Specifications

Anja Sautmann

Tuesday, June 10

1 Introduction

Introduction

- We want to collect comparable effects from each paper that follow statistical best practice and convention.

- Many papers report multiple treatment effects for the same outcome:
  - Different estimands
  - Different samples
  - Different estimation specifications

- IDEAL uses a “priority” approach:
  - One “IDEAL preferred” treatment effect estimate
  - If different: the authors’ preferred estimate

- This session: how to identify estimands, specifications, and samples used to pick out the right TE estimates.

Some terms

- The **estimand** refers to our quantity of interest.
- An **estimator** is a method to approximate this quantity.
- The result of our estimation is called an **estimate**.

The “ideal” estimand

An **Average Treatment Effect (ATE)** is the expected/average causal effect of a program, intervention, or policy for the entire study population.

The elusive ATE

When *compliance* is perfect (and there is no selective *attrition)*, the ATE can be estimated using the difference in means between treatment and control groups.

What is an empirical specification or model?

An empirical model that specifies the relationships between variables in the data, including any adjustments in estimation to incorporate study design and population characteristics. This may appear as an equation or just a description in text.

Muralidharan et al. 2021

C. *Estimation*

We report intent-to-treat estimates, comparing mean outcomes in treatment and control areas. We discuss MAO beliefs and their implications for interpretation in our cost-benefit analysis below. We thus estimate

$$y_{ivmsd} = \alpha + \beta T_{msd} + \delta_{sd} + \gamma \mathbf{X}_{ivmsd} + \epsilon_{ivmsd}, \quad (1)$$

where y is an outcome, T an indicator for assignment to treatment, and $\mathbf{X}$ a vector of prespecified covariates. In practice, there is only one covariable: the size of landholdings, binned into 40 evenly sized bins (i.e., 0 to 2.5th percentile of landholdings, 2.5th to 5th percentile, etc.).[27] Indices denote individual i in village v in mandal m in stratum s in district d. Treatment is strictly exogenous conditional on the randomization stratum fixed effects δ_{sd}. We cluster standard errors at the level of treatment assignment (the MAO) and conduct randomization inference as a robustness check. When using call center data, we reweight estimates by the inverse probability of being sampled.

Betancourt et al. (2020)

Statistical analysis

We compared trajectories of outcomes over time among families receiving the *Sugira Muryango* intervention with UC using linear mixed models for continuous outcomes and generalised linear mixed models with a logit link for binary outcomes. To account for clustering, we included random effects for randomisation cluster and child for outcomes assessed at the child level. For outcomes assessed at the caregiver level we included random effects for cluster and caregiver, and for outcomes assessed at the household level we included random effects for cluster and household. Following intention-to-treat analysis, we used chained equation imputations in STATA to account for missing data [26]. We report coefficients for the time-by-treatment interaction term and standardised effect sizes estimated based on the modelling results (Cohen's *d* for continuous outcomes, odds ratios (OR) for dichotomous outcomes) with 95% confidence intervals. Analyses were conducted using StataSE version 15 (StataCorp, College Station, TX). Intraclass correlations can be found in Additional File 1, and further analyses examining whether a family's enrolment in either ePW or cPW moderated intervention effects can be found in Additional File 2.

2 Full-sample ITT

What estimates are we collecting?

Noncompliance

	Assigned treatment	Received treatment	Complier
Observation 1	1	1	1
Observation 2	1	0	0
Observation 3	1	1	1
Observation 4	0	0	1
Observation 5	0	1	0
Observation 6	0	0	1

In practice, compliance is seldom perfect (and sometimes unknown).

Intention-to-treat (ITT)

The **intent-to-treat estimand** is the effect of being assigned to treatment vs. control.

- Sometimes actually the object of interest
- Other times, approximates the ATE

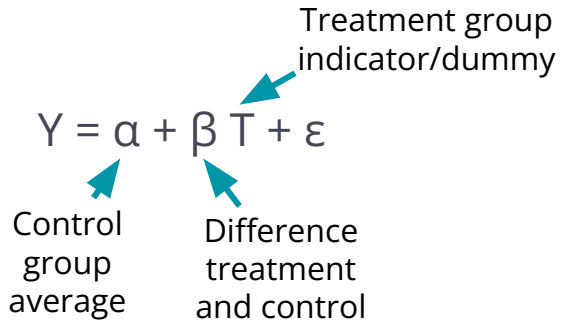
Estimating ITT

The simplest way to estimate the ITT is the difference in means between treatment and control groups.

