External Validity and Meta-Analysis

Tara Slough | NYU

Scott A. Tyson | Rochester

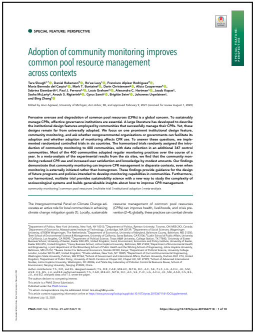
Motivation

- Evidence in Governance and Politics (EGAP) **Metaketa** initiative.
  - **Prospective meta-analysis** of coordinated field experiments
  - Goals: "accumulation" and to address "crisis of external validity."


SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials


SPECIAL FEATURE: PERSPECTIVE

Adoption of community monitoring improves common pool resource management across contexts


RESEARCH

RESEARCH ARTICLE SUMMARY

COMMUNITY POLICING

Community policing does not build citizen trust in police or reduce crime in the Global South

The challenge of meta-studies

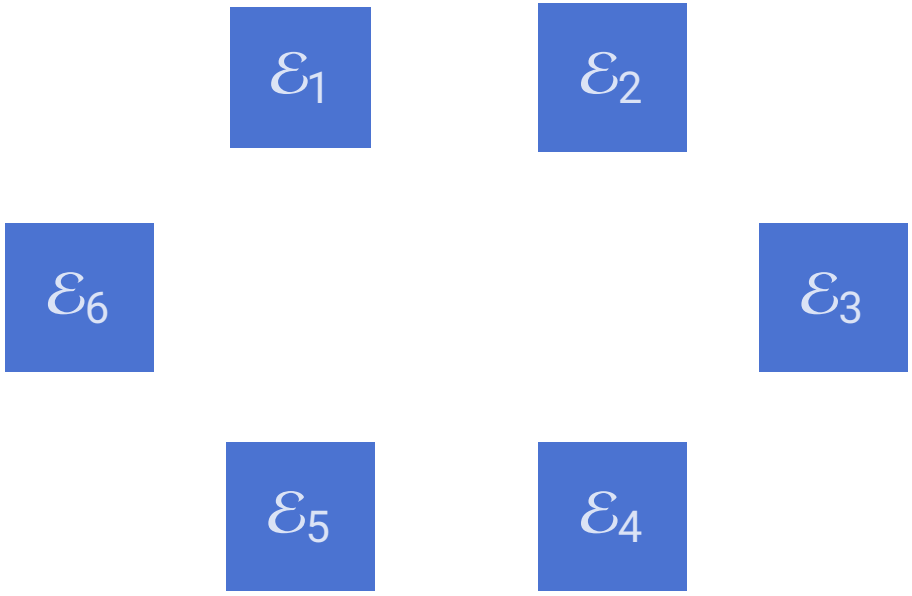
- Any two experiments on a phenomenon will produce **different estimates** of treatment effects.

- Three possible explanations for different estimates:
  1. **Statistical noise**, i.e., sampling variability.
  2. Differences in **experiment**, i.e., different outcome measures.
  3. Phenomenon is not **generalizable**.

- The challenge:
  - Noise is always present, limits our ability to assess #2 and #3.
  - A focus on estimation (in meta-analysis) largely focuses on #1.
  - **Our focus**: A (more) systematic treatment of #2 and #3.

