Meta-Analysis Overview

Steve Glazerman, IPA

June 11, 2025

GOAL OF THIS LECTURE

Familiarity – not mastery

Motivate your extraction/coding

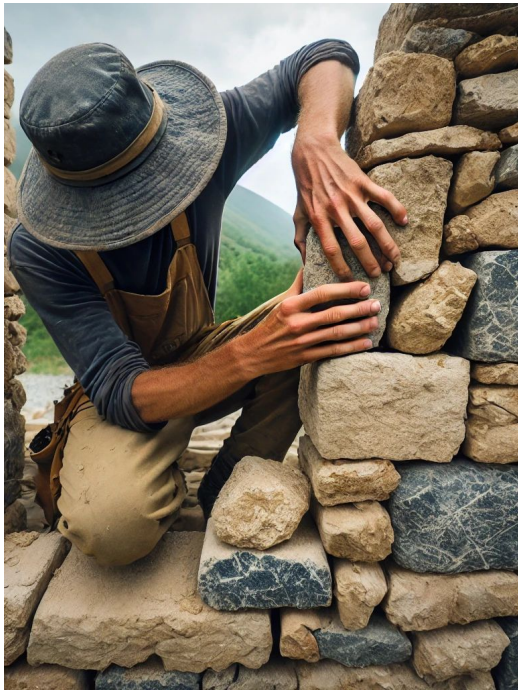
Hone your judgement

Outline

1. Background and example
2. Design of the meta-analysis
3. Implementation
4. Analysis and interpretation

1 Background and Example

What is meta-analysis and why is it useful?



WHAT

- Define a research question and (statistically) summarize all the available evidence on the topic
- Requires identifying studies that meet criteria and harmonizing them and putting them together to analyze as if it were a “study of studies”,
    - “Evidence aggregation”, “systematic review”, “narrative review”

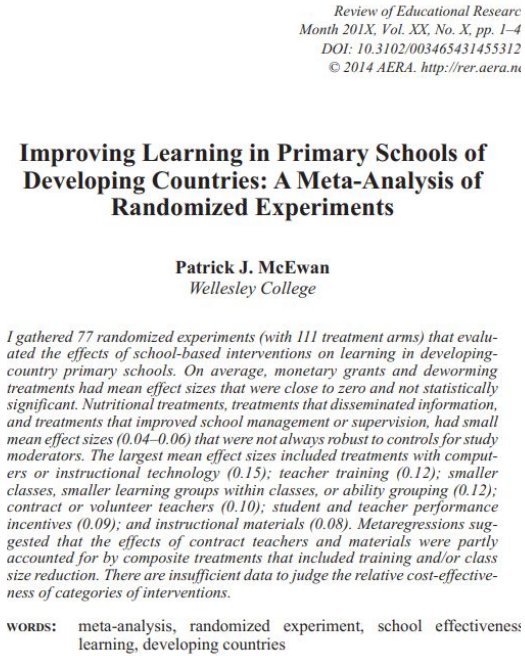
WHY

- Science is cumulative, we strive for consensus on the current state of knowledge
- Help policymakers benefit from what’s already known
- Identify gaps in knowledge

HOW…

Example, McEwan (RER 2014)

- Question: What does the evidence say about the whole range of interventions for improving primary school student achievement?
- Sample: 77 RCTs with 111 treatment arms in LMIC settings
- Results: Reports the average effect size for each of several categories of interventions, like monetary grants, deworming, school management, ed tech, teacher training, class size/ability grouping, performance incentives, instructional materials



Review of Educational Researc
Month 201X, Vol. XX, No. X, pp. 1–4
DOI: 10.3102/0034654314553122
© 2014 AERA. http://rer.aera.n

Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments

Patrick J. McEwan
Wellesley College

I gathered 77 randomized experiments (with 111 treatment arms) that evaluated the effects of school-based interventions on learning in developing-country primary schools. On average, monetary grants and deworming treatments had mean effect sizes that were close to zero and not statistically significant. Nutritional treatments, treatments that disseminated information, and treatments that improved school management or supervision, had small mean effect sizes (0.04–0.06) that were not always robust to controls for study moderators. The largest mean effect sizes included treatments with computers or instructional technology (0.15); teacher training (0.12); smaller classes, smaller learning groups within classes, or ability grouping (0.12); contract or volunteer teachers (0.10); student and teacher performance incentives (0.09); and instructional materials (0.08). Metaregressions suggested that the effects of contract teachers and materials were partly accounted for by composite treatments that included training and/or class size reduction. There are insufficient data to judge the relative cost-effectiveness of categories of interventions.