Estimating the ITT in practice

- Some fields simply report mean (and SE, N) of all study arms, plus statistical test of the difference

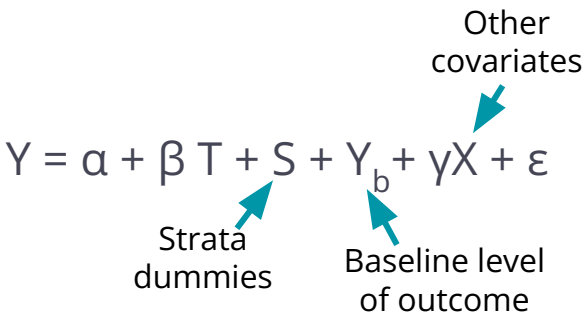
- In economics: typically OLS (linear regression) approach
- In the simplest case, the two are equivalent



Estimating the ITT in practice

- Most commonly: include covariates
  - Strata fixed effects
  - Covariates, including a baseline measure of the outcome if available
- Sometimes called **ANCOVA**: “analysis of covariance” specification

Note: covariates must be measured prior to the intervention (“at baseline”) or invariant over the intervention period (e.g. location of a school) to be exogenous.



What is the IDEAL-preferred specification?

IDEAL prioritizes ITT estimate with

- Strata fixed effects (dummies)
- Baseline outcome measure
- *No other controls.*

Parsimonious specification with good statistical power.

Which specifications do we collect?

Authors often report multiple specifications. While coding a paper:

- First, report all specification(s) you see in the paper
- Then the survey will ensure based on your responses that we have captured the best available specification from the IDEAL ranking of specifications
- **If different, we will also capture the "author preferred" specification.**
  - Reasons: want to capture results highlighted in abstract or intro, communication about the paper
  - Reasonable people can disagree on valid covariates

Empirical Specifications: Ranking

The following **priority rules** determine which empirical specifications are collected as “IDEAL-preferred” specification:

- *Stratification adjustment > No stratification adjustment*
- *Baseline outcome included > Baseline outcome not included*
- *No other controls included > Most parsimonious specification with (pre-treatment or static) controls*

Note: specifications with ***controls that are measured post intervention and not static over the intervention period*** are only collected if they are preferred by the authors, see below.

Empirical Specifications: Ranking

Based on these rules, the IDEAL-preferred specification ranking is as follows, from highest (1) to lowest (8) preferred:

#	Strata correction	Baseline outcome	Other controls	No other controls
1	✓	✓		
2	✓			✓
3	✓	✓	✓	
4	✓		✓	
5		✓		
6				✓
7		✓	✓	
8			✓	

What is “author preferred”?

Author-preferred:

- **Explicit statement** in the text. E.g.
  - “Our preferred estimates show…”
  - “We prefer this specification because…”
  - “The treatment on the treated effect is the most relevant for…”
- **Highlighting one specific treatment effect** in the abstract or introduction over the others

Picking out specifications

IDEAL preferred specification is used (and author-preferred).

Riley (2024)

E. *Empirical Strategy*

Survey Data.—I follow McKenzie (2012) and estimate intent-to-treat (ITT) effects using an ANCOVA specification:

$$Y_{is1} = \alpha_0 + \alpha_1 T_{1i} + \alpha_2 T_{2i} + \boldsymbol{\alpha}_s + Y_{is0} + \epsilon_{is}, \quad (1)$$

where Y_{is1} is the outcome of interest for individual i in stratum s, T_1 is a dummy variable for assignment to the Mobile Account treatment, T_2 is a dummy variable for assignment to the Mobile Disbursement treatment, $\boldsymbol{\alpha}_s$ is a set of randomization strata dummies (Bruhn and McKenzie 2009), Y_{is0} the baseline value of the outcome (if measured at baseline, otherwise excluded), and ϵ_{is} is a random error term. Hypothesis tests will use standard errors that allow for heteroskedasticity. The parameters of interest are α_1 and α_2, which give the effect of assignment to the Mobile Account and Mobile Disbursement groups respectively on the outcome variable. I also test whether the treatments have a significantly different effect from each other ($\alpha_1 = \alpha_2$).