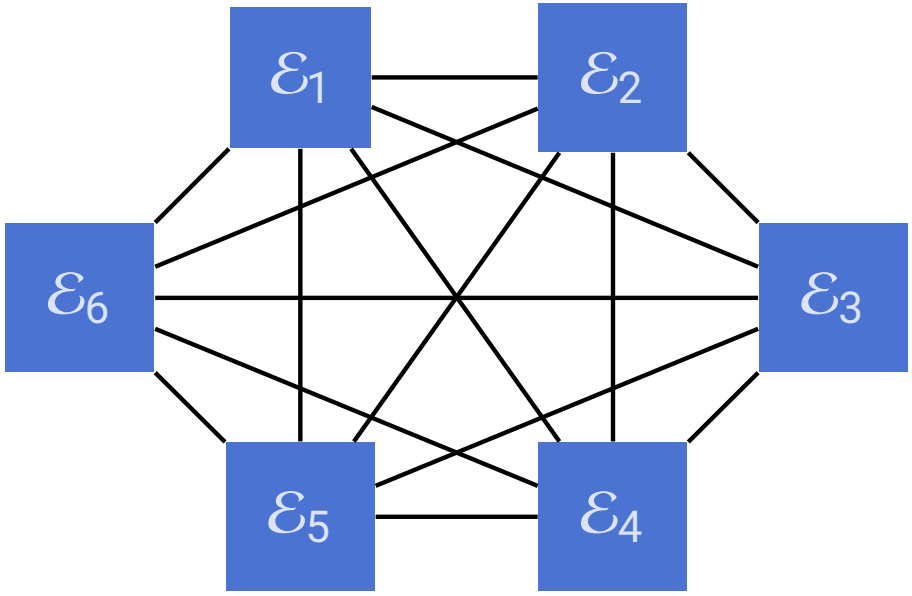
This paper

- Conceptual **framework** for meta-studies



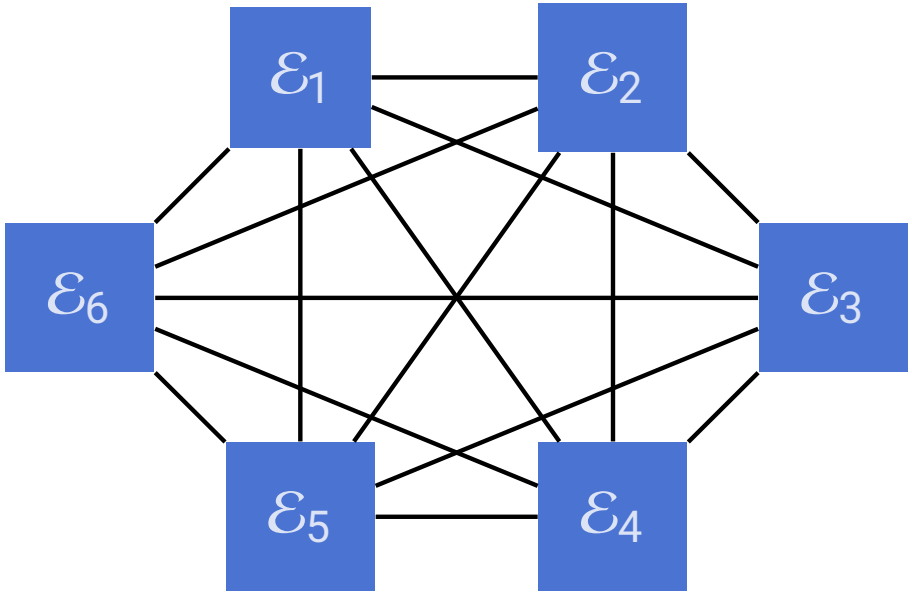
This paper

- Conceptual **framework** for meta-studies

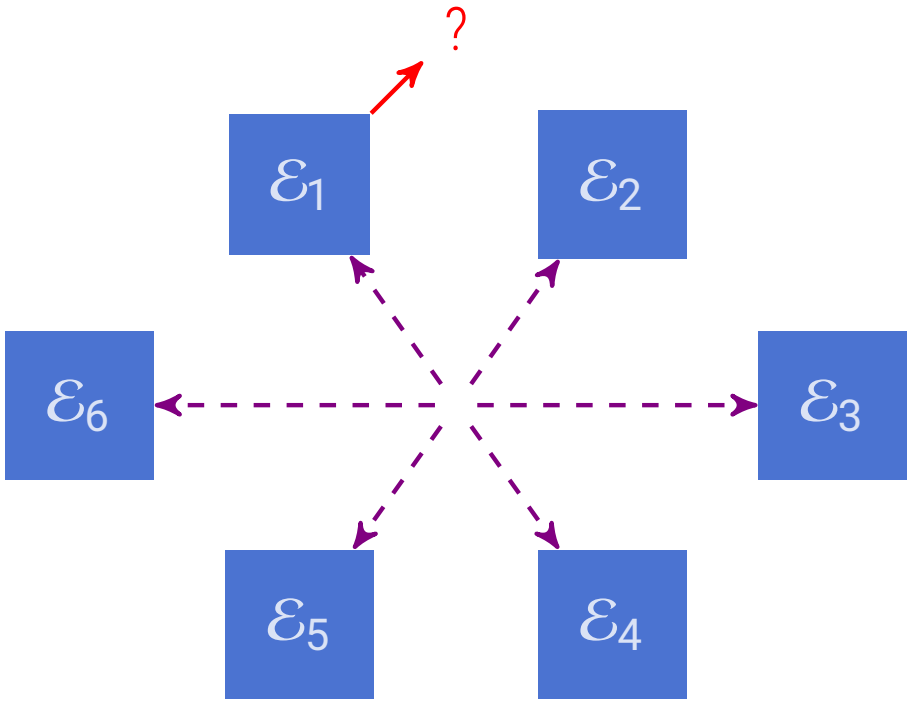


This paper

- Conceptual **framework** for meta-studies



- Objective: Understand the conditions under which multiple internally valid experiments produce **comparable** estimands
  - Design features: **harmonization**
