Sign Congruence, External Validity, and Replication

Tara Slough—NYU Scott A. Tyson—Emory

November 2, 2022

The external validity problem

Progress: credible measurement of causal effects.

The external validity problem

Progress: credible measurement of causal effects.

Critique: "lack of" **external validity**.

The external validity problem

Progress: credible measurement of causal effects.

Critique: "lack of" **external validity**.

A central problem for:

- Policymakers who want to use evidence
- Social scientists who want to understand general phenomena

The conventional wisdom: replicate it!

> *"To address...concerns about generalization, actual* ***replication studies*** *need to be carried out. Additional experiments have to be conducted in different locations, with different teams."* *Banerjee and Duflo (2009), p. 160*

The conventional wisdom: replicate it!

> *"To address...concerns about generalization, actual* ***replication studies*** *need to be carried out. Additional experiments have to be conducted in different locations, with different teams."* *Banerjee and Duflo (2009), p. 160*

In contrast, we argue that:

- Replications not necessarily informative about external validity.

The conventional wisdom: replicate it!

> *"To address...concerns about generalization, actual* ***replication studies*** *need to be carried out. Additional experiments have to be conducted in different locations, with different teams."* *Banerjee and Duflo (2009), p. 160*

In contrast, we argue that:

- Replications not necessarily informative about external validity.
- Without attention to design, replications can mislead.

The conventional wisdom: replicate it!

> *"To address...concerns about generalization, actual* ***replication studies*** *need to be carried out. Additional experiments have to be conducted in different locations, with different teams."* *Banerjee and Duflo (2009), p. 160*

In contrast, we argue that:

- Replications not necessarily informative about external validity.
- Without attention to design, replications can mislead.
- We provide formal definitions of external validity...

The conventional wisdom: replicate it!

> *"To address...concerns about generalization, actual* ***replication studies*** *need to be carried out. Additional experiments have to be conducted in different locations, with different teams."* *Banerjee and Duflo (2009), p. 160*

In contrast, we argue that:

- Replications not necessarily informative about external validity.
- Without attention to design, replications can mislead.
- We provide formal definitions of external validity...
- ... and provide guidance on how to evaluate it.

Example: "Power to the People"

2004 Ugandan under-5 child mortality: 106/1000 live births.

Example: "Power to the People"

2004 Ugandan under-5 child mortality: 106/1000 live births.

Björkman and Svensson (2009) study effects of community oversight of public primary healthcare providers.

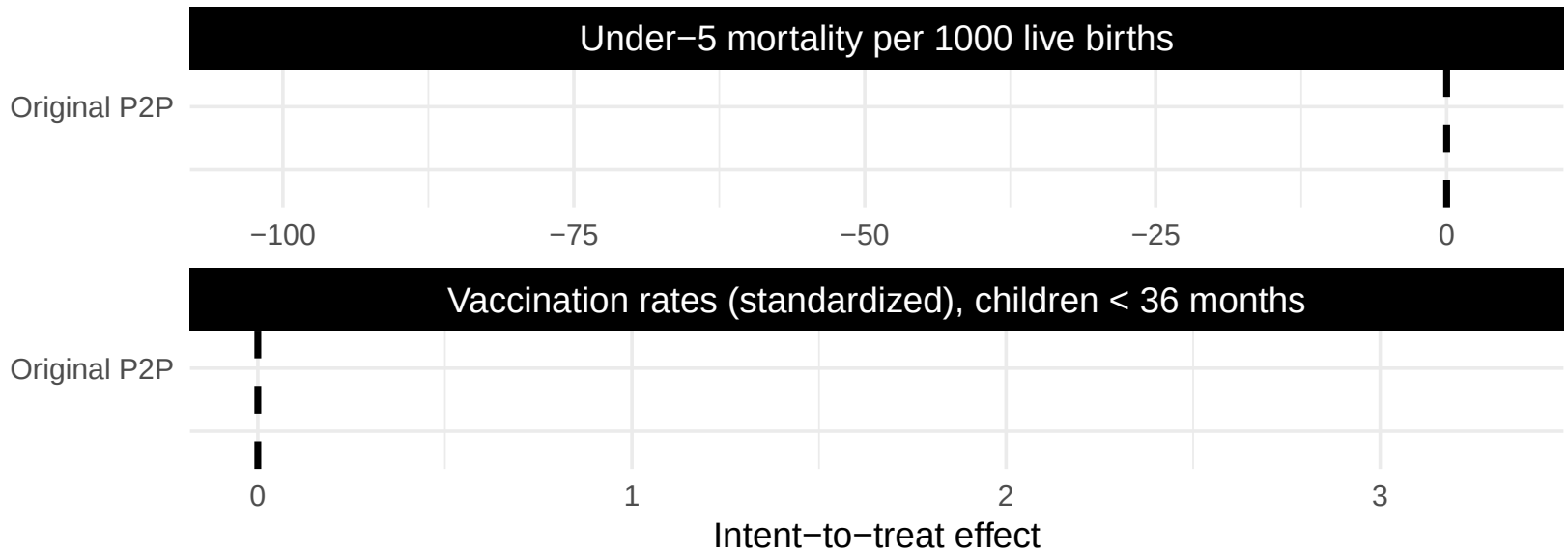
- 50 communities in rural Uganda (25 in treatment)
- One year implementation, 2004-2005

Example: "Power to the People"

2004 Ugandan under-5 child mortality: 106/1000 live births.

Björkman and Svensson (2009) study effects of community oversight of public primary healthcare providers.

- 50 communities in rural Uganda (25 in treatment)
- One year implementation, 2004-2005

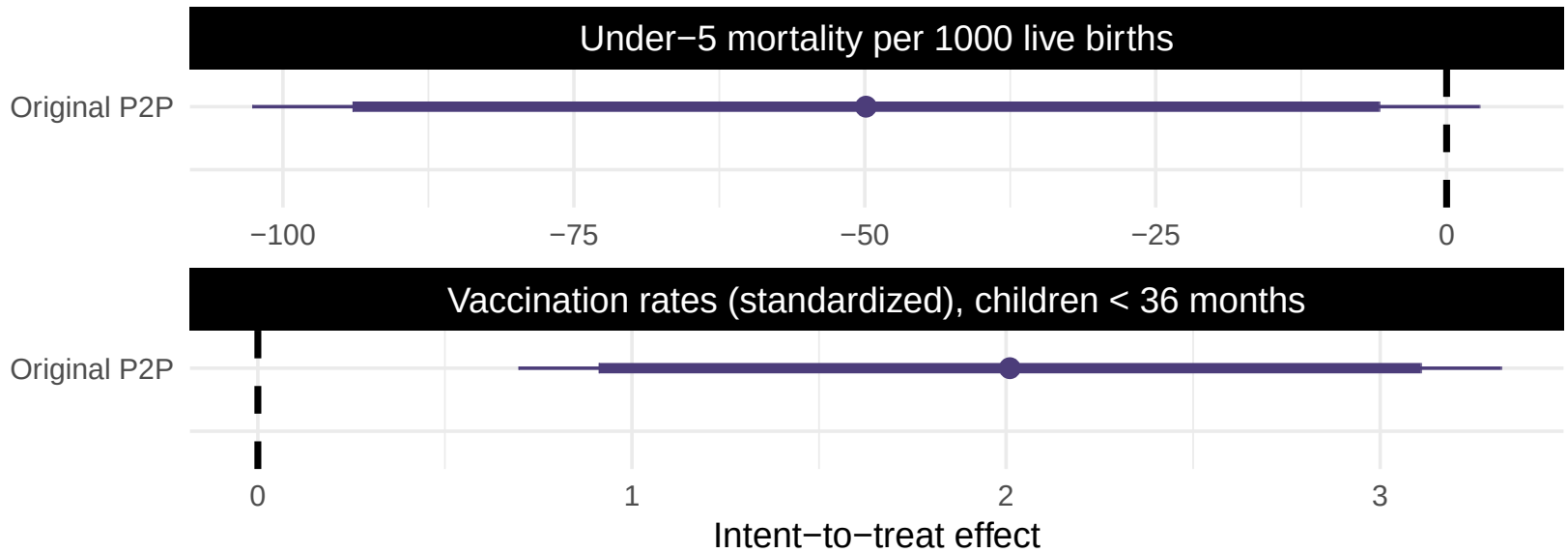


Example: "Power to the People"

2004 Ugandan under-5 child mortality: 106/1000 live births.

Björkman and Svensson (2009) study effects of community oversight of public primary healthcare providers.

