IDEAL

Impact Data and Evidence Aggregation Library

Concept review: Estimates

Alaka Holla

June 11, 2025

Agenda

1. Definition of an estimate
2. Null hypothesis
3. Point estimate and effect size
4. Precision
5. Sample size

1 Estimate

How to define an estimate in IDEAL

Glossary

- The ***estimand*** refers to our quantity of interest.
  Examples: ITT, TOT, LATE

- An ***estimator*** is a method to approximate this quantity
  Examples: Mean difference, odds ratio

Glossary

- An ***estimation model*** is a statistical technique to predict this quantity.

  Examples: OLS regression, hierarchical linear modelling, t-test (mean comparison)

- The result of our estimation is called an ***estimate***.

  This is the number!

Estimate

In IDEAL, an estimate refers to the result of the estimation of a treatment effect.

An estimate is defined by a number of components.

Estimate

Stages 1 & 2

- Estimand
- Contrast
- Outcome
- Unit of analysis
- Period
- Specifications

Stage 3

- Estimator
- Estimation model
- Null hypothesis
- Point estimate
- Precision

In SurveyCTO

Information you have entered in Stage 1 & Stage 2 will appear in Stage 3

Exhibit Label: Table 6

Outcome: Any prenatal care

Treatment: Performance based payment of health care providers

Control: Payments equivalent to the average amount of P4P payments

Unit of analysis: Parent

Type: IDEAL-preferred

Estimand: TOT / LATE

Empirical Specification: Strata Fixed Effects + Static Controls

Round of data collection: 25 months after study baseline

Estimator

Estimation model

Null hypothesis

Point estimate

Precision

Stage 3

2

Null hypothesis

Different types of null hypotheses

Null hypothesis

A null hypothesis (often denoted as H0) is a statement about a population parameter that we assume to be true until we have enough evidence to reject it.

- H0 = 0, for most cases
- H0 = 1, for odds ratio, risk ratio, etc.

Why do we want to know this?
Interpret 0.03 and 1.13

Sharp null

- A strong and specific statement:

The treatment has **no effect on any unit or individual** in the study.

Null: $\beta_i = 0, \forall i$ (sharp null hypothesis)

Null: $\beta_i = 1, \forall i$ (sharp null hypothesis)

Null: $\beta_i = constant, \forall i$ (sharp null hypothesis), specify

Example - where to find

Effect of a home-visiting parenting program to promote early childhood development and prevent violence: a cluster-randomized trial in Rwanda

Author affiliations • Sarah KG Jensen [1], Matias Placencio-Castro [2], Shauna M Murray [1], Robert T Brennan [1 3], Simo Goshev [4], Jordan Farrar [1], Aisha Yousafzai [5], Laura B Rawlings [6], Briana Wilson [6], ... 3 Show all authors

Abstract

Introduction Families living in extreme poverty require interventions to support early-childhood development (ECD) due to broad risks. This longitudinal cluster randomised trial examines the effectiveness of Sugira Muryango (SM), a home-visiting intervention linked to Rwanda's social protection system to promote ECD and reduce violence compared with usual care (UC).

Methods Families with children aged 6–36 months were recruited in 284 geographical clusters across three districts. Cluster-level randomisation (allocated 1:1 SM:UC) was used to prevent diffusion. SM was hypothesised to improve child development, reduce violence and increase father engagement. Developmental outcomes were assessed using the Ages and Stages Questionnaire (ASQ-3) and the Malawi Development Assessment Tool (MDAT) and anthropometric assessments of growth. Violence was assessed using questions from UNICEF Multiple Indicators Cluster Survey (MICS) and Rwanda Demographic and Health Surveys (DHS). Father engagement was assessed using the Home Observation for Measurement of the Environment. Blinded enumerators conducted interviews and developmental assessments.

