-food)
  - Savings and transfers (received and given)

Experimental Evidence on Financial Transfers

	Treatment effect	Control mean
Willingness to help		
Community member would help you	0.030	0.85
	(0.020)	
... and you would ask for help	0.025	0.83
	(0.020)	
Community member would ask you for help	0.109***	0.57
	(0.033)	
... and you would help them	0.109***	0.53
	(0.032)	
Actual transfers		
Receive any transfer from community member	−0.024	0.57
	(0.022)	
Amount, if any transfer received (Tk)	1,821***	4808
	(678)	
Give any transfer to community member	0.036**	0.15
	(0.018)	
Amount, if any transfer given (Tk)	1,310**	2001
	(558)	

Notes: The sample includes households from the 2011 survey. Each cell is a separate regression of the effect of treatment on whether the source denoted in the row would behave as described. Each regression also controls for upazila (county). Standard errors, clustered by village, are in parentheses, and the mean of the control group is in square brackets. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Experimental Evidence on Financial Transfers

- There is a strong norm that households would provide and receive financial assistance among each other
- The migration experiment significantly increased the willingness of households to participate in these arrangements as well as actual transfers between households
- This increase is not limited to households that were induced to migrate
- Migration strengthened informal relationships within a village more broadly

ITT Analysis on the Transmission of Income to Consumption

	Round 4				Diff in Diff	
	(1)	(2)	(3)	(4)	(5)	(6)
Log income (round 4)	0.157***	0.169***	0.130***	0.140***	0.112**	0.109**
	(0.027)	(0.028)	(0.028)	(0.029)	(0.054)	(0.046)
Treatment effect on log income	−0.073***	−0.066**	−0.072***	−0.061**	−0.077	−0.099**
	(0.027)	(0.027)	(0.027)	(0.026)	(0.061)	(0.046)
Village-round FE	X	X	X	X	X	X
Household FE					X	X
Household head controls		X		X		
Resource controls			X	X		
Includes baseline					X	X
Includes 2013						X
Observations	1857	1857	1857	1857	2166	4371
R^2	0.186	0.221	0.217	0.267	0.791	0.721

Notes: Table presents coefficients of the effect of log annual per-capita income on log annual per capita consumption and the interaction with treatment (β_0 and β_1 from equation 2). All models control for village fixed effects and all other interactions between treatment and log income as well as log income interacted with 2011 treatments. Column (2) additionally adds household head controls, column (3) adds household resource controls, and column (4) adds both household head and resource controls. Columns (5) and (6) show the result of difference-in-difference specifications, with the first coefficient shown being the interaction between log income and round 4, and the second coefficient shown being the interaction between treatment, log income, and a post-experiment indicator. Column (5) includes baseline data, and column (6) includes both baseline and 2013 data, and both include household fixed effects. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

ITT Analysis on the Transmission of Income to Consumption

- Migration treatment reduced the effect of HH income on consumption by over 7 pp, or 40% of the exposure in the control group
- Alternative specifications suggest results are not driven by differences in permanent income (check also Appendix A.1 with derivations of the bias)
- Similar effects for non-migrant HHs
- No effect on savings (and mean savings are also very small)

What is the Rationale of the Model?

- Different pieces of evidence all point to the experiment causing a substantial improvement in the willingness and ability to share risk in treatment villages
- Why did this happen?
- Migration subsidies interact with the underlying risk environment to generate either positive or negative spillovers
  - Subsidies increase social return to migration (crowd-in)
  - If migration is relatively safe, then the migrant may not need the safety net provided by the network (crowd-out)
- Welfare effects of policy are heavily context dependent

A Joint Model of Risk Sharing and Migration

- Starting point is Morten (JPE, 2019)
  - HHs make migration decisions taking into account the returns to migrating, including risk-sharing transfers
  - Risk sharing is constrained by limited commitment frictions (Ligon et al, ReStud 2002)
- Authors extend this framework to allow for a migration asset (i.e. a job connection at destination)
  - Migrants tend to return to the same employer
  - A one-time experiment led to persistent effects on migration
- This generates an additional motive for migrating: by allowing individuals to update their migration asset it provides a dynamic payoff for the future