- 50 communities in rural Uganda (25 in treatment)
- One year implementation, 2004-2005



Example: "Power to the People" Replication

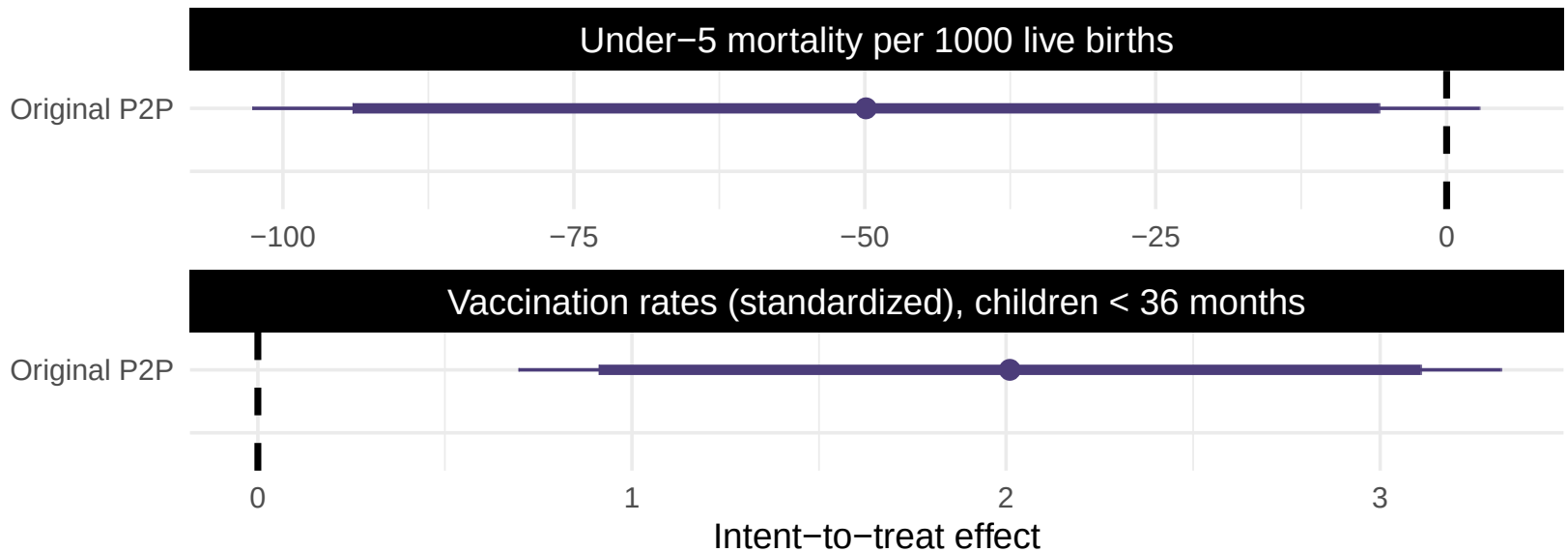
Raffler et al. (2022) sought to replicate "Power to the People":

- 187 communities in rural Uganda (92 in treatment)
- 18-month implementation, 2014-2016

Example: "Power to the People" Replication

Raffler et al. (2022) sought to replicate "Power to the People":

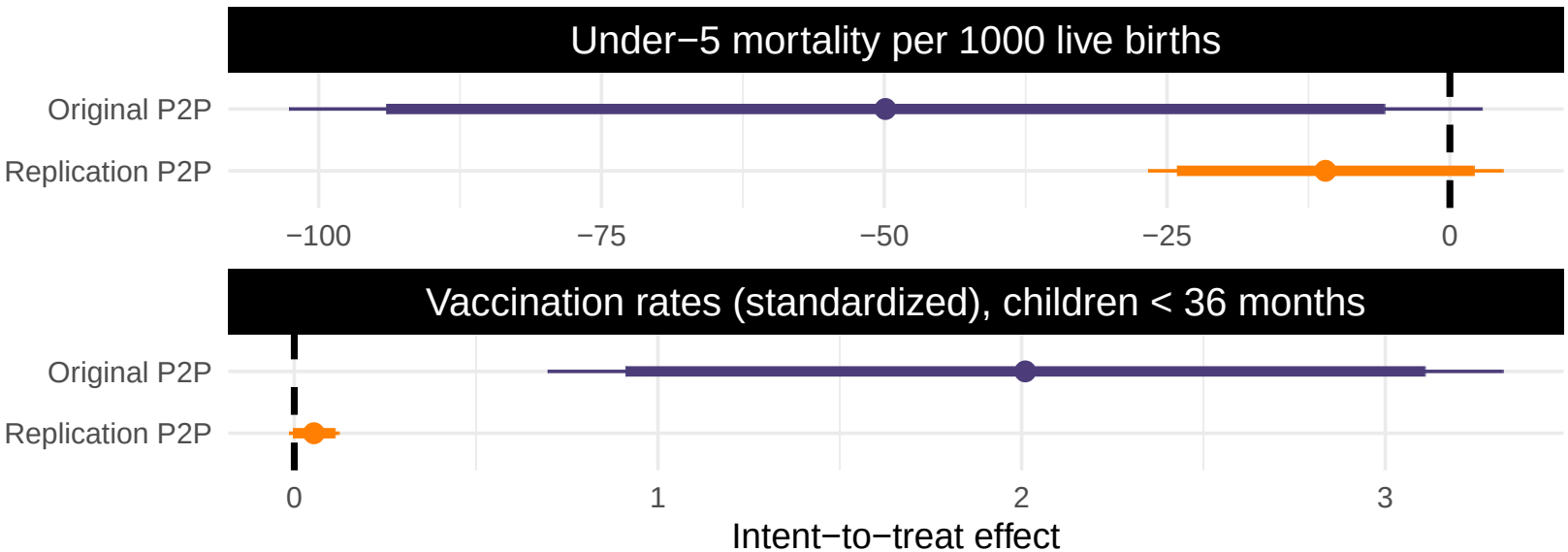
- 187 communities in rural Uganda (92 in treatment)
- 18-month implementation, 2014-2016



Example: "Power to the People" Replication

Raffler et al. (2022) sought to replicate "Power to the People":

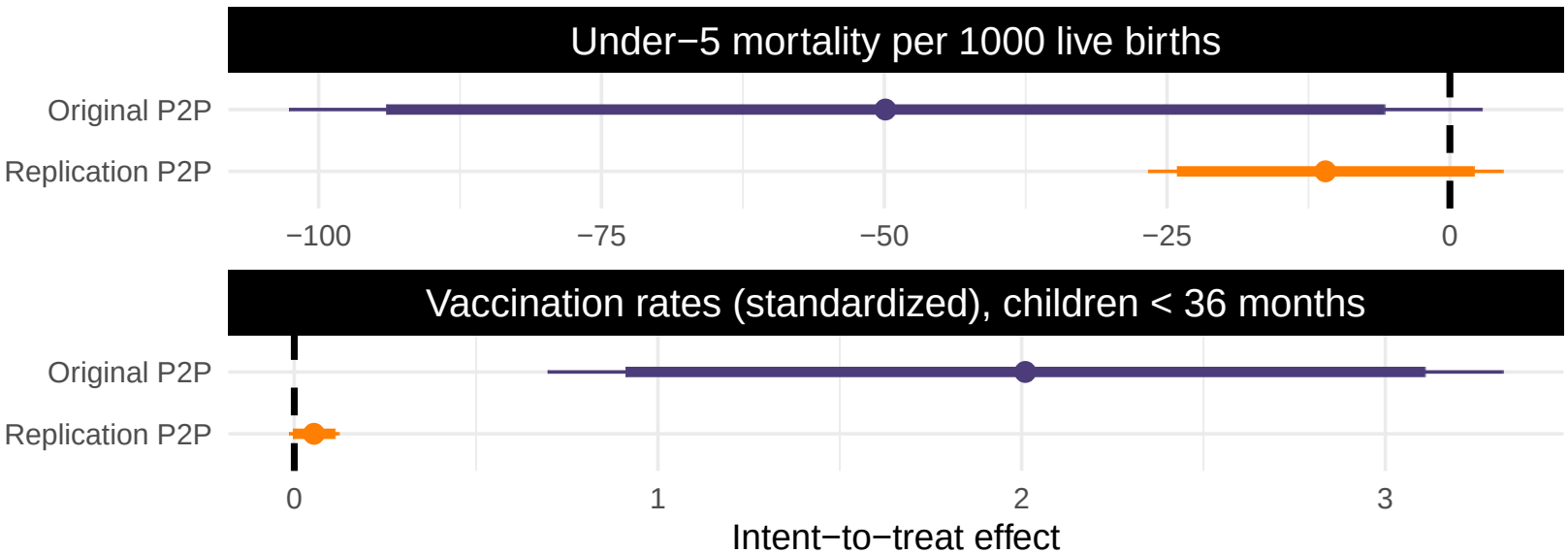
- 187 communities in rural Uganda (92 in treatment)
- 18-month implementation, 2014-2016



Example: "Power to the People" Replication

Raffler et al. (2022) sought to replicate "Power to the People":

- 187 communities in rural Uganda (92 in treatment)
- 18-month implementation, 2014-2016



Their interpretation: effect of P2P lacks **external validity**.

Overview of talk

When does the comparison of results from replication studies provide information about **external validity**?

External validity

Statistical tests

Overview of talk

Missing ingredient: **empirical targets** depend on choices of treatment(s) and outcome(s) in each study.