Results A total of 541 SM families and 508 UC families were enrolled and included in the analyses. Study attrition (2.0% children; 9.6% caregivers) was addressed by hot deck imputation. Children in SM families improved more on gross motor (d=0.162, 95% CI 0.065 to 0.260), communication (d=0.081, 95% CI 0.005 to 0.156), problem solving (d=0.101, 95% CI 0.002 to 0.179) and personal-social development (d=0.096, 95% CI −0.015 to 0.177) on the ASQ-3. SM families showed increased father engagement (OR=1.592, 95% CI 1.069 to 2.368), decreased harsh discipline (incidence rate ratio IRR=0.741, 95% CI 0.657 to 0.835) and intimate partner violence (IRR=0.616, 95% CI:0.458 to 0.828). There were no intervention-related improvements on MDAT or child growth.

Conclusion Social protection programmes provide a means to deliver ECD intervention.

Trial registration number NCT02510313.

Null hypothesis

Problem!

Null hypothesis is rarely explicitly stated.

It is implied:

- Authors interpretation
- Estimator, estimation model

...is significantly different from zero...

...violent punishment goes down by 24 percent...

3 Point estimate

Point estimate versus effect size

	Point estimate	Effect size
Definition	A single value of the estimate.	The magnitude of the treatment effect.
Unit	Same as the unit of outcome measure.	A standard measure, e.g. cohen's *d*, hedge's *g*.
Objective		Create a common metric to include different outcome measures in the same synthesis.

Papers may report point estimates and/or effect sizes. The outcome variable would be a *standardized* outcome, if an effect size was reported.