A Joint Model of Risk Sharing and Migration

Model period	Activities	State of the world
Start of period t —	Start-of-period job contacts (A_t)	$h_t = \{s_t, A_t\}$
Before migration —	Observe village state (s_t)	
	Migration decision (j_t)	
After migration —	Exogenous creation of job contacts ($\hat{A}_t$)	$\hat{h}_t = \{h_t, j_t, q_t, \hat{A}_t(h_t, j_t)\}$
	Observe migration state (q_t)	
	Make risk-sharing transfers	
	Consume	
	Return to village	
	Exogenous separation of job contacts (A_{t+1})	
Start of period $t+1$ —	Start-of-period job contacts (A_{t+1})	$h_{t+1} = \{s_{t+1}, A_{t+1}\}$
	Observe village state (s_{t+1})	

Figure 1: Model timeline

Optimization Problem without Risk Sharing

- The before-migration value is the expected utility at the time the household is deciding whether or not to migrate:

$$\Omega^i(h) = \max_{\mathbb{I}^i} \sum_{\hat{h}} \pi_{\hat{h}|h,\mathbb{I}^i} \left[u(\hat{y}^i(\hat{h}, d^{\text{fin}})) - \mathbb{I}^i(\hat{h}) d^{\text{utility}} + \beta \sum_{h'} \pi_{h'|\hat{h}} \Omega^i(h') \right]$$

- The after-migration value is the expected utility once the migration decision has been made and the household learns if it has a job contact, and then learns the state of the world in the destination:

$$\hat{\Omega}^i(h) = u(\hat{y}^i(\hat{h}, d^{\text{fin}})) - \mathbb{I}^i(\hat{h}) d^{\text{utility}} + \beta \sum_{h'} \pi_{h'|\hat{h}} \Omega^i(h')$$

- $\Omega^i(h)$ and $\hat{\Omega}^i(h)$ determine the credible threat points in the full endogenous risk-sharing model

Optimization Problem with Risk Sharing

- The optimization problem involves migration choices of both HHs and the net transfer from HH1 one to HH2, τ, to maximize total welfare
- The after-migration problem is the following:

$$\widehat{V}(\hat{h}, \widehat{U}(\hat{h})) = \max_{\tau(\hat{h}, d^{\text{fin}}), U(h')} u(\hat{y}^2(\hat{h}, d^{\text{fin}}) + \tau(\hat{h}, d^{\text{fin}})) - \mathbb{I}^i(\hat{h}) d^{\text{utility}} + \beta \sum_{h'} \pi_{h'|\hat{h}} V(h', U(h'))$$

- Subject to a promise-keeping constraint for HH1

$$(\widehat{\lambda}_{\hat{h}}): \; u(\hat{y}^1(\hat{h}, d^{\text{fin}}) - \tau(\hat{h}, d^{\text{fin}})) - \mathbb{I}^i(\hat{h}) d^{\text{utility}} + \beta \sum_{h'} \pi_{h'|\hat{h}} V(h', U(h')) \geq \widehat{U}(\hat{h})$$

- And incentive compatibility constraints in period 1 for both HHs associated to autarky case (with LMs $\phi^1_{h',\hat{h}}$ and $\phi^2_{h',\hat{h}}$)

Optimization Problem with Risk Sharing

- FOCs yield (see Appendix A.2)

$$\frac{u_1(c^2(\hat{h}))}{u_1(c^1(\hat{h}))} = \hat{\lambda}_{\hat{h}}$$

$$V_2(h', U(h')) = \widehat{\lambda}_{\hat{h}} \frac{(1 + \phi^1_{h',\hat{h}})}{(1 + \phi^2_{h',\hat{h}})}$$

- The envelope condition yields

$$\widehat{V}_2(\hat{h}, \widehat{U}(\hat{h})) = \hat{\lambda}_{\hat{h}}$$

Optimization Problem with Risk Sharing

- The before-migration problem is the following:

$$V(h, U(h)) = \max_{j} \left\{ \max_{\widehat{U}(\hat{h})} \left[\sum_{\hat{h}} \pi_{\hat{h}|h,j} \hat{V}(\widehat{h}, \widehat{U}(\hat{h})) \right] \right\}$$