WORDS: meta-analysis, randomized experiment, school effectiveness learning, developing countries

2 Design

Meta-analysis starts like any study

- Define the research question
- Map out interventions, treatment contrast, outcomes, populations of interest
- Theory of change, logic model highly desirable

Theory of change

Example from McEwan

Provides a mapping o interventions and other explanatory variables on outcomes

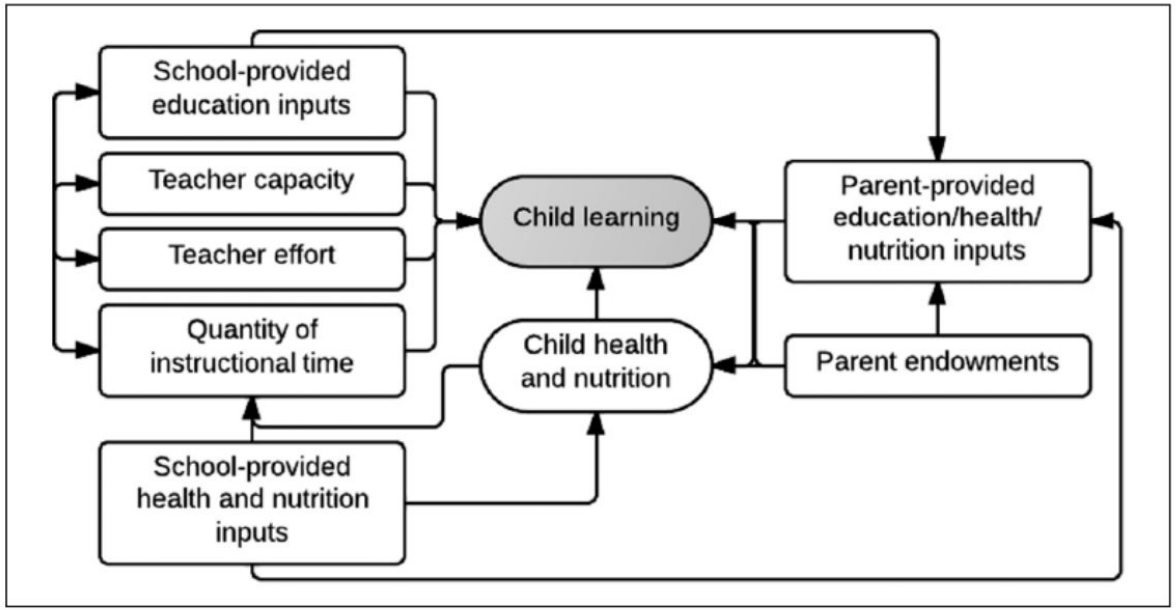


FIGURE 1. *The direct and indirect determinants of child learning.*

Study parameters unique to meta-analysis

- Study selection (inclusion/exclusion) criteria
  - Relevance
  - Methods, e.g. RCTs only or RCTs + certain admissible quasi-experiments
  - Minimum quality and reporting completeness
- Search criteria and scope
  - Databases to search
  - Keywords and Boolean logic
- Protocol
  - Data to extract, controlled vocabularies
  - Coding guidelines and training/supervision
- Analysis plan
  - Aggregation method, e.g. regression
  - Lumping & splitting (hard to pre-specify, often have to refine after seeing the universe of studies)
  - Heterogeneity

Inclusion/Exclusion example

Criteria for Study Inclusion and Exclusion

I included studies if they (a) were conducted in a low- to upper middle-income country, as defined by the World Bank in 2012[1]; (b) were conducted in primary schools, broadly defined to include Grades 1 to 8 (or ages 6 to 14, if the grades were not reported); (c) randomly assigned children (or clusters of children) to an education or health intervention in a school setting, or "business-as-usual" in the same setting; (d) reported results for at least one continuously measured learning outcome in language or reading, mathematics, or a composite assessment including either outcome; and (e) reported sufficient data to calculate the treatment's effect size and standard error, in the full experimental sample.