  - **Assumptions** within and across constituent studies

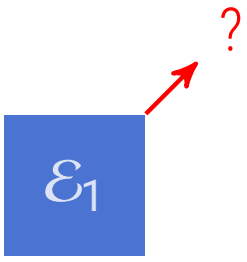


External Validity

What is external validity?

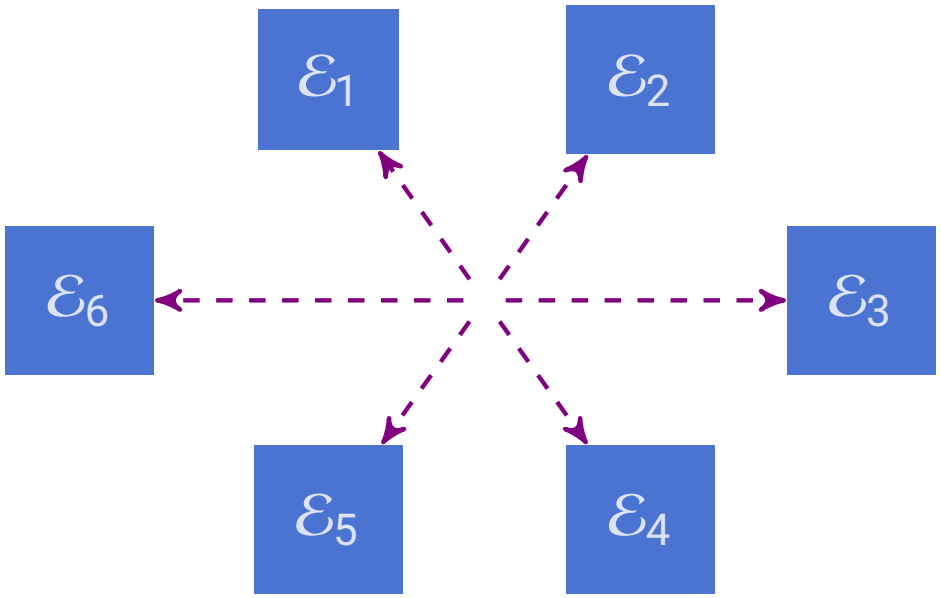
- Many views.
- We classify existing accounts into:
  - **Projective** concepts
    - External validity as a property of a single study (or estimate).
  - **Deductive** concepts
    - External validity as a relational property between a cross-section of studies.