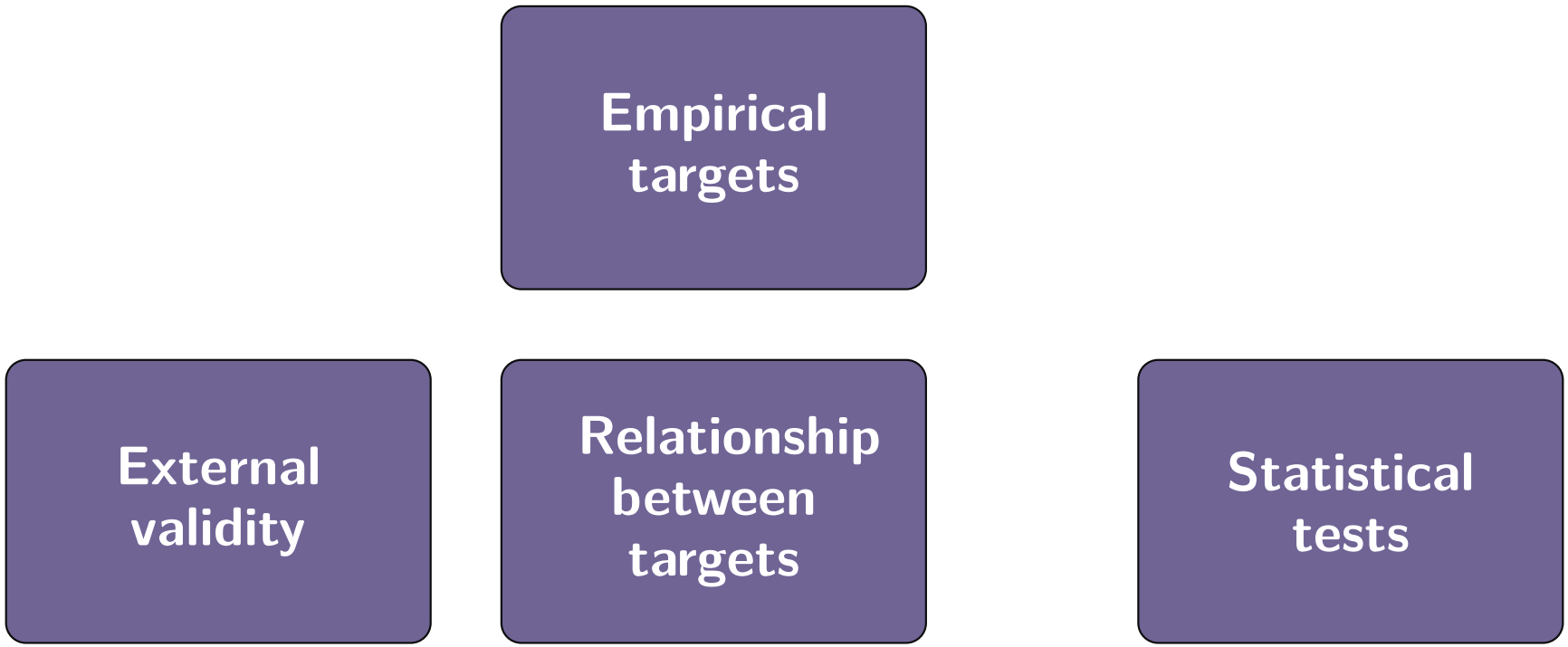
Empirical targets

External validity

Statistical tests

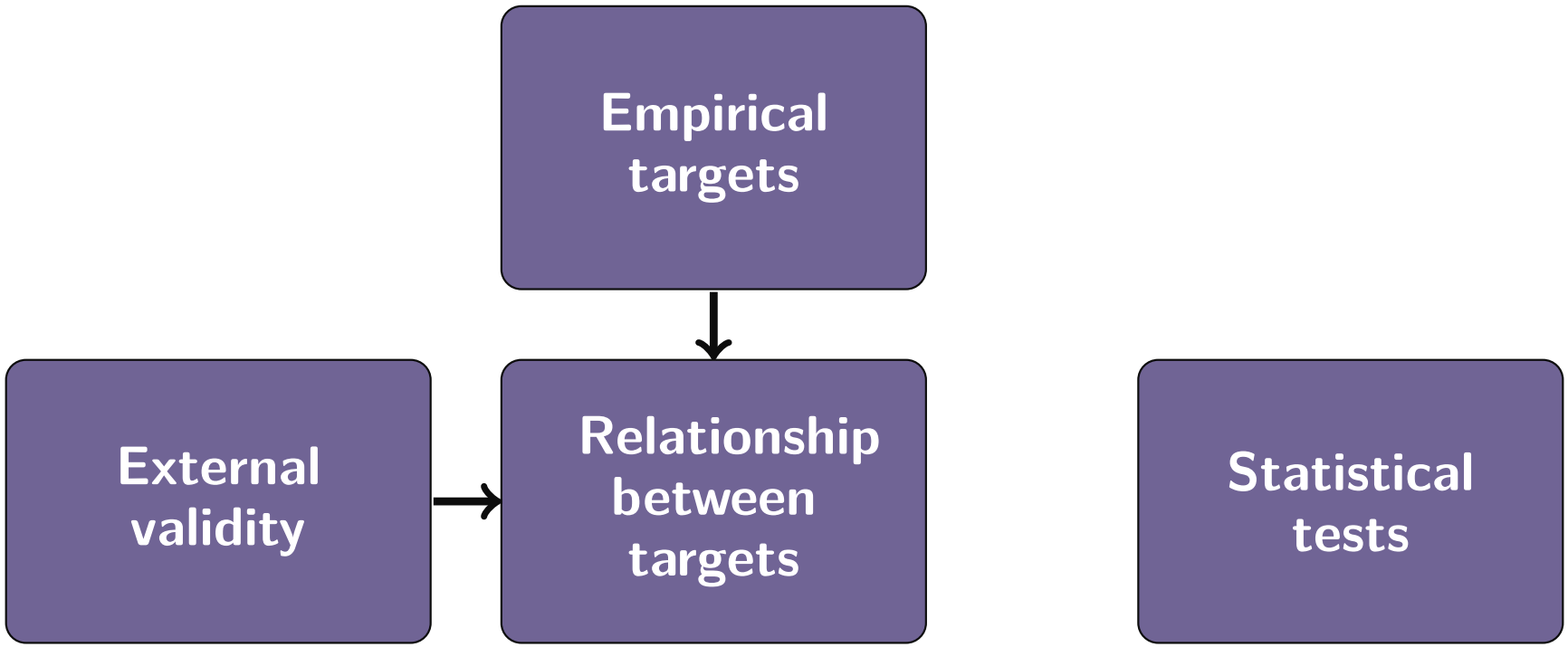
Overview of talk

To compare, we need to know how targets relate to each other.



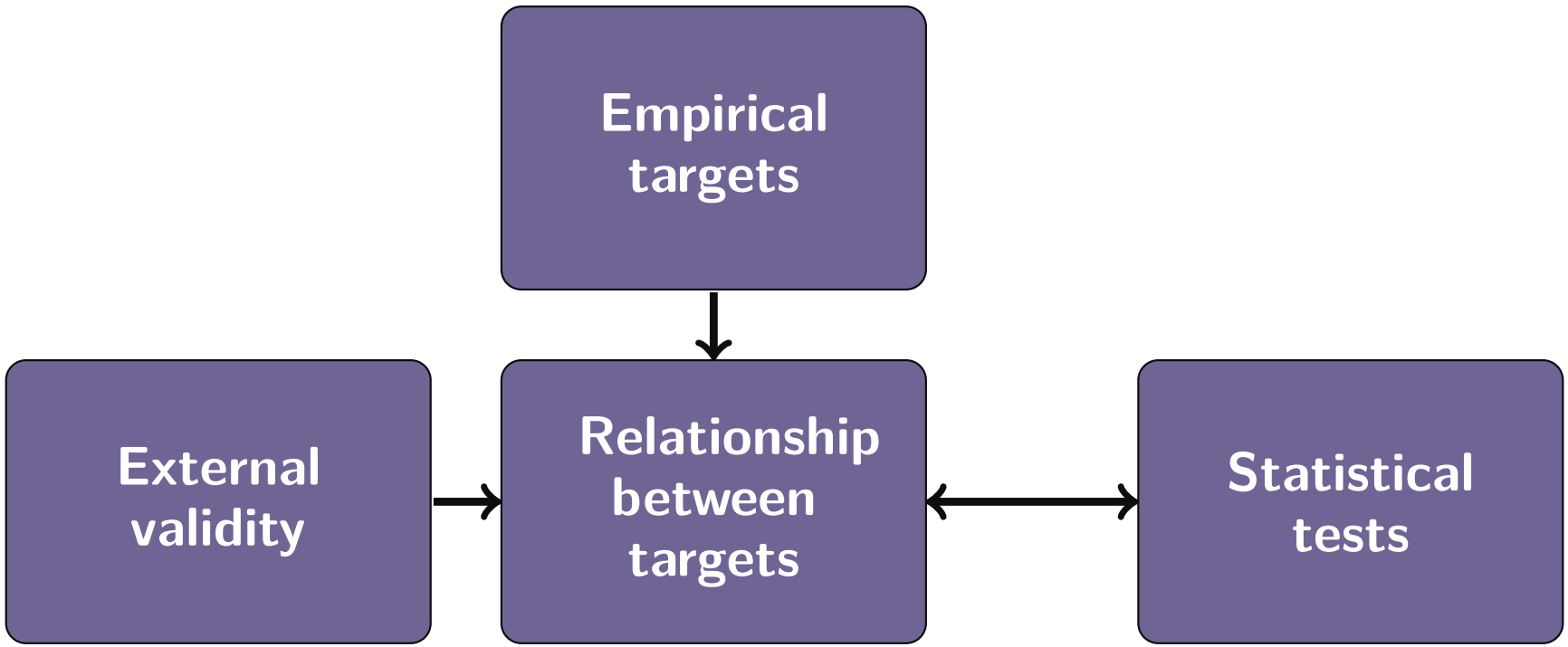
Overview of talk

Relationship between targets depends on both **external validity** of the mechanism and **empirical targets** (research design).