Exclusion reasons related to reporting, design deficiencies

A remaining set of studies met the previous criteria but did not report sufficient data to estimate effect sizes and/or standard errors. Several studies did not report sufficient data to estimate the mean difference between treatment and control groups at each follow-up (Pollitt, Hathirat, Kotchabhakdi, Missell, & Valyasevi, 1989; Sungthong, Mo-suwan, Chongsuvivtwong, & Geater, 2004; Whaley et al., 2003), in one case because statistically nonsignificant results were not reported in the paper (Nga et al., 2011). In other cases, I could not convert mean differences (or a regression coefficient estimating a similar parameter) to effect sizes, given the lack of data on the standard deviation of the outcome variable (Adelman, Alderman, Gilligan, & Lehrer, 2008; Kazianga, de Walque, & Alderman, 2012; Pandey, Goyal, & Sundararaman, 2009; Vazir, Nagalla, Thangiah, Kamasamudram, & Bhattiprolu, 2006). Finally, I excluded cluster-randomized experiments in which standard errors did not correctly account for the unit of assignment (Chandler, Walker, Connolly, & Grantham-McGregor, 1995; Lai, Zhang, Hu, et al., 2012; Piper & Korda, 2011).

3 Implementation

“How-to” meta-analysis

Steps involved

- Conduct search and build the universe
- Winnow down: Title and abstract review
- Further selection: Full text review
- Coding
- Analysis

Key conceptual issues to understand

- Publication bias, grey literature, p-hacking
- Study design features: Randomization issues, differential attrition, failed balance checks, issues with clustering
- Multiple comparison problems

Harmonization: outcomes

- Outcome types: continuous, dichotomous
- Combining Natural units
  - E.g. monthly wage, business profits, annual household income
- “Unit-free” metrics: Cohen’s *d*, Hedges’ *g* – standard deviation units

$$d = \frac{\bar{Y}_T - \bar{Y}_C}{s_{\text{pooled}}}$$

$$SE_d = \sqrt{\frac{n_T + n_C}{n_T n_C} + \frac{d^2}{2(n_T + n_C)}}$$

Harmonization: intervention, outcome, population

Study A	Study B	Study C	Study D
Training vs. nothing	Soft vs. hard skills training vs. nothing	Low vs. high intensity training	3x2 factorial: high/low/no intensity; soft/hard skills focus
Y = Earnings	Y = Employment, Earnings, Productivity	Y =Earnings at 3, 24 and 60 months	Y = Job satisfaction
Full sample only	Results by gender	Results by gender, baseline education, and county	Full sample only