- Subject to promise-keeping that needs to hold for each migration outcome with LM λ_j, and incentive compatibility constraints with LMs $\alpha^1_{\hat{h}}$ and $\alpha^2_{\hat{h}}$
- FOC and envelope condition yield:

$$\widehat{V}_2(\hat{h}, U(\hat{h})) = -\lambda_j \frac{(1 + \alpha^1_{\hat{h}})}{(1 + \alpha^2_{\hat{h}})} \ \forall \hat{h}$$

$$V_2(h, U(h)) = -\lambda_j \ \forall j$$

Updating Rules

- The optimization problem implies that the Pareto weight follows a simple updating rule:

$$\frac{u_1(c^2(\hat{h}_{t+1}))/u_1(c^1(\hat{h}_{t+1}))}{u_1(c^2(\hat{h}_t))/u_1(c^1(\hat{h}_t))} = \frac{\hat{\lambda}_{t+1}}{\hat{\lambda}_t} = \frac{(1+\phi^1_{h_{t+1},\hat{h}_t})(1+\alpha^1_{h_{t+1}})}{(1+\phi^2_{h_{t+1},\hat{h}_t})(1+\alpha^2_{h_{t+1}})}$$

- If neither HH is constrained ($\alpha^1_{h_{t+1}} = \alpha^2_{h_{t+1}} = \phi^1_{h_{t+1}} = \phi^2_{h_{t+1}} = 0$), then the growth rate of relative marginal utility is zero
- Otherwise, the change in marginal utility (which pins down consumption levels and optimal transfers) adjusts towards whichever household has the binding constraints

Solving the Model

- The model can be extended from 2 HHs to N HHs
  - The Pareto frontier traces out the utility to household N, given promised utility to households $1, .., N-1$ (i.e. an aggregated "rest of the village" HH)
  - Simulate N households who each follow the policy rule derived for the two-household case
  - One additional parameter, which scales the marginal utility of unconstrained households in order to satisfy the economy-wide budget constraint
- Computational algorithm is quite involved (check Appendix A.3)
  - Largely follows Morten (JPE, 2019) with migration asset and temporary experiment shock as two additional state variables
  - Solve for the equilibrium policy functions by value function iteration until convergence (i.e. $\mid V_{1j}^i - V_{0j}^i \mid < \epsilon$) for before- and after-migration grid

From the Experiment to the Model

1. The experiment changes the financial cost of migrating, d_t^{fin} (fixed in estimation)
2. It may change the utility cost of migrating, $\Delta d_t^{\text{utility}}$ (e.g. generating a utility benefit of migrating with friends)
3. It changes the value of autarky for HHs (i.e. the threat points derived under the no risk sharing scenario)

Identification (focusing on the experimental variation)

- Parameters of the migration asset ($\pi^{\text{get contact}}, \pi^{\text{lose contact, nomig}}, \pi^{\text{lose contact, mig}}$) are identified off migration transitions conditional on earlier migration
  - The experiment induces a number of people to migrate who did not have any previous migration history and thus helps identify the probability of obtaining a contact by observing the re-migration rate following the first migration episode
- Migration parameters (opportunity cost, utility cost and subsidy)
  - Treatment effect on migration rates, both during and after the experiment
- Other preference parameters (risk aversion and discount factor)
  - Moments related to risk sharing: indirect inference based on consumption regression on simulated control and treatment group

Parameter Estimates

Preferences	
CRRA parameter	1.88
	(0.037)
Opportunity cost of migration	0.15
	(0.088)
Utility cost of migrating	0.075
	(0.0051)
Utility subsidy	0.075
	(0.025)
Decay rate of utility subsidy	0.15
	(0.83)
Income processes	
Mean home income	2.23
	(0.23)
Std. home income	0.58
	(0.0040)
Mean city income with contact	0.38
	(0.17)
Std. city income with contact	0.78
	(0.074)
Dynamics	
Prob. get contact	0.79
	(0.28)
Prob. lose contact if migrate	0.44
	(0.85)
Prob. lose contact if don't migrate	0.66
	(5.62)
Model criterion	1.715

Notes: The table shows parameter estimates and standard errors. The parameter estimates arise from estimating the model by simulated method of moments. The analytical standard errors are computed by numerical differentiation. The mean level of utility in control villages is 3.23.