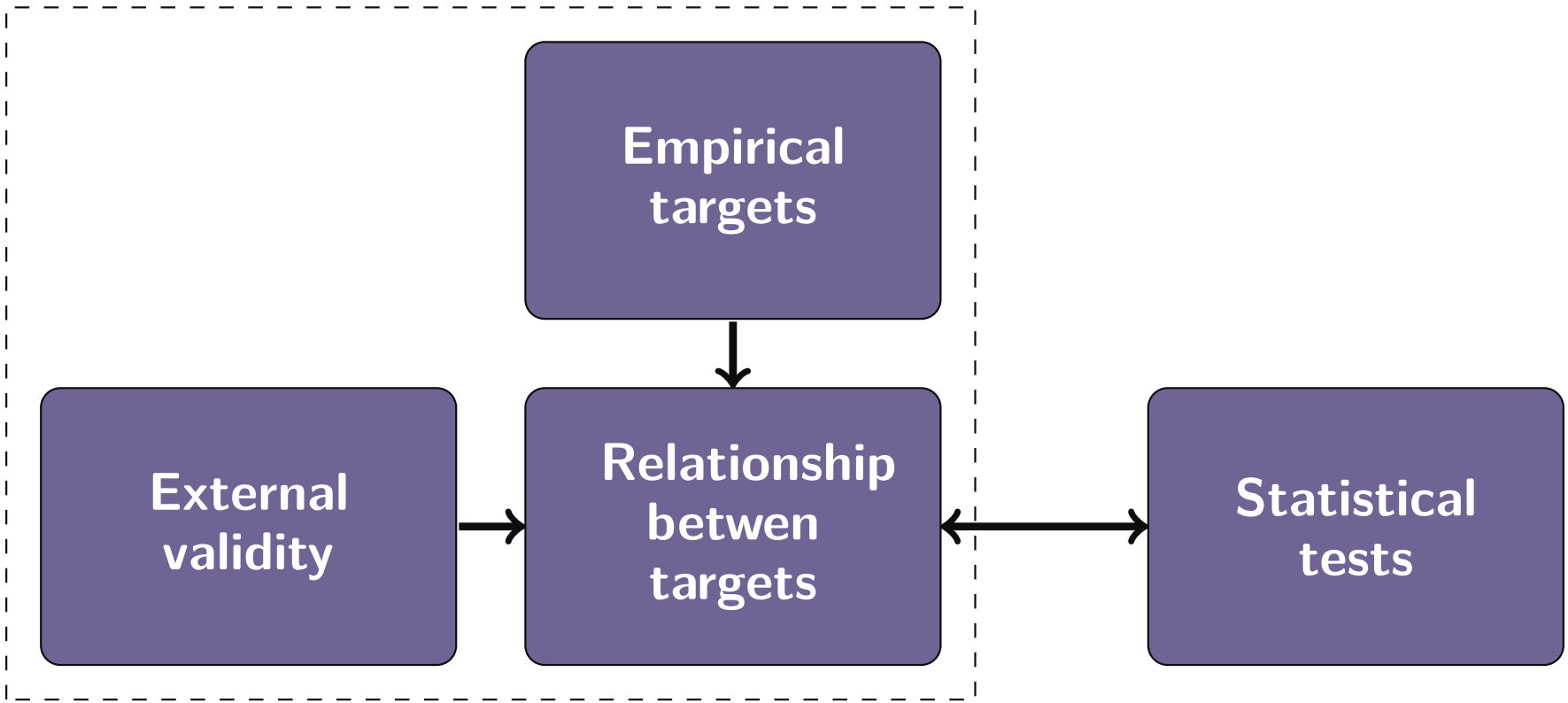
Overview of talk

We provide conditions under which statistical tests used in replications provide information about **external validity**.



Overview of talk

Framework suggests two approaches for the design of replications.



Our contribution to the literature

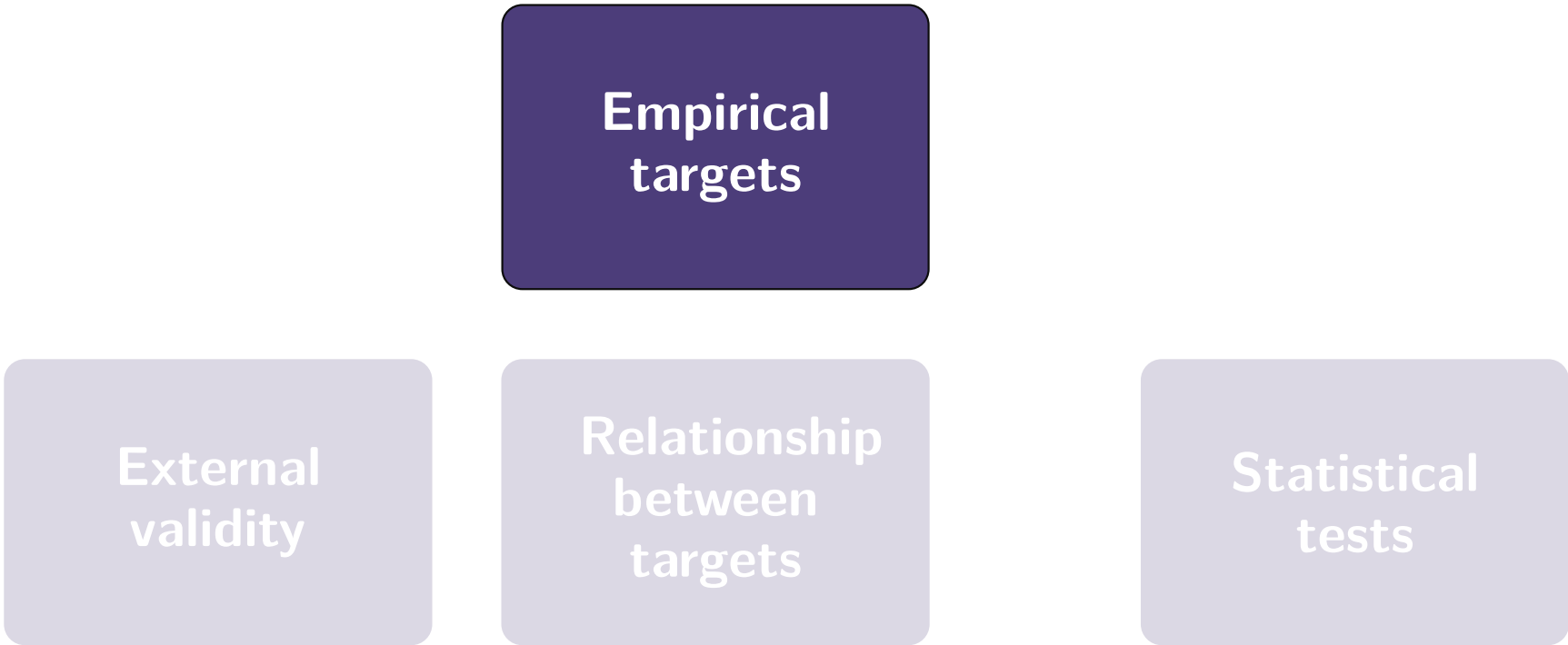
- Concepts of **external validity** (Smith, 1982; Shadish, Cook and Campbell, 2002; Guala 2005; Deaton 2010; Pearl and Bareinboim 2011, 2014; Bisbee et al., 2017; Findley, Kikuta, and Denley 2021; Dehejia, Pop-Eleches, and Samii, 2021; Egami and Hartman, 2022)
- Theoretical framework for **multi-study** research design (Meager 2019; Gechter and Meager 2021; Slough and Tyson, 2022; Wilke and Samii, 2022)
- Practical guidance for **replication** (Collins, 1992; Guala, 2005; Schmidt, 2009; Nosek and Errington, 2017)

Framework for Research Design

A Conceptual Framework for Research Design

Requires a framework that incorporates study-level and cross-study design features.

- Builds upon Slough and Tyson (2022).



A study

A study is a triple:

1. A **setting**, θ
   $\rightarrow$ Contextual features, population, time, etc.
2. A **measurement strategy**, m
   $\rightarrow$ Outcome choice and measurement components
3. A **contrast**, (ω', ω'')
   $\rightarrow$ Comparison of interest (e.g., treatment/control)

A study

A study is a triple:

1. A **setting**, θ
   $\rightarrow$ Contextual features, population, time, etc.
2. A **measurement strategy**, m
   $\rightarrow$ Outcome choice and measurement components
3. A **contrast**, (ω', ω'')
   $\rightarrow$ Comparison of interest (e.g., treatment/control)

Two studies are **harmonized** if the measurement strategy and contrast are the same.