Projective concepts of external validity

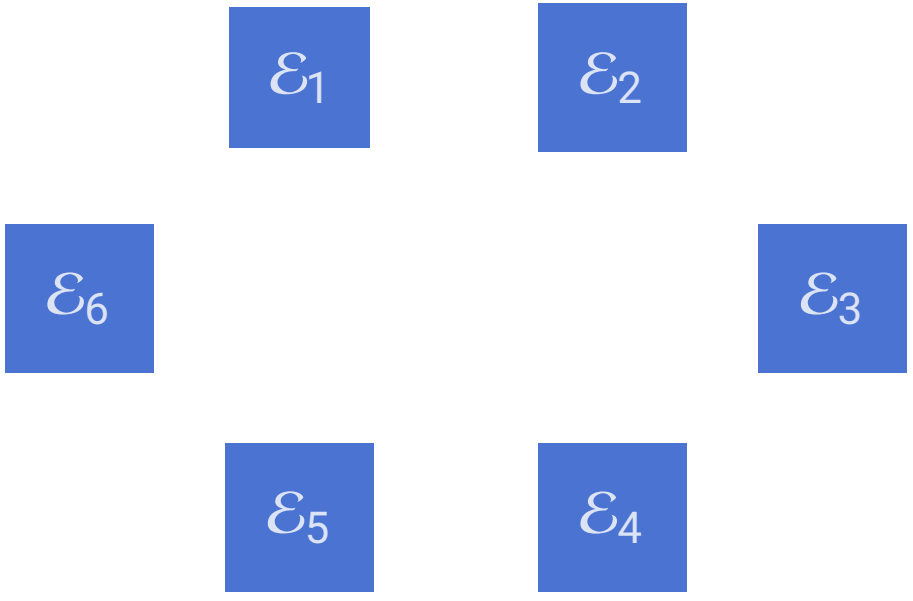


- External validity is a property of a **study** (or estimate):
  1. **Extrapolation** of estimates to different settings, samples, outcomes, or treatments. Shadish, Cook, and Campbell (2002); Fariss and Jones (2018)
  2. **Transportability** of estimates to a different setting. Pearl and Bareinboim (2014)
  3. **Sample → population** estimands. Egami and Hartman (2020); Findley, Kikuta, and Denly (2021)
  4. **Parallelism** between findings in artificial (lab) and natural (field) settings, also between methods. Smith (1982); Guala (2005); Pritchett and Sandefur (2015)

Deductive concepts of external validity



- External validity as a **relational** property between a cross-section of studies. Lucas (2003); Gailmard (2021)
  - No specific definitions in the literature.
- Meta-study practitioners seem to invoke a deductive concept of external validity:
  - If not, why spend the money/time to do multiple studies?
  - Much cheaper, easier, to extrapolate from a single study.



Framework: Constituent Studies

Studies, Meta-Analyses

- The objective:
  - of a **study**: measure the effect of a mechanism where it may be present.
  - of a **meta-analysis**: combine the findings from different studies that pertain to the same mechanism in multiple contexts or samples.

Building blocks: individual studies

- Three elements in a study designed by the researcher:
  1. A **setting**, $\theta \in \Theta$
     - Θ represents settings where a mechanism could operative, or within "scope conditions" of theory/argument
  2. A set of **measurement strategies**, M
     - Set of outcome measures that (may) reveal presence of mechanism.
  3. A **contrast**, $C = \{(\omega', \omega'') | \omega', \omega'' \in \Omega\}$
     - $\Omega \in \mathbb{R}$ is the set of possible instruments
     - Think of ω' as control, ω'' as treatment.

- Definitions:
  - A **study** is a triple, $\mathcal{E} = \{m, (\omega', \omega''), \theta\} \in M \times C \times \Theta$.
  - A **meta-study** is a collection of studies, $\mathcal{M}(\mathcal{I}) = \{\mathcal{E}_i\}_{i \in \mathcal{I}}$.

Treatment Effects

- Each study estimand is some **treatment effect**:
  - For a study $\mathcal{E} = \{m, (\omega', \omega''), \theta\}$, a treatment effect is a smooth mapping $\tau_m(\omega', \omega'' \mid \theta) : M \times C \times \Theta \to \mathbb{R}$.
  - We assume that the derivative of T has full rank for almost all contrasts.