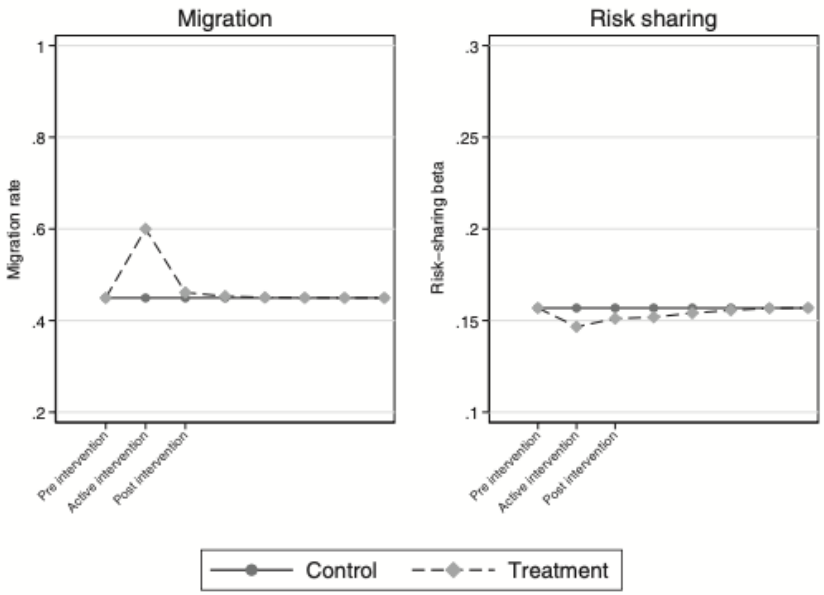
Model Fit

	Data	Model
Risk sharing (control)	0.16	0.16
Risk sharing (treatment effect)	−0.073	−0.055
Mean migration rate	0.39	0.45
Mig. treatment effect (during RCT)	0.22	0.35
Mig. treatment effect (after RCT)	0.094	0.15
Migrate neither during/after RCT (control)	0.49	0.32
Migrate during and after RCT (control)	0.23	0.22
Migrate neither during/after RCT (treatment effect)	−0.17	−0.16
Migrate during and after RCT (treatment effect)	0.15	0.14
Mean log home income (migrant)	1.80	1.57
Std. log home income (migrant)	0.67	0.29
Mean log home income (nonmigrant)	2.13	2.63
Std. log home income (nonmigrant)	0.56	0.33
Log std. mig. income (migrant)	0.27	0.23
Log mean mig. income (nonmigrant prior pd.)	0.60	0.39
Log mean mig. income (migrant prior pd.)	0.73	0.51
Model criterion		1.72

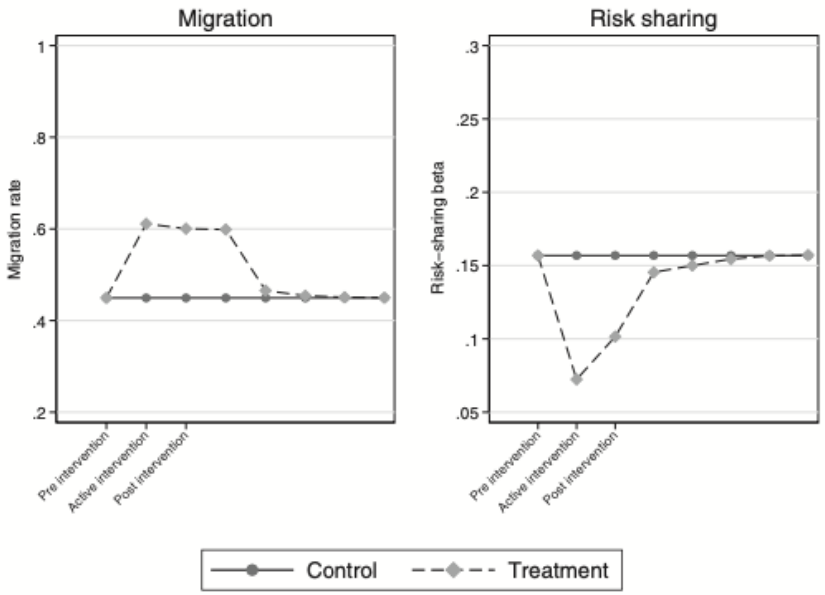
Notes: The table shows the targeted moments in the data (column (1)) and in the estimated model (column (2)).