Empirical target

Treatment effects measure the influence of a mechanism.

- Consistent with "effects of causes" view of causal research.

Empirical target

Treatment effects measure the influence of a mechanism.

- Consistent with "effects of causes" view of causal research.

The **treatment effect function**, $\tau_m(\omega', \omega'' \mid \theta)$:

Empirical target

Treatment effects measure the influence of a mechanism.

- Consistent with "effects of causes" view of causal research.

The **treatment effect function**, $\tau_m(\omega', \omega'' \mid \theta)$:

1. Is **symmetric** in contrasts: $\tau_m(\omega', \omega'' \mid \theta) = -\tau_m(\omega'', \omega' \mid \theta)$.
   $\rightarrow$ A feature of standard causal estimands.

Empirical target

Treatment effects measure the influence of a mechanism.

- Consistent with "effects of causes" view of causal research.

The **treatment effect function**, $\tau_m(\omega', \omega'' \mid \theta)$:

1. Is **symmetric** in contrasts: $\tau_m(\omega', \omega'' \mid \theta) = -\tau_m(\omega'', \omega' \mid \theta)$.
   $\rightarrow$ A feature of standard causal estimands.
2. Is **smooth** almost everywhere.
   $\rightarrow$ Facilitates analysis, not restrictive.

Empirical target

Treatment effects measure the influence of a mechanism.

- Consistent with "effects of causes" view of causal research.

The **treatment effect function**, $\tau_m(\omega', \omega'' \mid \theta)$:

1. Is **symmetric** in contrasts: $\tau_m(\omega', \omega'' \mid \theta) = -\tau_m(\omega'', \omega' \mid \theta)$.
   $\rightarrow$ A feature of standard causal estimands.
2. Is **smooth** almost everywhere.
   $\rightarrow$ Facilitates analysis, not restrictive.
3. Its derivative has **full rank** in measurement strategies and contrasts.
   $\rightarrow$ Measured effects depend on research design.

Measured treatment effects

We do not observe $\tau_m(\omega', \omega'' \mid \theta)$ directly.

Measured treatment effects

We do not observe $\tau_m(\omega', \omega'' \mid \theta)$ directly.

Our measured effect in study j is given by:

$$e_j = \tau_{m_j}(\omega_j{}', \omega_j{}'' \mid \theta_j) + \varepsilon_j^{n_j}.$$

- $\varepsilon_j^{n_j}$ is observation error

Measured treatment effects

We do not observe $\tau_m(\omega', \omega'' \mid \theta)$ directly.

Our measured effect in study j is given by:

$$e_j = \tau_{m_j}(\omega_j{}', \omega_j{}'' \mid \theta_j) + \varepsilon_j^{n_j}.$$

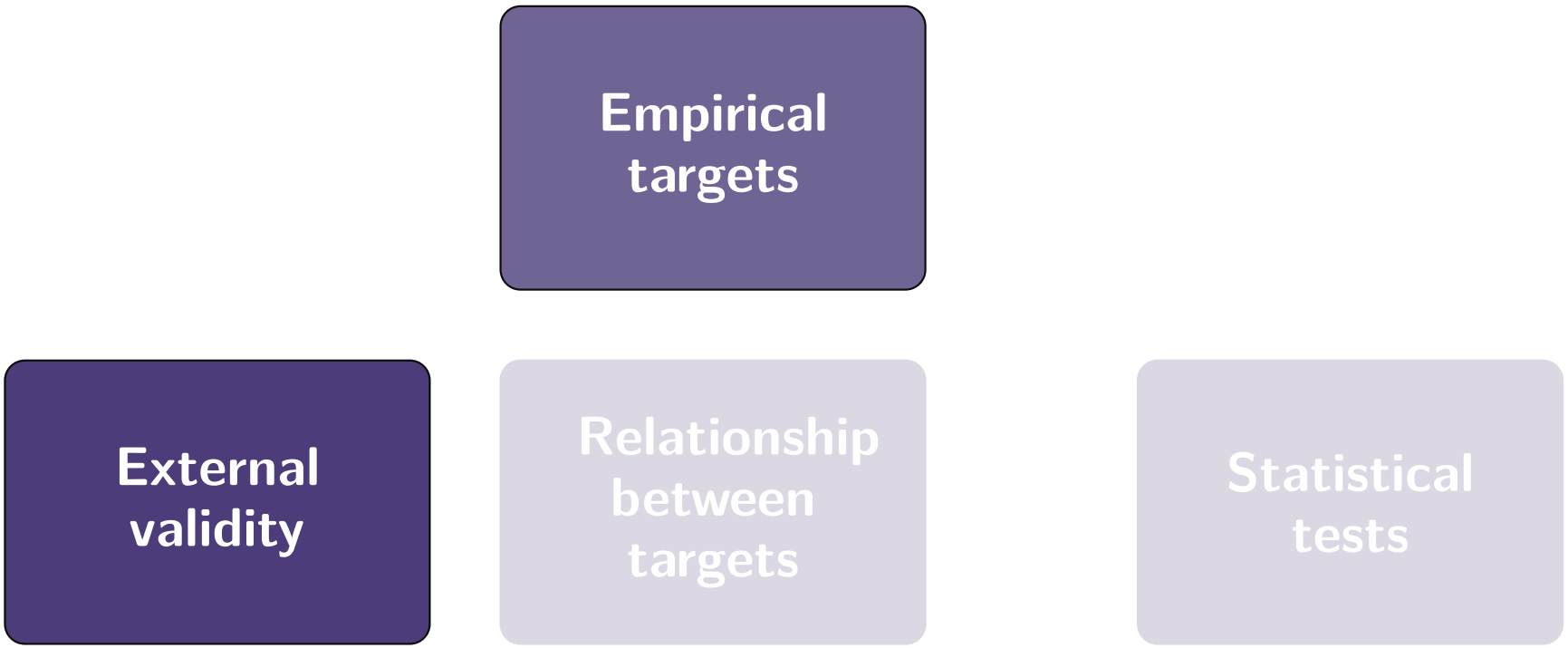
- $\varepsilon_j^{n_j}$ is observation error
- **Unbiased** when $\mathbb{E}[\varepsilon_j^{n_j}] = 0$
- **Consistent** when $\mathbb{E}(\varepsilon_i^{n_i} - \mathbb{E}[\varepsilon_j^{n_i}])^2 \to 0$ (in probability) as $n_i \to \infty$.

Concepts

Concepts of External Validity

External validity is a cluster of concepts.

- Cross-sectional concepts: external validity, sign-congruent external validity.



External Validity

Definition (Slough and Tyson, 2022)

*A mechanism has **external validity** from setting θ to setting θ' if for almost every measurement strategy and contrast,*

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

External Validity

Definition (Slough and Tyson, 2022)

*A mechanism has **external validity** from setting θ to setting θ' if for almost every measurement strategy and contrast,*

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

Fix the research design.

External Validity

Definition (Slough and Tyson, 2022)

*A mechanism has **external validity** from setting θ to setting θ' if for almost every measurement strategy and contrast,*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Why might we want a weaker concept of external validity?

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Why might we want a weaker concept of external validity?

- Mechanism is only activated for a subset of sample.

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Why might we want a weaker concept of external validity?

- Mechanism is only activated for a subset of sample.
- We have a directional prediction.

Sign-Congruent External Validity

Definition

*A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy and contrast*

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

Fix the research design.

- ...but consider the effect of mechanism in different settings.

Why might we want a weaker concept of external validity?

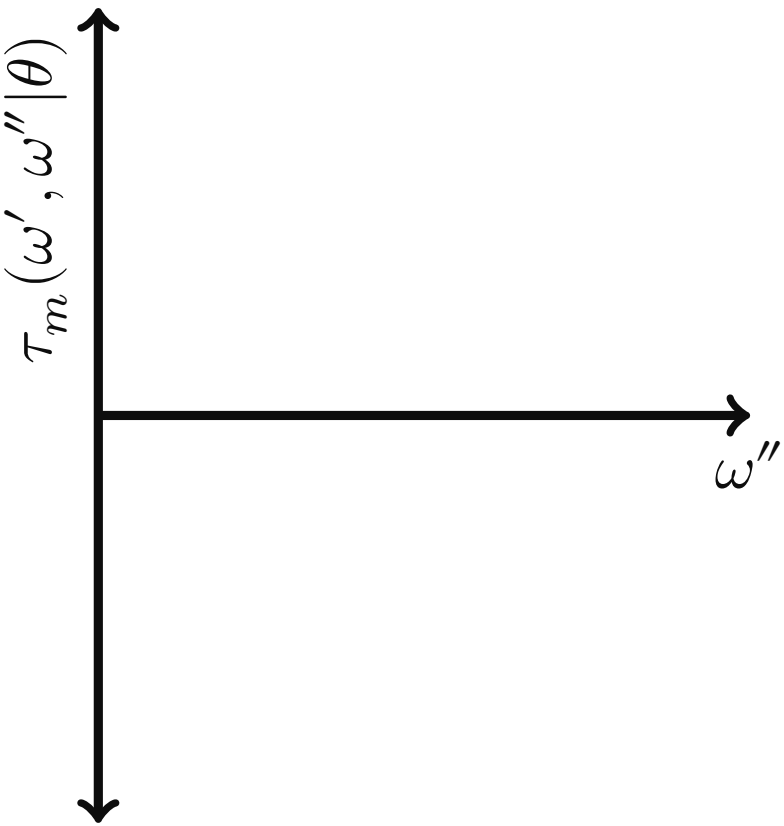
- Mechanism is only activated for a subset of sample.
- We have a directional prediction.
- (Evaluated by comparison that we make.)