4 Analysis and interpretation

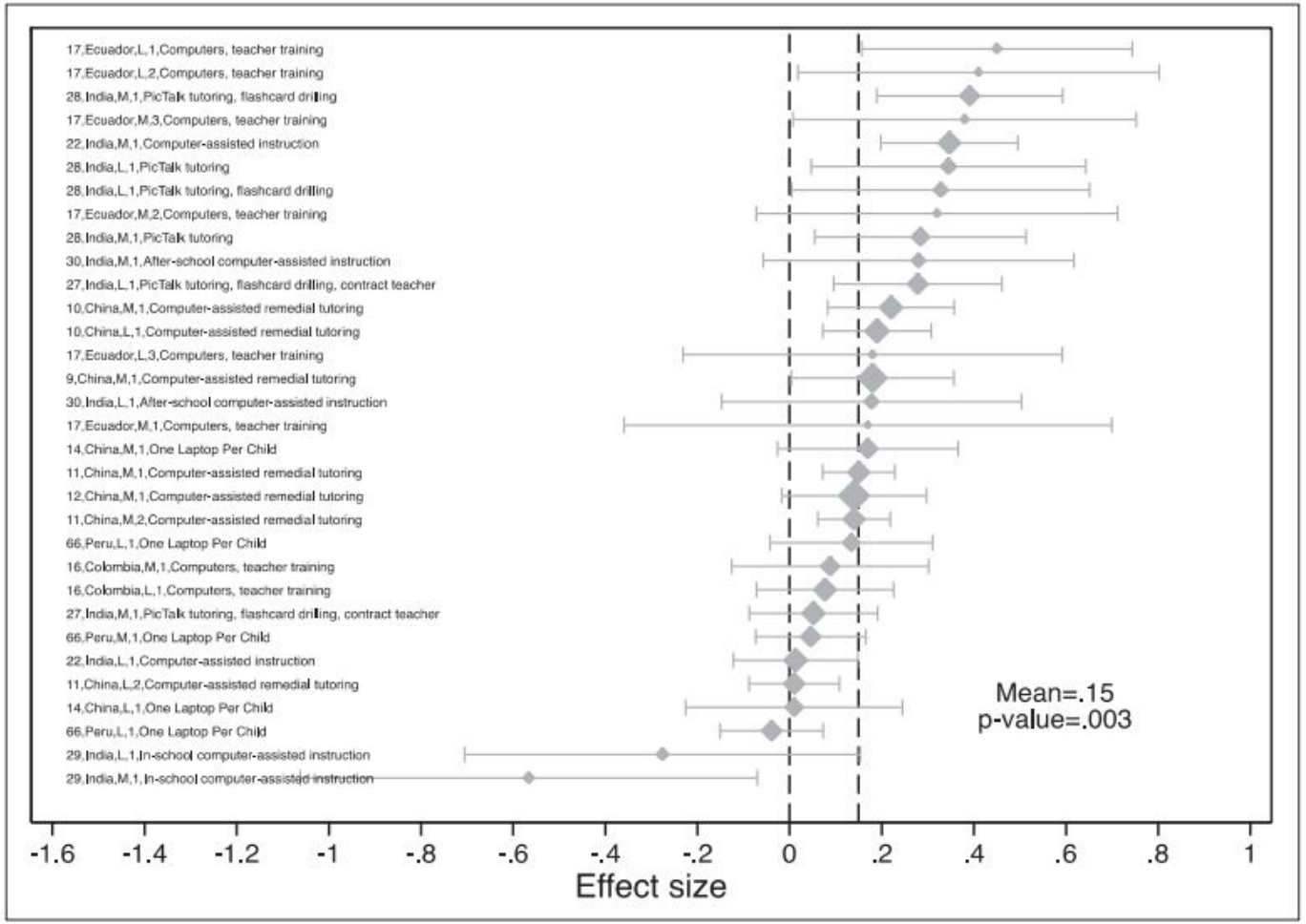
...including summary statistics and visualizations

Quantitatively combining results across studies

- Re-analyze the micro data if possible, gives most flexibility
- Work with impact estimates and standard errors or p-values, most practical
- If all else fails, can count estimates by sign and significance, but not advised