Standardization of effect size

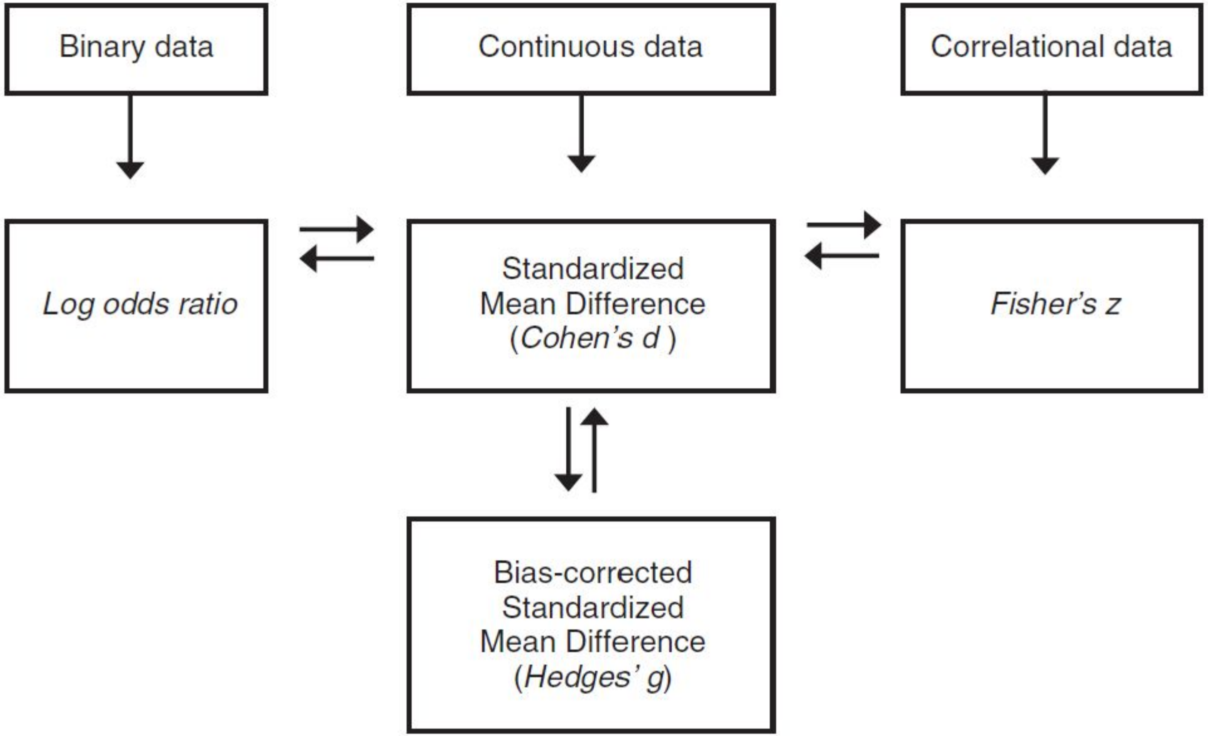
Effect size
based on the estimator

- **Mean difference**
- **Odds ratio**
- Correlations (less common in RCT)

Estimator of the treatment effect indicates which effect size to calculate.

- Mean Difference (Final Values)
- Mean Difference (Net)
- Median Difference (Final Values)
- Median Difference (Net)
- Hazard Ratio (HR)
- Hazard Ratio, Log
- Odds Ratio (OR)
- Odds Ratio, Log
- Risk Difference (RD)
- Risk Ratio (RR)
- Risk Ratio, Log
- Slope
- Other, specify

Converting among effect sizes



4 Precision

Different measures of precision

Measures of precision for standardization

Standard error

T-statistics

Z-statistics
(binary outcome)

Always collect

P-value

Confidence interval

Standard deviation

F-ratio

Collect these alternative if any of the “always” is missing.

The precision statistics can help calculate the effect size.

Calculating ES using different statistics

Useful formulas for calculating *ESsm* from a range of statistical data

Formula	*Data needed and definition of terms*
Direct calculation formula for ES_{sm}	
(1) $ES_{sm} = \dfrac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}$ $s_{pooled} = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	Means ($\overline{X}$), standard deviations (s), and sample sizes (n) for each group.
Algebraically equivalent formulas for ES_{sm}	
(2) $ES_{sm} = t\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}$	Independent t-test (t) and sample sizes (n) for each group.
(3) $ES_{sm} = \dfrac{2t}{\sqrt{N}}$	Independent t-test (t) and total sample size (N). Assumes $n_1 = n_2$.

A range of formulas can be used depending on the available statistics.

Formula	*Data needed and definition of terms*
(15) $s_{pooled} = \dfrac{\overline{X}_1 - \overline{X}_2}{t\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}}$	Means ($\overline{X}$) and sample sizes (n) for each group, and associated t-value (t).
(16) $s = se\sqrt{n-1}$	Standard error of the mean (se) and sample size (n) for any group.
(17) $s_{pooled} = \sqrt{\dfrac{MS_b}{F_{oneway}}}$ $MS_b = \dfrac{\Sigma n_j \overline{X}_j^2 - \frac{(\Sigma n_j \overline{X}_j)^2}{\Sigma n_j}}{k-1}$	F-ratio (F) from a one-way ANOVA with k groups and the mean ($\overline{X}$) and sample size (n) for each group (j).

Precision adjustments

This information helps assess potential bias in estimate of precision.

- Conventional (no adjustment)
- Robust
- Clustered robust

Standard error, confidence interval or p-value

Example - where to find

Table 3—Results of Different Targeting Methods on Error Rate Based on Consumption

		By income status		By detailed income status				
Sample:	Full population (1)	Inclusion error (2)	Exclusion error (3)	Rich (4)	Middle income (5)	Near poor (6)	Very poor (7)	Per capita consumption of beneficiaries (8)
Community treatment	0.031*	0.046**	0.022	0.028	0.067**	0.49	−0.013	9.933
	(0.017)	(0.018)	(0.028)	(0.021)	(0.027)	(0.038)	(0.039)	(18.742)
Hybrid treatment	0.029*	0.037**	0.009	0.020	0.052**	0.031	−0.008	−1.155
	(0.016)	(0.017)	(0.027)	(0.020)	(0.025)	(0.037)	(0.037)	(19.302)
Observations	5,753	3,725	2,028	1,843	1,882	1,074	954	1,719
Mean in PMT treatment	0.30	0.18	0.52	0.13	0.23	0.55	0.48	366