Picking out specifications

Term ITT not explicitly used; random intercept at the school level accounts for cluster randomization: specification (8).

Wolf et al., 2019

Level 1 (child-level) model:

$$Y_{ijk} = B_{0jk} + B_{1jk}'X_{ijk} + e_{ijk}$$

Where X_{ijk} is the vector of child covariates (gender, age, and baseline score).

Level 2 (classroom-level) model:

$$B_{0jk} = \gamma 00k + u0jk$$

Where B_{0jk} is the classroom-level random intercept.

Level 3 (school-level) model:

$$\gamma_{00k} = \pi_{000} + \pi_{001}TT_k + \pi_{002}TTPA_k + \pi_{003}'Z_k + v_{00k}$$

Where γ_{00k} is the school-level random intercept; Z_k is the vector of school-level covariates (district dummies, private or public status, and four dummy variables for different school mobility scenarios [i.e., three treatment schools combined their separate KG classrooms into one KG1 and one KG2 classroom; one school split into two schools between baseline and follow-up; nine schools started with a combined KG1 and KG2 class and split into two separate classrooms midyear; and 12 teachers switched to teach a different KG class within the school midyear]); TT_k is an indicator for schools assigned to the teacher training condition; and $TTPA_k$ an indicator for schools assigned to teacher training plus parental awareness.

Picking out specifications

IDEAL specification (2) is shown, as is an author-preferred estimate that is equivalent to specification (3).

Ozler et al., (2018)

years old, respectively. We analyze these outcomes at the individual level. To estimate intention-to-treat (ITT) effects of each intervention on child outcomes, we employ a regression model of the following form for each round of follow-up data collection:

$$Y_{ij} = \alpha + \gamma^2 T_j^2 + \gamma^3 T_j^3 + \gamma^4 T_j^4 + \beta X_{ij} + \varepsilon_{ij}, \quad (1)$$

For each measure of our primary outcomes in rounds 2 and 3, we estimate two versions of the model in equation (1). In the "unadjusted" regressions, we only include indicators for the strata used to perform block randomization – namely, the "district x bin" fixed effects, where bins refer to the groups of four CBCCs on the envelopes used during the public lotteries that were described in detail above (Bruhn and McKenzie, 2009).[22] In our "adjusted" regressions, we add indicator variables for child age in months and the baseline (lagged) value of the child development measure to the X_{ij} vector. These variables were chosen because they are strongly predictive of performance at follow-up and, as a result, improve the precision of the impact estimates. We prefer this analysis of covariance specification to a difference-in-difference estimation because of the large gains in power (McKenzie, 2012).

Table 3
Impacts on child assessments - 18-month follow-up.