Visualization

Forest Plot from McEwan study

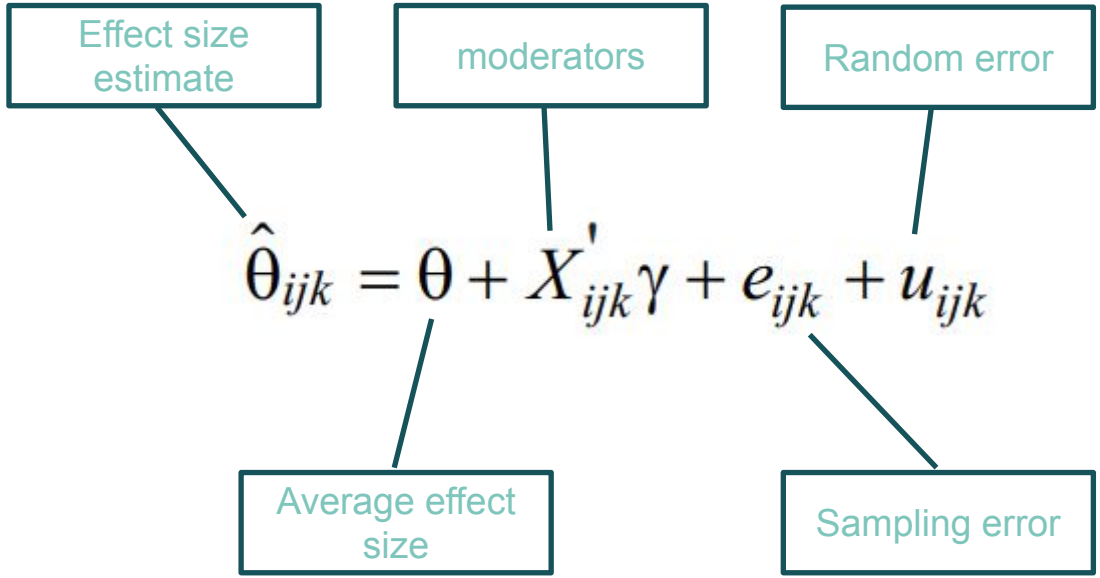


Estimating equation: random effects meta-regression

For unit i, cluster j, study k

Estimate average effect size

- Controlling for covariates/moderators
- Taking into account the precision of each contributing effect size by using inverse variance weights
- Can also weight for number of estimates per study



Weight = $\frac{1}{v_{ijk} + \hat{\sigma}_{\theta}^{2}} \times \frac{1}{n_{ijk}}$

Results in table format

Author grouped interventions and categories and sub-categories

Note the clustering of effect sizes within experiments within studies

TABLE 4

Mean effect sizes by subcategory of treatment

	Mean effect size	*p*	Sample sizes: Studies	Experiments	Effect sizes
Instructional					
Computers or technology	0.150	.003	10	13	32
Teacher training	0.123	<.001	17	23	75
Class size or composition	0.117	.018	6	8	34
Instructional materials	0.078	<.001	15	19	69
Monetary grants	−0.011	.723	4	4	14
Health or nutrition					
Food, beverages, and/or micronutrients	0.035	.054	12	12	38
Deworming drugs	0.013	.388	5	7	22
Incentives					
Contract or volunteer teachers	0.101	<.001	8	11	41
Student/teacher performance incentives	0.089	.044	8	9	26
School management or supervision	0.055	.168	5	10	32
Informational treatments	0.049	.240	7	8	28

Note. The mean effect size is obtained with Equation (6). The *p* value is obtained with the wild cluster bootstrap-*t*, clustering by the number of studies. A single study may include more than one experiment if the experiments share samples and/or treatments. A single experiment is defined as one or more treatment arms and one control group. Each experiment contains $T * F * O$ effect sizes, where T is the number of treatment arms, F is the number of follow-ups, and O is the number of outcome measures.

Metaregression results: covariates interpretation

- Intervention types
  - Ed tech
  - Teacher training
  - Class size
  - Etc.
- Intervention characteristics
  - Government vs. university implementer
  - Years implemented
  - Region
  - Etc.

TABLE 5

Random effects metaregression

	6
Materials	−0.033 (0.022)
Computers or technology	0.076** (0.037)
Grants	0.030 (0.033)
Teacher training	0.086*** (0.021)
Class size, small-group instruction, tracking	0.069* (0.039)
Food, beverage, micronutrients	0.044 (0.048)
Deworming	−0.018 (0.043)
Other health	0.070 (0.059)
Information	0.002 (0.030)
Performance incentives	0.056* (0.033)
Contract or volunteer teachers	−0.022 (0.033)
School management, supervision	−0.021 (0.030)
Months of treatment exposure at follow-up	−0.003* (0.001)
Follow-up conducted >1 month after treatment?	−0.009 (0.022)
Government implementer	−0.007 (0.048)
University/researcher implementer	−0.038 (0.033)
Post-1990s	−0.039 (0.043)
Latin America and Caribbean	0.125** (0.061)
East Asia and Pacific	0.054 (0.039)