Notes: All regressions include stratum fixed effects. Robust standard errors in parentheses, clustered at the village level. All coefficients are interpretable relative to the PMT treatment, which is the omitted category. The mean of the dependent variable in the PMT treatment is shown in the bottom row. All specifications include stratum fixed effects.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Other types of precision measures

This is a rapidly evolving field, so we would like to record all the types of p-values reported in the paper.

- Precision measures adjusted for multiple hypotheses
- Small-sample correction p-value
- Random inference-based p-value
- Bootstrapped
- Permutation tests p-value
- Other (including if only significance sign is reported, e.g. *, **, ***)

Example - Random inference

TABLE 3—IMPACTS ON STUDENT LEARNING, LINEAR MIXED EFFECTS MODEL

	Pooled	Year 1	Year 2
Model A. Direct effects only			
Advertised P4P (τ_A)	0.01	−0.03	0.04
	[−0.04, 0.08]	[−0.06, 0.03]	[−0.05, 0.16]
	(0.75)	(0.20)	(0.31)
Experienced P4P (τ_E)	0.11	0.06	0.16
	[0.02, 0.21]	[−0.03, 0.15]	[0.04, 0.28]
	(0.02)	(0.17)	(0.00)
Experienced P4P × incumbent (λ_E)	−0.06	−0.05	−0.09
	[−0.20, 0.07]	[−0.19, 0.11]	[−0.24, 0.06]
	(0.36)	(0.54)	(0.27)
Model B. Interactions between advertised and experienced contracts			
Advertised P4P (τ_A)	0.01	−0.02	0.03
	[−0.05, 0.14]	[−0.06, 0.07]	[−0.05, 0.21]
	(0.46)	(0.62)	(0.22)
Experienced P4P (τ_E)	0.12	0.06	0.18
	[0.05, 0.25]	[−0.01, 0.19]	[0.08, 0.33]
	(0.01)	(0.10)	(0.00)
Advertised P4P × experienced P4P (τ_{AE})	−0.03	−0.01	−0.04
	[−0.17, 0.09]	[−0.15, 0.10]	[−0.22, 0.13]
	(0.51)	(0.65)	(0.58)
Experienced P4P × incumbent (λ_E)	−0.08	−0.05	−0.11
	[−0.31, 0.15]	[−0.30, 0.18]	[−0.36, 0.14]
	(0.43)	(0.56)	(0.38)
Observations	154,594	70,821	83,773

Notes: For each estimated parameter, or combination of parameters, the table reports the point estimate (stated in standard deviations of student learning), 95 percent confidence interval in brackets, and p-value in parentheses. Randomization inference is conducted on the associated z-statistic. The measure of student learning is based on the empirical Bayes estimate of student ability from a two-parameter IRT model, as described in Section IIC.

[Leaver et al. 2021]