		Dependent Variable:									
		Attending CBCC: 2012-13		Enrolled in Primary: 2012-13		Malawi Developmental Assessment Tool Score					
						Total		Language Skills		Fine Motor/ Perception Skills	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
T2 (teacher training)		0.084**	0.090***	−0.105***	−0.109***	−0.092	−0.063	−0.041	−0.031	−0.134*	−0.107*
		(0.034)	(0.030)	(0.034)	(0.029)	(0.080)	(0.060)	(0.079)	(0.063)	(0.074)	(0.059)
T3 (T2 + teacher incentives)		0.085**	0.097***	−0.086**	−0.097***	0.013	−0.003	0.088	0.085	−0.081	−0.115*
		(0.035)	(0.031)	(0.034)	(0.031)	(0.081)	(0.067)	(0.084)	(0.072)	(0.075)	(0.065)
T4 (T2 + parenting training)		0.065**	0.078***	−0.064**	−0.075***	0.115	0.126*	0.183**	0.185**	0.008	0.012
		(0.031)	(0.028)	(0.031)	(0.027)	(0.086)	(0.067)	(0.087)	(0.071)	(0.075)	(0.061)
Lagged Dependent Variable (Baseline)			−0.092***		0.090***		0.510***		0.426***		0.444***
			(0.015)		(0.015)		(0.034)		(0.031)		(0.034)
Any Treatment (T2, T3, or T4) - **separate regression**		0.077***	0.088***	−0.083***	−0.092***	0.019	0.027	0.084	0.087	−0.064	−0.065
		(0.027)	(0.024)	(0.027)	(0.023)	(0.071)	(0.055)	(0.071)	(0.058)	(0.065)	(0.051)
Mean and Standard Deviation of dependent variable in the control group		0.695		0.297		0.000		0.000		0.000	
		(0.461)		(0.458)		(1.000)		(1.000)		(1.000)	
F-test for Equality of Parameters (p-value)	T2 = T3	0.974	0.843	0.586	0.701	0.191	0.367	0.131	0.106	0.454	0.889
	T2 = T4	0.567	0.687	0.234	0.276	0.007***	0.002***	0.005***	0.001***	0.033**	0.049**
	T3 = T4	0.546	0.543	0.503	0.478	0.195	0.058*	0.266	0.168	0.199	0.045**
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations		1925	1925	1925	1925	1936	1936	1936	1936	1936	1936

Notes: 0.01 - ***; 0.05 - **; 0.1 - *; OLS regressions at the child level using standardized test scores at the 18-month follow-up and baseline covariates with standard errors (SEs) in parentheses. Child assessments at the 18-month follow-up were conducted at the end of the 2012-13 school year. The covariate adjustment referred to as the 'lagged dependent variable' is the baseline value of the dependent variable, except in columns (1)–(4), where it is the baseline value of the Malawi Developmental Assessment Tool: Total Score. SEs are clustered at the CBCC level and observations are weighted using sampling weights and tracking weights (for 42 observations randomly assigned to tracking). 'Any treatment (T2, T3, or T4)' refers to a separate OLS regression, presenting the average impact of being in one of the three treatment arms in comparsion with the control group.

Picking out specifications

Ozler et al., (2018)

years old, respectively. We analyze these outcomes at the individual level. To estimate intention-to-treat (ITT) effects of each intervention on child outcomes, we employ a regression model of the following form for each round of follow-up data collection:

$$Y_{ij} = \alpha + \gamma^2 T_j^2 + \gamma^3 T_j^3 + \gamma^4 T_j^4 + \beta X_{ij} + \varepsilon_{ij}, \quad (1)$$

For each measure of our primary outcomes in rounds 2 and 3, we estimate two versions of the model in equation (1). In the "unadjusted" regressions, we only include indicators for the strata used to perform block randomization – namely, the "district x bin" fixed effects, where bins refer to the groups of four CBCCs on the envelopes used during the public lotteries that were described in detail above (Bruhn and McKenzie, 2009).[22] In our "adjusted" regressions, we add indicator variables for child age in months and the baseline (lagged) value of the child development measure to the X_{ij} vector. These variables were chosen because they are strongly predictive of performance at follow-up and, as a result, improve the precision of the impact estimates. We prefer this analysis of covariance specification to a difference-in-difference estimation because of the large gains in power (McKenzie, 2012).

Picking out specifications

Table 3

Impacts on child assessments - 18-month follow-up.