Metaregression results: covariates interpretation

- Population characteristics
  - Younger/older children
- Outcome type
  - Math vs. Reading vs. composite
  - Government vs. commercial exam
- Study characteristics
  - Published vs. gray literature
  - High attrition
  - Differential attrition

	6
South Asia	0.005 (0.042)
Log(gross domestic product per capita in baseline year, US$ (2000))	−0.087*** (0.029)
Grades 5–8 only	−0.016 (0.031)
Both Grades 1–4 and 5–8	−0.030 (0.026)
Uncertain grades	0.025 (0.041)
Convenience sample	−0.051* (0.026)
Composite test score	−0.021 (0.035)
Math test score	−0.007 (0.015)
Commercial/international test	−0.001 (0.030)
Government or school test	−0.027 (0.033)
Uncertain test	0.005 (0.043)
Experiment is published	−0.018 (0.025)
Alternating list assignment	0.037 (0.038)
Attrition at follow-up	0.018 (0.108)
Missing indicator	−0.032 (0.025)
Absolute value of differential attrition	−0.513 (0.311)
Missing indicator	0.026 (0.018)
Constant	0.730*** (0.213)

Note. The sample for each regression incluSee the text for details of sample, weights, and estimation.
*$p < .1$. **$p < .05$. ***$p < .01$.

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	

Thank you for your attention

Steven Glazerman
sglazerman@poverty-action.org

IDEAL
Impact Data and Evidence Aggregation Library

1

Section Title

Room for a short description or subtitle.

Core Theoretical Concepts

- **Collective Action:** any action taken by an individual for the advancement of the group as a whole
- **Perceived Efficacy:** people's own beliefs about their ability to influence change and transform the environment. Not dependent on the outcomes of their efforts for change.
- **Empowerment:** ability of the group to influence external affairs without external support.
- **Sustainability:** ability of the group to continue to influence external affairs.
- **Autonomy:** ability of the group to function without external sources of support.

Summary of Results Light

- This is all to give an idea of spacing. Let's say that social media algorithms **strongly influence users' emotions** across age groups.
  - We observe this trend from engagement with Facebook, X, and Instagram.
  - These algorithms elicit strong emotional responses, eliciting both positive and negative feelings.
- However, we did not find evidence that **reducing access to social media improved mental health**.
  - This is all bogus information.
  - Users reported anxiety about being disconnected from their communities, or missing out on important information.

Summary of Results Dark

- This is all to give an idea of spacing. Let's say that social media algorithms **strongly influence users' emotions** across age groups.
  - We observe this trend from engagement with Facebook, X, and Instagram.
  - These algorithms elicit strong emotional responses, eliciting both positive and negative feelings.
- However, we did not find evidence that **reducing access to social media improved mental health**.
  - This is all bogus information.
  - Users reported anxiety about being disconnected from their communities, or missing out on important information.

Introducing a Concept Light

Why is this important? What is the value add of this to the field? This is important for IDEAL because it allows us to gauge research transparency.

Adding more context here and bolding **keywords** that are important for understanding how this concept applies.

Maybe this doesn't apply to all types of experimental design.

Limitations

- Bullets here
- Or a paragraph of things to consider

Primary Font and Font Sizes

Open Sans, which is one of the standard Google fonts. Content text and titles are both the “normal” option of the font menu.

If you don’t see this in your Font menu,click the dropdown font bar to search for Open Sans.

- Title text is set to size 28
- Contect text is set to sizes 17 and 15

Introducing a Concept Dark

Why is this important? What is the value add of this to the field? This is important for IDEAL because it allows us to gauge research transparency.

Adding more context here and bolding **keywords** that are important for understanding how this concept applies.

Maybe this doesn't apply to all types of experimental design.

Limitations

- Bullets here
- Or a paragraph of things to consider

Primary Question

Impact. What is the primary outcome we are measuring? This field will capture:

1. Variable A
2. Variable B

Secondary Questions

Mechanism. How did different mechanisms relate to these outcomes?