Example - multiple SE & p-values

Table 1
Program Impacts on Leblango Early Grade Reading Assessment Scores
(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Leblango EGRA Score Index†	(2) Letter Name Knowledge	(3) Initial Sound Recognition	(4) Familiar Word Recognition	(5) Invented Word Recognition	(6) Oral Reading Fluency	(7) Reading Comprehension
Full-cost program	0.638***	1.014***	0.647***	0.374**	0.215	0.476**	0.445**
S.E.	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.128)	(0.113)
R.I. p-value	[0.005]	[0.006]	[0.007]	[0.010]	[0.161]	[0.025]	[0.030]
q-value	--	{0.040}	{0.040}	{0.040}	{0.276}	{0.072}	{0.072}
Reduced-cost program	0.129	0.407	0.076	-0.002	0.031	0.071	0.045
S.E.	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
R.I. p-value	[0.327]	[0.106]	[0.415]	[0.994]	[0.675]	[0.444]	[0.668]
q-value	--	{0.212}	{0.592}	{0.994}	{0.736}	{0.592}	{0.736}
Number of students	1460	1476	1481	1474	1471	1467	1481
Number of schools	38	38	38	38	38	38	38
Adjusted R-squared	0.149	0.219	0.103	0.066	0.075	0.074	0.058
Difference between treatment effects	0.509**	0.607**	0.570***	0.376***	0.184	0.405**	0.400**
S.E.	(0.127)	(0.159)	(0.128)	(0.092)	(0.093)	(0.117)	(0.120)
R.I. p-value	[0.010]	[0.020]	[0.006]	[0.007]	[0.212]	[0.021]	[0.038]
q-value	--	{0.032}	{0.021}	{0.021}	{0.212}	{0.032}	{0.046}
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces. † PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Kerwin and Thornton, 2021

Example - More p-values!

Table 4

Inference Results: Perry Preschool Intervention

Variable (1)	No C (2)	No T (3)	Ctr. M. (4)	Treat. M. (5)	Diff. Ms. (6)	Asy. p-val. (7)	Naive p-val. (8)	Blk. p-val. (9)	Per. S.D. (10)	Blk. p-val. (11)	IPW p. S.D. (12)	Bonf. p-val. (13)
Lifestyles: diet and physical activity at 40 y.o. – males												
Physical activity	35	30	0.457	0.367	0.090	0.766	0.779	0.584	0.584	0.545	0.545	1.000
Healthy diet	35	29	0.229	0.379	0.151	*0.097*	0.113	*0.015*	*0.033*	*0.020*	*0.072*	*0.040*
Lifestyles: smoking at 27 y.o. – males												
Not a daily smoker	39	31	0.462	0.581	0.119	0.164	0.160	*0.092*	0.092	*0.089*	*0.089*	0.267
Not a heavy smoker	39	31	0.615	0.903	0.288	*0.003*	*0.002*	*0.004*	*0.005*	*0.004*	*0.005*	*0.012*
No. of cigarettes	39	31	8.744	4.291	4.453	*0.011*	*0.010*	*0.008*	*0.009*	*0.006*	*0.011*	*0.018*

[Conti, Heckman, and Pinto, 2016]

5 Sample size

Sample size

- Needed for standardization of effect size (in most cases) and meta-analysis (e.g. weighting).
- Provides information on study scale, retention, and attrition.

IDEAL only collects analytical sample size, i.e. the N of observations entered estimation. This could be different from target sample size.

Sample size fields

- At baseline by study arm or combined

[baseline_values]: Please report the following information for the **Baseline Period** or the reference round of data collection associate with this treatment effect:

Important:

- Sometimes, authors choose to report less commonly used precision statistics for the treatment effect. Please enter the information on those statistics in this field, and if available include the reason for which the authors chose to report them.

- Please provide the details of precision statistics reported for the treatment effect, including the type and values.

	Mean	Stand. Deviation	Stand. Error	Sample Size
Evaluation Arm				
Reference Arm				
Both combined				

Sample size fields

- For the data rounds in the period for the treatment effect by study arm or combined

[period_values]: Please report the following information for round "**Follow up**" associated with this treatment effect:

Important:
- Sometimes, authors choose to report less commonly used precision statistics for the treatment effect. Please enter the information on those statistics in this field, and if available include the reason for which the authors chose to report them.
- Please provide the details of precision statistics reported for the treatment effect, including the type and values.

	Mean	Stand. Deviation	Stand. Error	Sample Size
Evaluation Arm				
Reference Arm				
Both combined				

Thank you for listening

Alaka Holla
aholla@worldbank.org

IDEAL

Impact Data and Evidence Aggregation Library