	Dependent Variable:									
	Attending CBCC: 2012-13		Enrolled in Primary: 2012-13		Malawi Developmental Assessment Tool Score					
					Total		Language Skills		Fine Motor/ Perception Skills	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
T2 (teacher training)	0.084**	0.090***	−0.105***	−0.109***	−0.092	−0.063	−0.041	−0.031	−0.134*	−0.107*
	(0.034)	(0.030)	(0.034)	(0.029)	(0.080)	(0.060)	(0.079)	(0.063)	(0.074)	(0.059)
T3 (T2 + teacher incentives)	0.085**	0.097***	−0.086**	−0.097***	0.013	−0.003	0.088	0.085	−0.081	−0.115*
	(0.035)	(0.031)	(0.034)	(0.031)	(0.081)	(0.067)	(0.084)	(0.072)	(0.075)	(0.065)
T4 (T2 + parenting training)	0.065**	0.078***	−0.064**	−0.075***	0.115	0.126*	0.183**	0.185**	0.008	0.012
	(0.031)	(0.028)	(0.031)	(0.027)	(0.086)	(0.067)	(0.087)	(0.071)	(0.075)	(0.061)
Lagged Dependent Variable (Baseline)		−0.092***		0.090***		0.510***		0.426***		0.444***
		(0.015)		(0.015)		(0.034)		(0.031)		(0.034)
Any Treatment (T2, T3, or T4) - **separate regression**	0.077***	0.088***	−0.083***	−0.092***	0.019	0.027	0.084	0.087	−0.064	−0.065
	(0.027)	(0.024)	(0.027)	(0.023)	(0.071)	(0.055)	(0.071)	(0.058)	(0.065)	(0.051)
Mean and Standard Deviation of dependent variable in the control group	0.695		0.297		0.000		0.000		0.000	
	(0.461)		(0.458)		(1.000)		(1.000)		(1.000)	
F-test for Equality of Parameters (p-value) T2 = T3	0.974	0.843	0.586	0.701	0.191	0.367	0.131	0.106	0.454	0.889
T2 = T4	0.567	0.687	0.234	0.276	0.007***	0.002***	0.005***	0.001***	0.033**	0.049**
T3 = T4	0.546	0.543	0.503	0.478	0.195	0.058*	0.266	0.168	0.199	0.045**
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	1925	1925	1925	1925	1936	1936	1936	1936	1936	1936

Notes: 0.01 - ***; 0.05 - **; 0.1 - *; OLS regressions at the child level using standardized test scores at the 18-month follow-up and baseline covariates with standard errors (SEs) in parentheses. Child assessments at the 18-month follow-up were conducted at the end of the 2012-13 school year. The covariate adjustment referred to as the 'lagged dependent variable' is the baseline value of the dependent variable, except in columns (1)–(4), where it is the baseline value of the Malawi Developmental Assessment Tool: Total Score. SEs are clustered at the CBCC level and observations are weighted using sampling weights and tracking weights (for 42 observations randomly assigned to tracking). 'Any treatment (T2, T3, or T4)' refers to a separate OLS regression, presenting the average impact of being in one of the three treatment arms in comparsion with the control group.

2 Heterogeneous Treatment Effects

Treatment effects for population subgroups

Heterogeneous Treatment Effects

- Treatment effects vary across the population
- Heterogeneity has implications for ability to estimate the ATE

- **Conditional Average Treatment Effects/CATE**
- Researchers may be explicitly *interested* in the effect on specific groups *or* the effect difference

Estimating CATEs

- **Split the sample into sub-samples (recorded through UoA)**
- **Use interaction terms that allow the treatment effect to differ by group**

What CATEs do we collect?

Authors may report treatment effects for multiple subsamples.

- **IDEAL prioritizes the full-sample effect**
- **Heterogeneous effects** collected if
  - These are the **only effects reported**
  - Effects in a sub-sample are part of the research question the paper answers → **highlighted** in abstract or introduction

- Example: some papers are explicitly interested in differential treatment effects by gender.

2 LATE and TOT

Alternative estimands under non-compliance

Noncompliance

	Assigned treatment	Received treatment	Complier
Observation 1	1	1	1
Observation 2	1	0	0
Observation 3	1	1	1
Observation 4	0	0	1
Observation 5	0	1	0
Observation 6	0	0	1

- **May want to explicitly estimate treatment effects among *compliers*:**
  - Local Average Treatment Effect (LATE)
  - Also sometimes called CACE (Conditional average causal effect)
  - Effect of treatment on the treated (TOT) (with one-sided non-compliance)

- **Problem:** non-compliers are selected – e.g. it's plausible that…
  - those with largest treatment effects over-comply
  - those with lowest treatment effects under-comply

Local Average Treatment Effect (LATE)

- LATE: the treatment effect among compliers
- Idea: use random assignment as an *instrument* for actual treatment receipt
- LATE vs. ITT:
  - Sometimes the object of interest
  - Sometimes an approximation of ATE

Estimating LATE

- **Instrumental-variable (IV) estimate**
- **Two-stage least square (2SLS)**
- **Wald estimator in TOT case (Scaling ITT by proportion of compliers)**

What is the IDEAL preferred estimand?