Context. This is all sample text to give ideas for this slide, anything here can be replaced.

Title Here

Experimental Studies

Impact. What is the primary outcome we are measuring? This field will capture:

1. Variable A
2. Variable B

Observational Studies

Mechanism. How did different mechanisms relate to these outcomes?

Context. This is all sample text to give ideas for this slide, anything here can be replaced.

Data Processing

Identifiers removed from collected data.

- Encrypted participant IDs
- More notes about de-identification
- Additional bullets and notes here
- Final point

Screenshots from SurveyCTO or papers.

Figure title here

Citation and additional context for the figure here.

Figure.

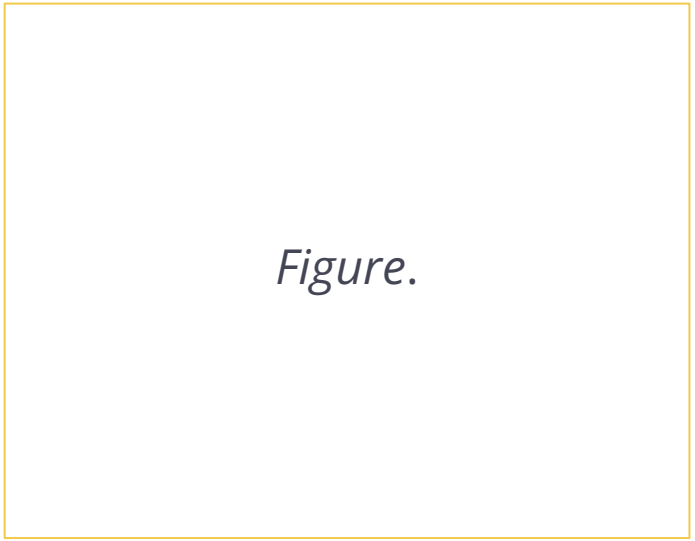
Figure/table.

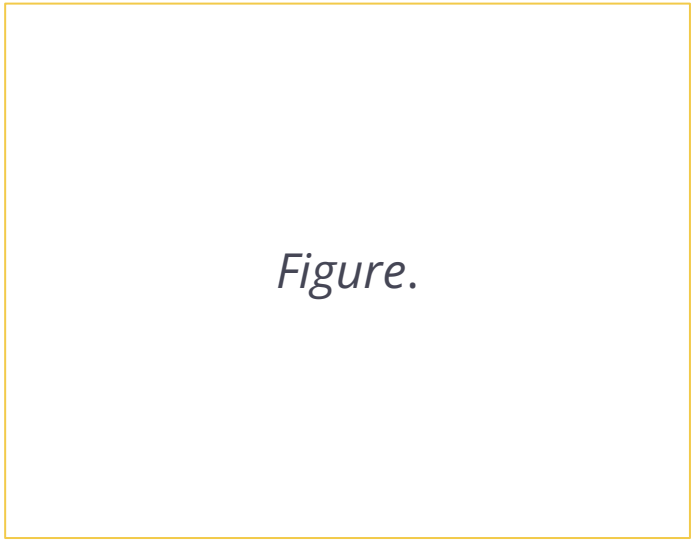
Larger Figure

Citation and additional context for the figure here.

Title Here

Central question for audience here?





Title Here

Central question for audience here?

Figure.

Figure.

Results

Title Here

Descriptor of the below timeline here. Additional context for the process outlined below.

Inputs and requirement descriptions here.

Inputs and requirement descriptions here.

Inputs and requirement descriptions here.

Inputs and requirement descriptions here.

Inputs and requirement descriptions here.

Emphasis slide

Additional information or a quote here.

Timeline with dates

First target outcome	September 5, 2025
Second target outcome	September 5, 2025
Third target outcome	September 5, 2025
Fourth target outcome	September 5, 2025

Title Here

Bucket 1

Bucket information and details.

Bucket 2

Bucket information and details.

Bucket 3

Bucket information and details.

Bucket 4

Bucket information and details.