- Relationship to potential outcomes:

$$\tau_m(\omega', \omega'' \mid \theta) = f(Y_m(\omega'')|\mathcal{D}, \theta) - f(Y_m(\omega')|\mathcal{D}, \theta)$$

  - $f(\cdot)$: some operator, usually expectation or quantile function.
  - $\mathcal{D}$: set of units for which (investigator thinks) the mechanism is operative.

Example

- Lind: *A Treatise on the Scurvy* (1753)
  - Early medical experiment on the effects of citrus (lemon + orange) on scurvy on ailing seamen.
  - Noted for clarity of the mechanism, **Vitamin C**.

Vitamin C → Scurvy

Building blocks: Example

- Simplifying Lind's original study:
  - Setting: the **ship** upon which the experiment was conducted in 1747.
  - Contrast: **lemon + orange treatment** vs. **pure control** ("do nothing").
  - Measurement strategy: indicator for the **incidence** of any scurvy symptom.

- Treatment effects:
  - Experiment was conditioned on seamen having scurvy symptoms.
  - We would expect the treatment to work on all participants, thus $\mathcal{D}$ includes all units.
  - If we care about averages, then $\tau_m(\omega', \omega'' \mid \theta)$ is the **ATE**.

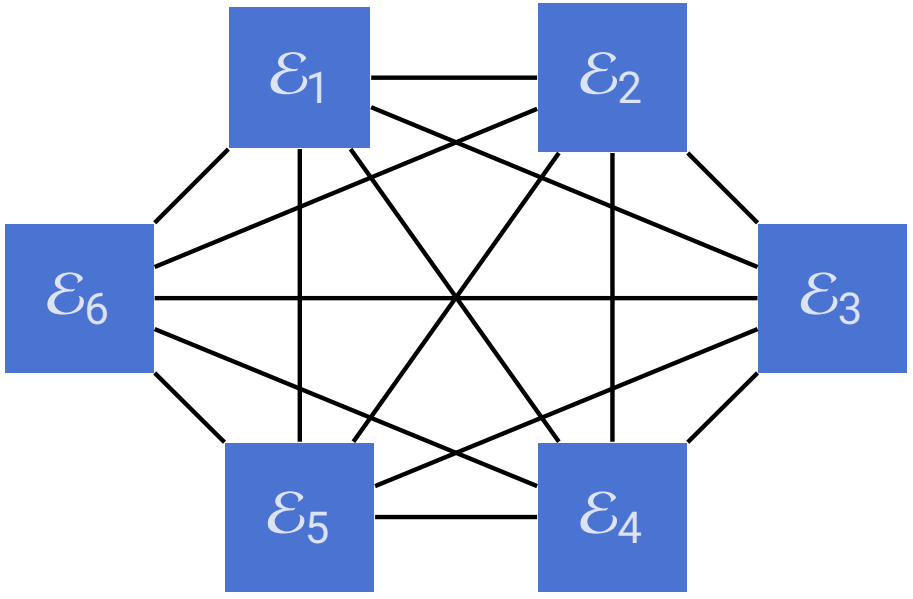
Divergent Validity

- **Divergent validity** holds between measurement strategies $m \in M$ and $m' \in M$, if

$$\tau_m(\omega', \omega'' \mid \theta) \neq \tau_{m'}(\omega', \omega'' \mid \theta).$$

- Important if we want to **combine** studies.

- Concretely, if divergent validity holds, we should expect different treatment effects if we use different measurement strategies, i.e.:
  - Any scurvy symptom
  - Bloody gums



Framework: Combining Studies

The Goal: Comparability

- Studies $\mathcal{E}_1 = \{m_1, (\omega_1', \omega_1''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2', \omega_2''), \theta_2\}$, are **comparable** if:

$$\tau_{m_1}(\omega_1', \omega_1'' \mid \theta_1) = \tau_{m_2}(\omega_2', \omega_2'' \mid \theta_2).$$

  A meta-study has **constituent comparability** if all constituent studies *i* in $\mathcal{M}(\mathcal{I})$ are comparable.