Authors may report multiple estimands.
LATE especially common in so-called “encouragement designs”.

- **IDEAL prioritizes the ITT**
- **LATE** collected if
  - It is the **only estimand reported**
  - LATE is the **author-preferred estimand**

Picking out specifications

Linhares et al. (2022)

Strengthening Bonds Program's Efficacy Findings

First, considering the intention-to-treat GEE analysis, which included the 27 dropout participants in the entire sample analysis (n = 119), there were no statistical differences between groups in parenting behavior outcomes (parental practices, maternal interactive behavior, maternal sense of competence) and child behaviors outcomes.

Second, Table 3 presents the treatment-on-the-treated GEE analysis, including the 92 participants who completed the study with the intervention and assessments concluded. In the treatment-on-the-treated, there were statistically significant effects of the intervention for decreasing mothers' coercive parenting practices (d = -0.54, medium effect size) and child behavior problems (d = -0.43, small effect size), when adjusting for socioeconomic status and mother's years of schooling. The treatment-on-the-treated without the adjustment showed a decrease only for mothers' coercive parenting practices. Additionally, there were no statistically significant effects of the intervention on the other mothers' parenting practices (parent-child relationship, positive encouragement, parental inconsistency, parental sense of competence, and maternal interactive behavior).

ITT results are discussed briefly in the paper but not reported in any exhibit.

Rank 8 in the IDEAL preference list for the TOT shown.

Table 3

Strengthening Bonds Program's Effects on Parenting and Child Behavior Outcomes (GEE Analysis)

Outcomes	Coefficient (*SE*)	*p*-value	[95% CI] Min	Max
Parenting components				
PAFAS[1] - Parent-child relationship	0.10 (0.24)	.68	-0.36	0.56
PAFAS - Positive encouragement	-0.06 (0.13)	.69	-0.39	0.26
PAFAS - Parental inconsistency	-0.26 (0.30)	.38	-0.85	0.32
PAFAS - Coercive parenting	-0.91 (0.34)	.01	-1.56	-0.25
PSOC[2] - Parental sense of competence	1.52 (0.96)	.11	-0.36	3.41
PICCOLO[3] - Maternal interactive behavior (structured-play situation)	-1.32 (1.79)	.46	-4.83	2.19
PICCOLO - Maternal interactive behavior (free-play situation)	-2.09 (2.03)	.30	-6.06	1.88
Child behavior				
Child behavior problems	-2.31 (1.12)	.04	-4.50	-1.11

Note. min = minimum; max = maximum; *SE* = standard error. [1]PAFAS = Parenting and Family Adjustment Scales; [2]PSOC = Parenting Sense of Competence Scale; [3]PICCOLO = Parenting Interactions with Children: Checklist of Observations Linked to Outcomes. Results adjusted for socioeconomic status score and mothers' years of schooling.

4 Putting it all together

Counting treatment effects

- Our approach ensures **at least one** treatment effect is collected per outcome.

- We may also collect **up to two** specifications per outcome:

  - IDEAL-preferred: e.g., ITT with strata fixed effects and no controls.
  - Author-preferred: e.g., TOT with strata fixed effects and additional controls.

- Sometimes there may be multiple rounds of data collection post intervention (year 1 follow-up, year 2 follow-up, etc.) used to measure outcomes:

  - If each round is associated with a distinct estimate, we count each as a separate treatment effect (called “periods” in the context of estimates).
  - This means there may be **more than one** treatment effect per outcome and specification.

Caveat

What we aren't collecting in the current survey version

*Some specification details are **not** currently implemented in the surveyCTO form:*

- **Stage 1**: Factorial designs with treatment interaction terms.
- **Stage 2:** LATE is author-preferred but LATE and ITT are reported in the same exhibit
- **Stages 1, 2, and 3**:
  - Treatment effects estimated using linear combinations of coefficients
  - Heterogeneous treatment effects that are estimated in the same regression specification (with interaction terms)

Thank you for listening

Anja Sautmann
asautmann@worldbank.org

IDEAL
Impact Data and Evidence Aggregation Library