Simulating the Experiment Inside the Model

(b) Financial component

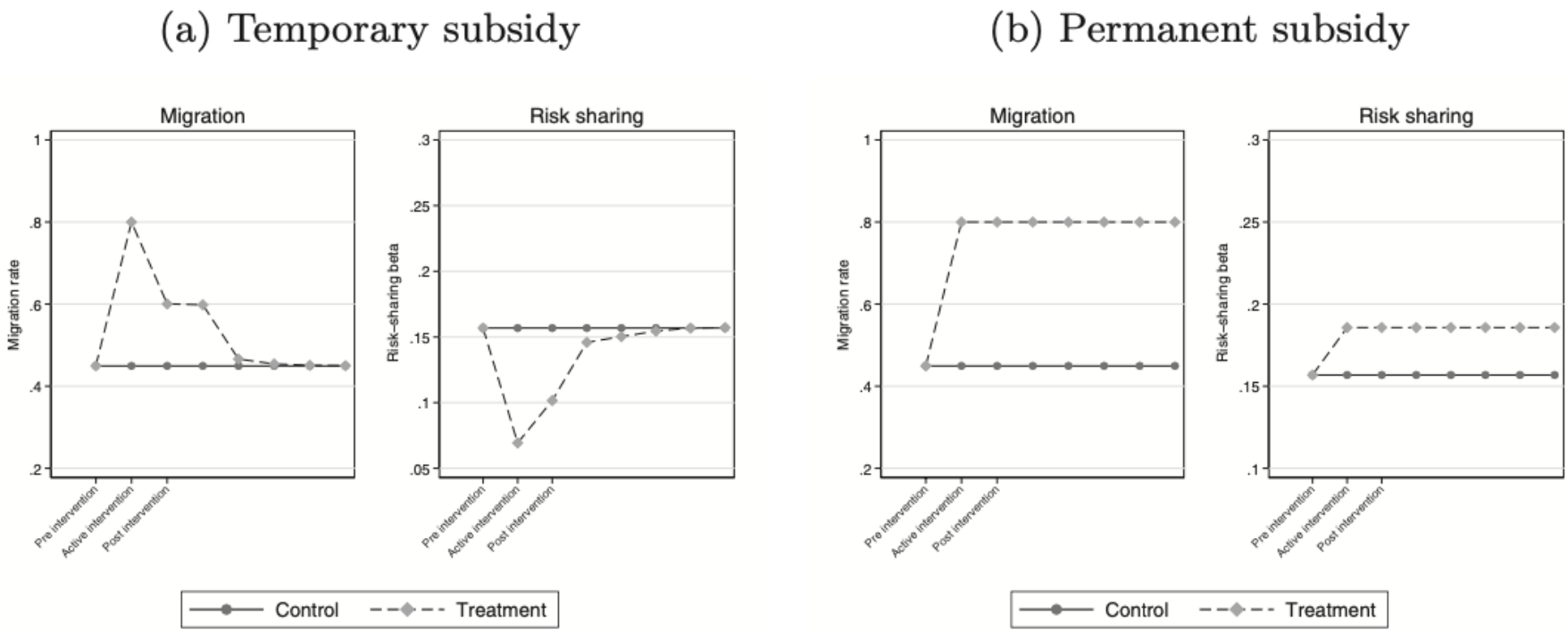


(c) Utility component



- The experiment led to an increase in welfare equivalent to a permanent 12.9% increase in consumption, net of the financial subsidy
- Welfare gains are three times higher when accounting for spillover effects of the experiment through risk sharing

Counterfactuals



- The permanent subsidy leads to a crowding-out of risk sharing through increased outside option + reduced risk of migration
- Difficulty of extrapolating from RCTs to alternative longer-term policies

Main Takeaways of the Paper

- New migration opportunities led to a positive spillover on risk sharing, generating larger welfare gains
  1. Increased transfers between households
  2. Increased the willingness of households to help others
  3. Reduced the exposure of household consumption on household income
- A dynamic migration model with limited commitment reveals the conditions under which risk sharing improves and quantifies the welfare gains
  - Key forces are the riskiness of the migration option and the expected value of city income
- Alternative policy experiments illustrate how temporary interventions may have very different impacts than longer-term policies

The Value of Combining an RCT with an Economic Model

- This paper uses existing experimental data to explore the trade-offs involved with introducing new income opportunities into a village
- The structural model draws on Morten (JPE, 2019) and adds a dynamic component making migration state-dependent
- The experiment both provides a genuine source of exogenous variation and allows the authors to account for the utility aspects of the subsidy that would otherwise not be identifiable