- Summary: two studies are comparable when they allow us to make apples-to-apples comparisons.

Comparability involves:

1. Cross-study design decisions: **harmonization**.

2. Assumptions about the manifestation of the mechanism as a treatment effect: **external validity**.

Design Elements: Harmonization

- Two forms of design **harmonization**

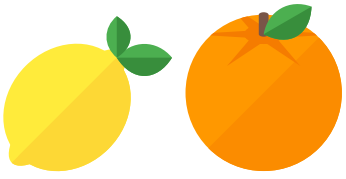
  1. Studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are **contrast harmonized** if $(\omega_1', \omega_1'') = (\omega_2', \omega_2'')$ in almost every setting.
     - Metaketas focus only on treatment harmonization, or ensuring that $\omega_1'' = \omega_2''$.

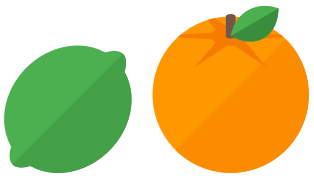
  2. Studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are **measurement harmonized** if $m_1 = m_2$ for almost every contrast and at almost every setting.
     - Measurement harmonization is a more fundamental issue than current treatments of rescaling/normalizing outcomes across sites.

- A meta-study $\mathcal{M}(\mathcal{I})$ is **harmonized** if every constituent study is contrast and measurement harmonized.

Harmonization Examples

- These two experiments are **not** contrast harmonized:

$\mathcal{E}_1$:  vs. Nothing

$\mathcal{E}_2$:  vs. Nothing

- These two experiments are not measurement harmonized:
  - $\mathcal{E}_1$: m_1: incidence of any scurvy symptom
  - $\mathcal{E}_2$: m_2: incidence of bloody gums (one scurvy symptom)

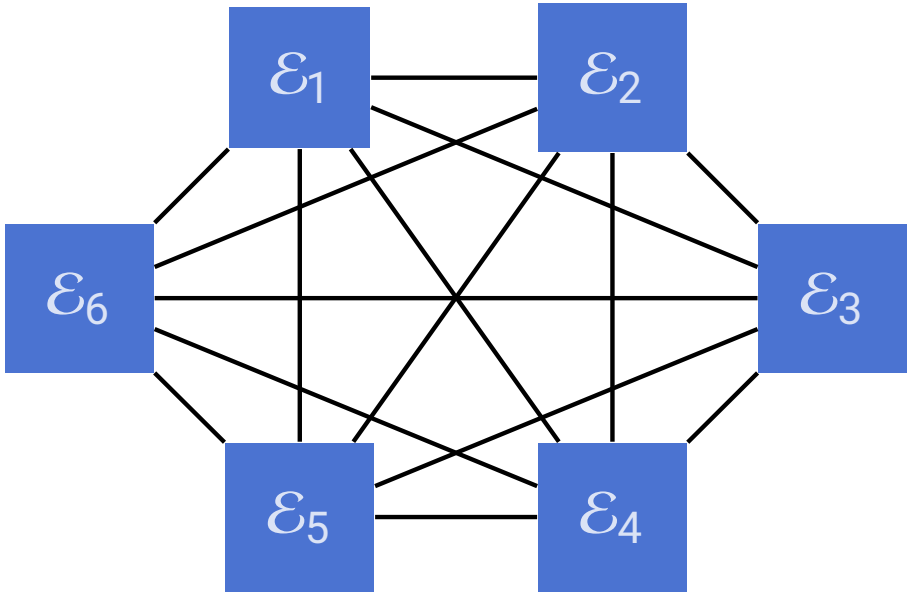
External Validity

- A mechanism has **external validity** from setting θ to setting θ' if for every measurement strategy, $m \in M$, and almost every contrast, (ω', ω''),

$$\tau_m(\omega', \omega'' \mid \theta) = \tau_m(\omega', \omega'' \mid \theta').$$

  A mechanism is **externally valid** if it has external validity across almost all settings $\theta \in \Theta$.

- Summary: A mechanism has external validity if it produces the same effect in two different settings with otherwise identical experimental conditions.



Results, Applications

Strategy

- We want to understand when meta-studies are **constituent comparable**.

- Two assumptions:
  1. **Divergent** validity (DV)
  2. **External** validity (EV)

- Allows us to ask: When are treatment effects produced by an externally valid mechanism comparable?
  - By extension, what quantity are we estimating in meta-analysis?

Result 1: given contrast harmonization

Theorem

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be contrast harmonized + EV + DV $\rightarrow$ $\mathcal{E}_1$ and $\mathcal{E}_2$ comparable iff ***measurement harmonized****.*

- Intuition: if two contrast-harmonized studies are comparable, either:
    1. Measurement strategies are harmonized.
    2. Two measurement strategies produce the same treatment effect $\rightarrow$ contradicts divergent validity.

Result 2: given measurement harmonization

Theorem

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be measurement harmonized + EV + DV $\rightarrow$ $\mathcal{E}_1$ and $\mathcal{E}_2$ comparable iff ***contrast harmonized****.*

- Intuition: suppose Lind had treated scurvy with a lime + orange on a different ship ($\mathcal{E}_2$).
  - How likely is it that a lemon + orange generated a **comparable** treatment effect to a lime + orange?
  - Limes contain less Vitamin C than lemons.
  - What is the likelihood that oranges on the lime ship precisely offset this difference in Vitamin C intake?
    - Exceedingly unlikely.

Result 3: When are studies comparable?

Theorem

*A meta-study is **constituent comparable** iff: EV + DV + measurement harmonization + contrast harmonization*

- Intuition: Follows directly from Theorems 1 and 2.
- Implication: Non-harmonized studies will not necessarily detect an externally valid mechanism even when that substantive mechanism is present in all the settings comprising the meta-analysis.

Application: Meta-Analysis

- Using meta-analysis on multiple experiments:
  - Treatment effects estimates are **reduced-form** estimates.
  - Standard meta-analysis estimators are **structural** estimators.

- **Fixed-effects** meta-analysis estimator:

$$\hat{t}_i = \mu + \epsilon_i$$

  - In the experimental context we examine, $\mu = \tau_m(\omega', \omega'' \mid \theta)$.

- **Implication**: The structural parameter μ is identified in a fixed effects meta-analysis iff:
  - Each measurement strategy satisfies divergent validity.
  - Mechanism is externally valid.
  - Every study in $\mathcal{M}(\mathcal{I})$ is both contrast and measurement harmonized.

Conclusion

- Causal meta-analyses often viewed as **agnostic** ways to cumulate knowledge.

- In additional to standard identification assumptions, for comparability of causal effects, we must:
  - Invoke additional within-site and between-site **assumptions** (divergent and external validity, respectively)
  - Pursue design **harmonization** beyond what is done in existing applications

- Limits to the agnosticism in the cumulation of evidence in meta-studies.

Thank you!

tara.slough@nyu.edu styson2@ur.rochester.edu

References

Egami, Naoki, and Erin Hartman. 2020. "Elements of External Validity: Framework, Design, and Analysis." Working paper.

Fariss, Christopher J, and Zachary M Jones. 2018. "Enhancing validity in observational settings when replication is not possible." *Political Science Research and Methods* 6 (2): 365–380.

Findley, Michael G, Kyosuke Kikuta, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science forthcomin* pp. 1–51.

Gailmard, Sean. 2021. "Theory, History, and Political Economy." *Journal of Historical Political Economy* 1 (1): 69–104.

Guala, Francesco. 2005. *The methodology of experimental economics*. Cambridge University Press.

Lucas, Jeffrey W. 2003. "Theory-testing, generalization, and the problem of external validity." *Sociological Theory* 21 (3): 236–253.

Pearl, Judea, and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science* 29 (4): 579–595.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from experiments when context matters." *American Economic Review* 105 (5): 471–75.

Shadish, William, Thomas D Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Smith, Vernon L. 1982. "Microeconomic systems as an experimental science." *The American Economic Review* 72 (5): 923–955.