Concepts of external validity: an illustration

For a fixed measurement strategy and "control" instrument:

Concepts of external validity: an illustration

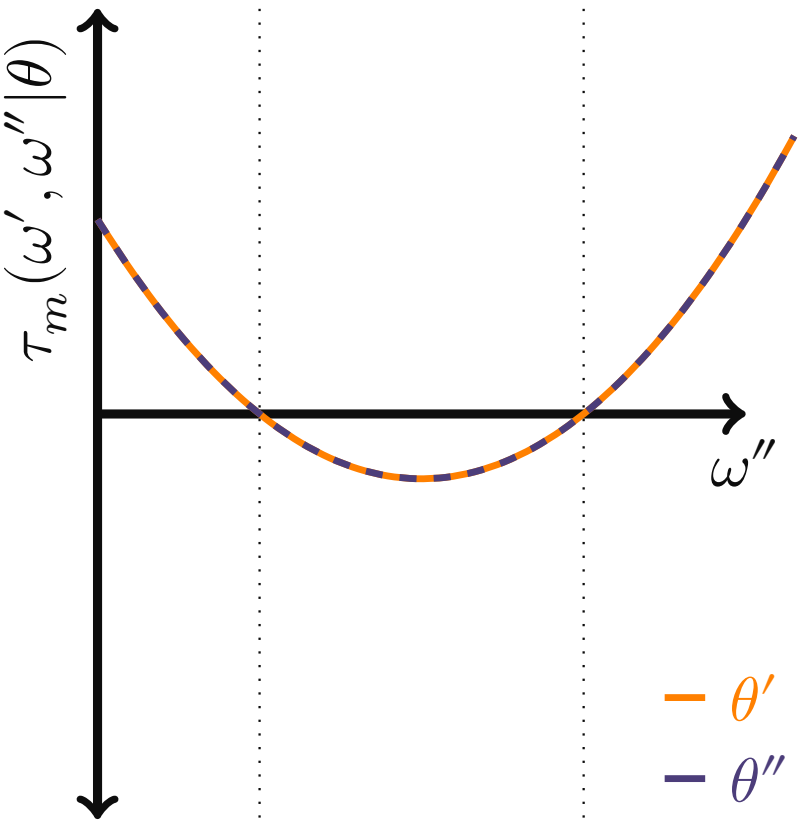
For a fixed measurement strategy and "control" instrument:



Concepts of external validity: an illustration

For a fixed measurement strategy and "control" instrument:

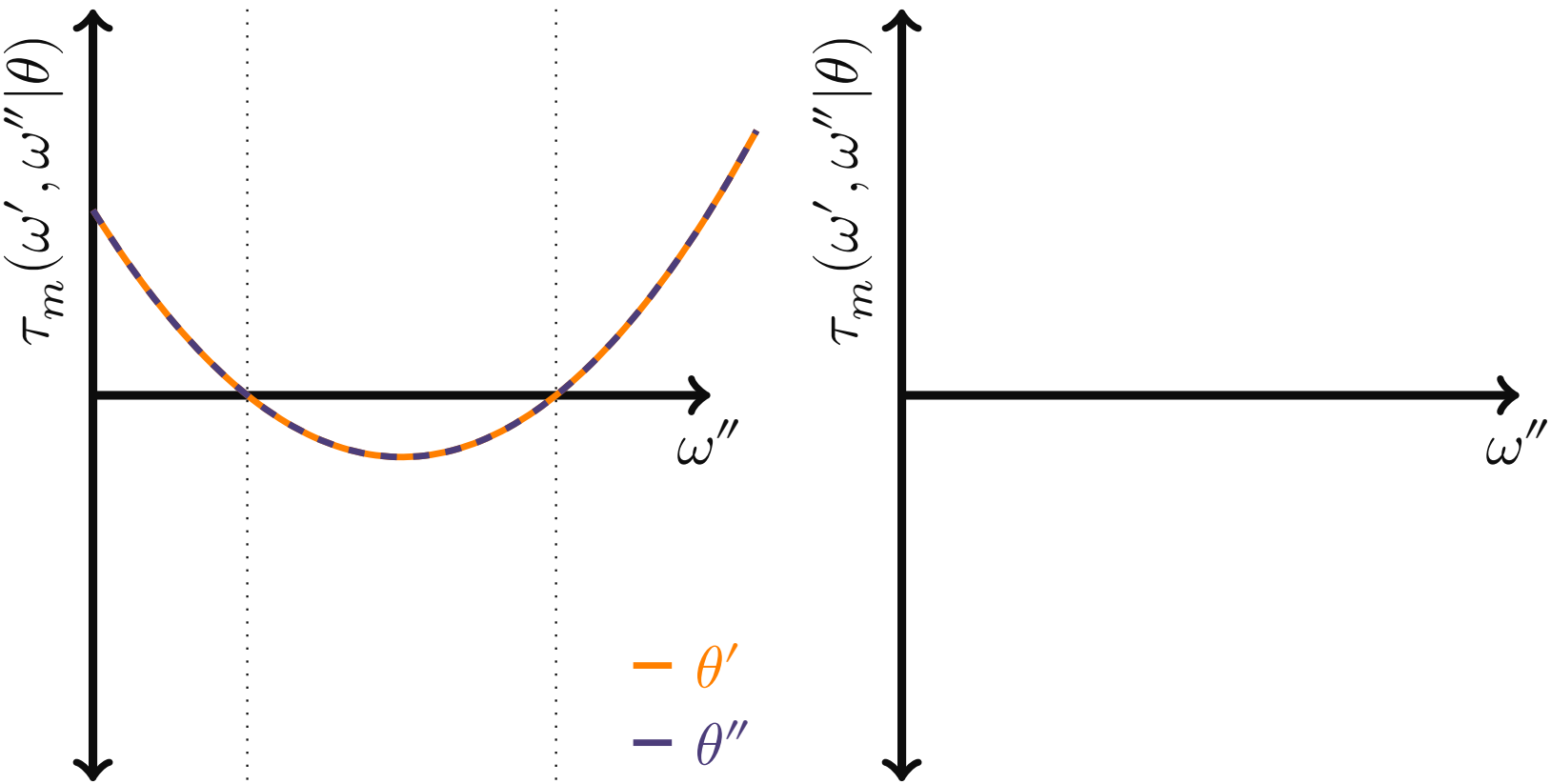
A. External Validity



Concepts of external validity: an illustration

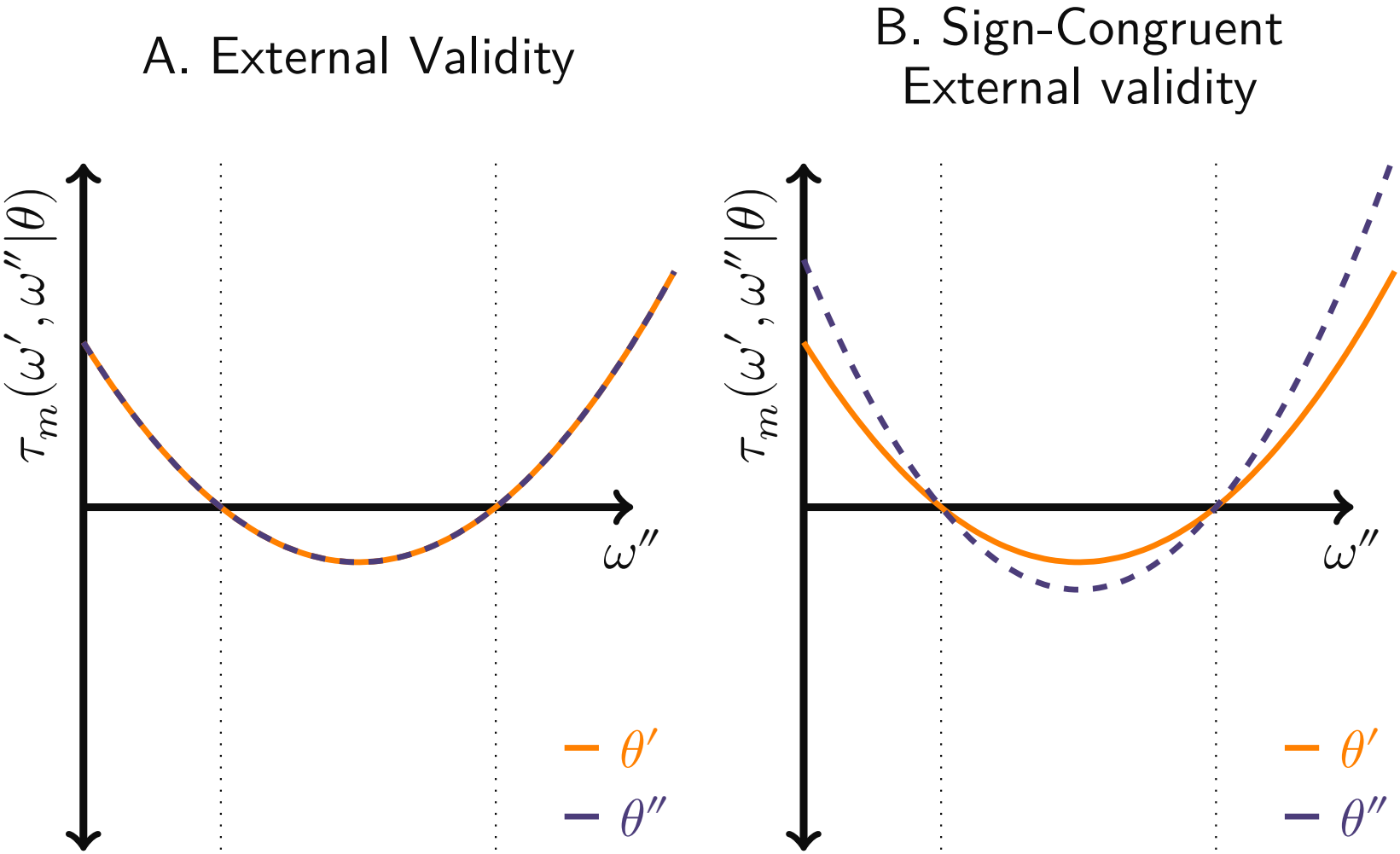
For a fixed measurement strategy and "control" instrument:

A. External Validity

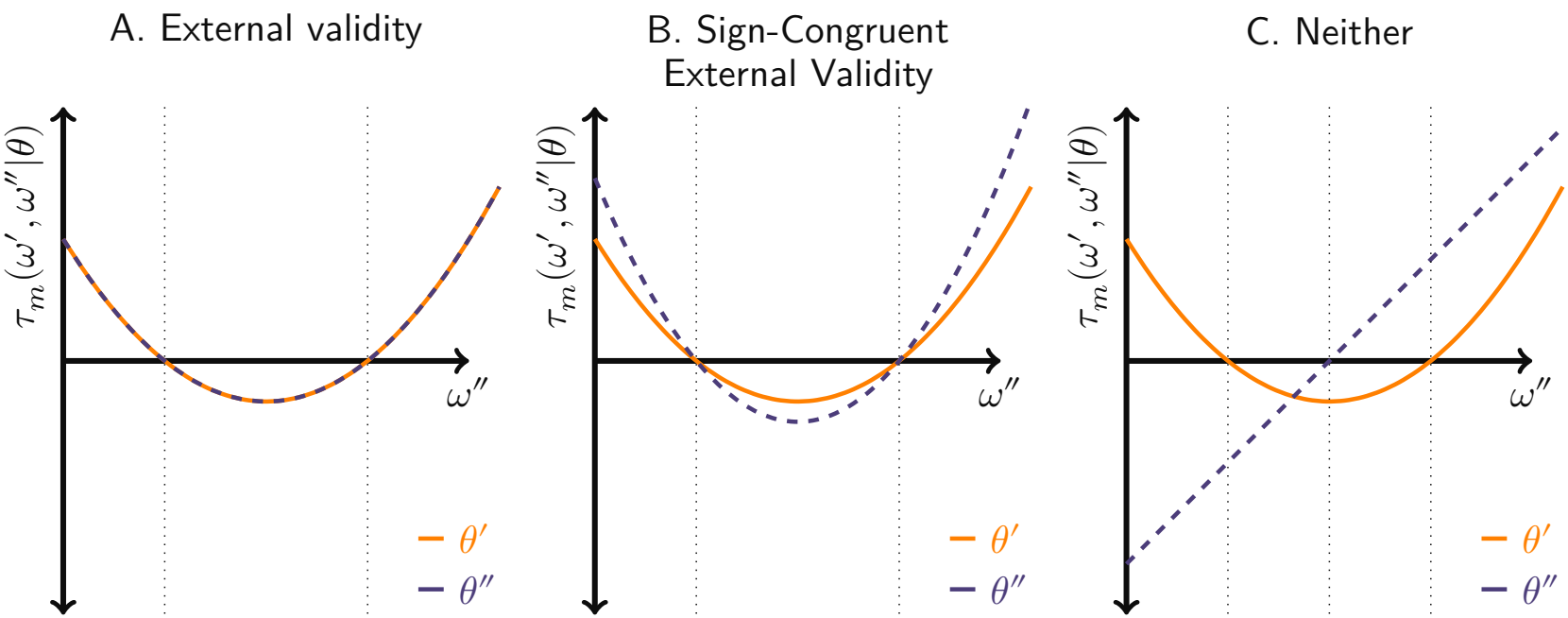


Concepts of external validity: an illustration

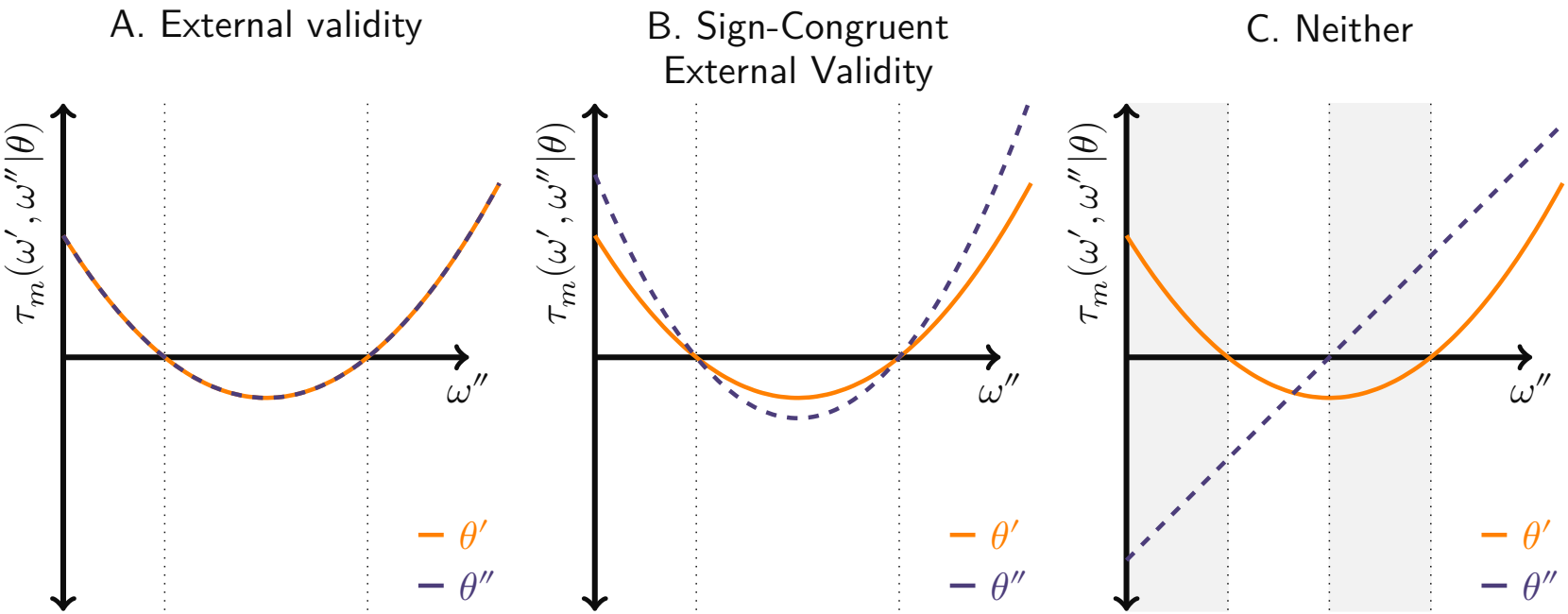
For a fixed measurement strategy and "control" instrument:



Comparing notions of external validity



Comparing notions of external validity

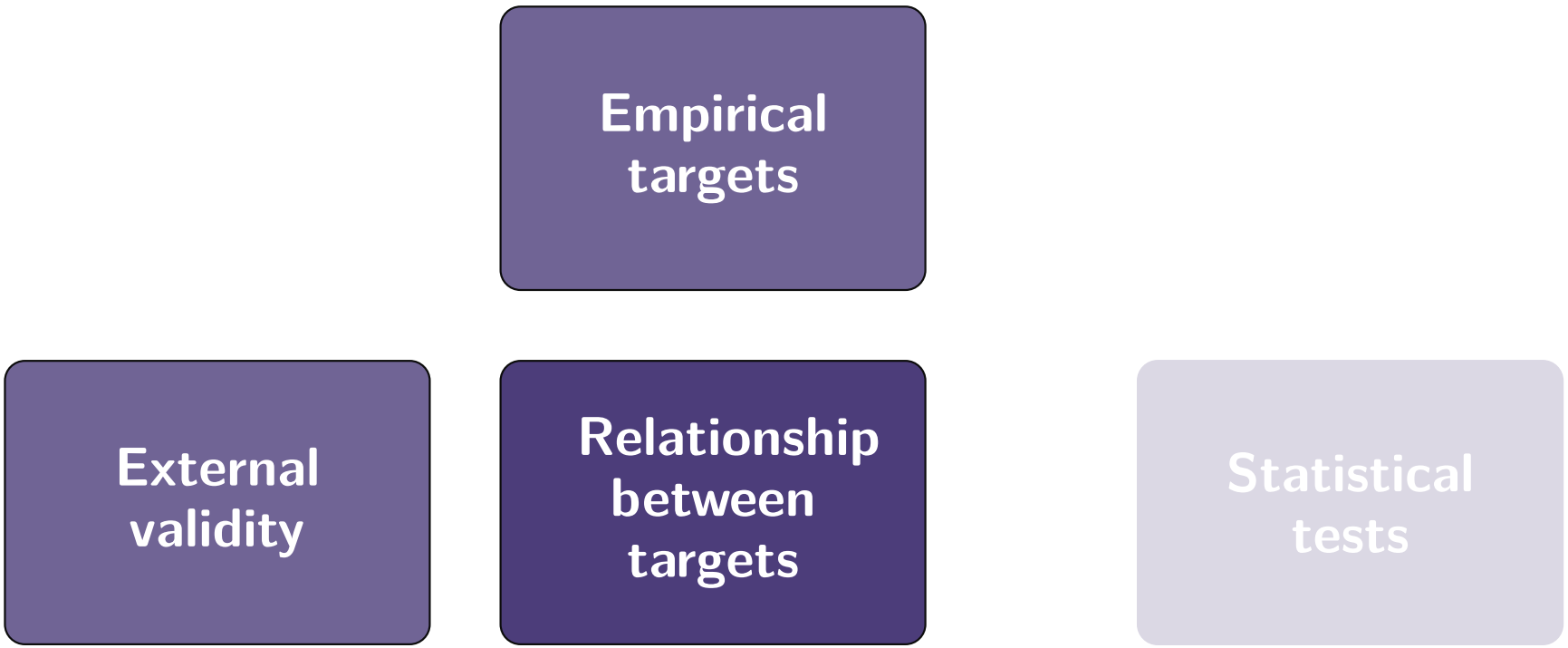


When **sign-congruent external validity** does not hold, the set of research designs where a harmonized design will produce effects with different signs in different settings has positive measure.

Relationship between targets

How do the empirical targets across studies relate to each other?

- Concepts of **target equivalence** and **target congruence**
- **Discrepancies** between targets



Target equivalence and congruence

Consider two studies: $\mathcal{E}_1 = \{m_1, (\omega_1{}', \omega_1{}''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2{}', \omega_2{}''), \theta_2\}$:

Target equivalence and congruence

Consider two studies: $\mathcal{E}_1 = \{m_1, (\omega_1{}', \omega_1{}''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2{}', \omega_2{}''), \theta_2\}$:

Definition

*$\mathcal{E}_1$ and $\mathcal{E}_2$ are **target-equivalent** if*

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

Target equivalence and congruence

Consider two studies: $\mathcal{E}_1 = \{m_1, (\omega_1{}', \omega_1{}''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2{}', \omega_2{}''), \theta_2\}$:

Definition

*$\mathcal{E}_1$ and $\mathcal{E}_2$ are **target-equivalent** if*

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

Definition

*$\mathcal{E}_1$ and $\mathcal{E}_2$ are are **target-congruent** if*

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

Target discrepancies

The **target discrepancy** from setting θ to θ' is

$$\Delta_{m,(\omega',\omega'')}(\theta,\theta') = \tau_m(\omega',\omega'' \mid \theta) - \tau_m(\omega',\omega'' \mid \theta').$$

Target discrepancies

The **target discrepancy** from setting θ to θ' is

$$\Delta_{m,(\omega',\omega'')}(\theta,\theta') = \tau_m(\omega',\omega'' \mid \theta) - \tau_m(\omega',\omega'' \mid \theta').$$

Non-random differences in effects across settings, holding constant the design.

Target discrepancies

The **target discrepancy** from setting θ to θ' is

$$\Delta_{m,(\omega',\omega'')}(\theta,\theta') = \tau_m(\omega',\omega'' \mid \theta) - \tau_m(\omega',\omega'' \mid \theta').$$

Non-random differences in effects across settings, holding constant the design.

Measure of departures from **external validity**.

- Target-equivalence implies **zero** target discrepancies.
- Target-congruence is when they take a **particular form**.

Artifactual discrepancies

For a fixed setting θ, the **artifactual discrepancy** is

$$\mathcal{A}_{ij}(\theta) = \tau_{m_i}(\omega_i{}', \omega_i{}'' \mid \theta) - \tau_{m_j}(\omega_j{}', \omega_j{}'' \mid \theta).$$

Artifactual discrepancies

For a fixed setting θ, the **artifactual discrepancy** is

$$\mathcal{A}_{ij}(\theta) = \tau_{m_i}(\omega_i{}', \omega_i{}'' \mid \theta) - \tau_{m_j}(\omega_j{}', \omega_j{}'' \mid \theta).$$

Non-random differences in treatment effects produced by:

- Different measurement strategies
- Different contrasts

Artifactual discrepancies

For a fixed setting θ, the **artifactual discrepancy** is

$$\mathcal{A}_{ij}(\theta) = \tau_{m_i}(\omega_i{}', \omega_i{}'' \mid \theta) - \tau_{m_j}(\omega_j{}', \omega_j{}'' \mid \theta).$$

Non-random differences in treatment effects produced by:

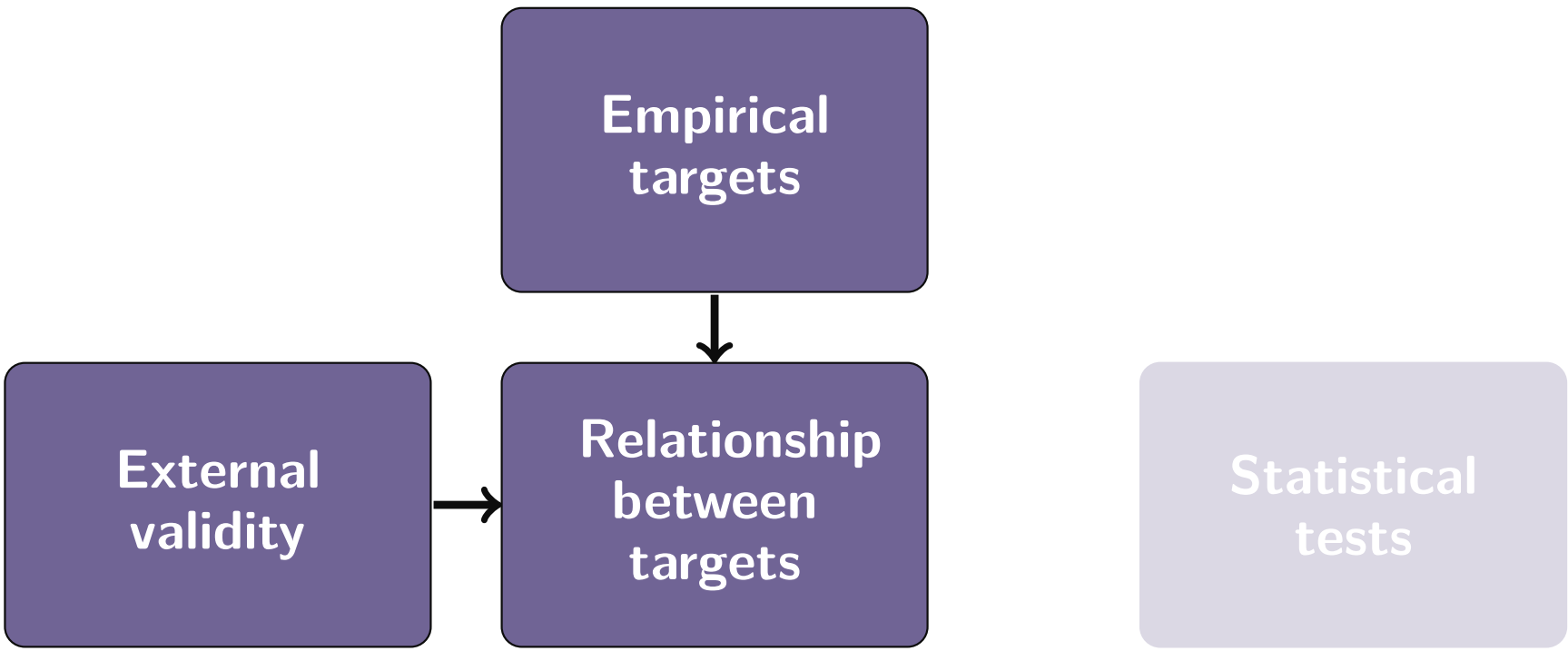
- Different measurement strategies
- Different contrasts

Remark: $\mathcal{A}_{ij}(\theta) = 0$ for almost every θ if and only if i and j are **harmonized**.

Results

Our goal

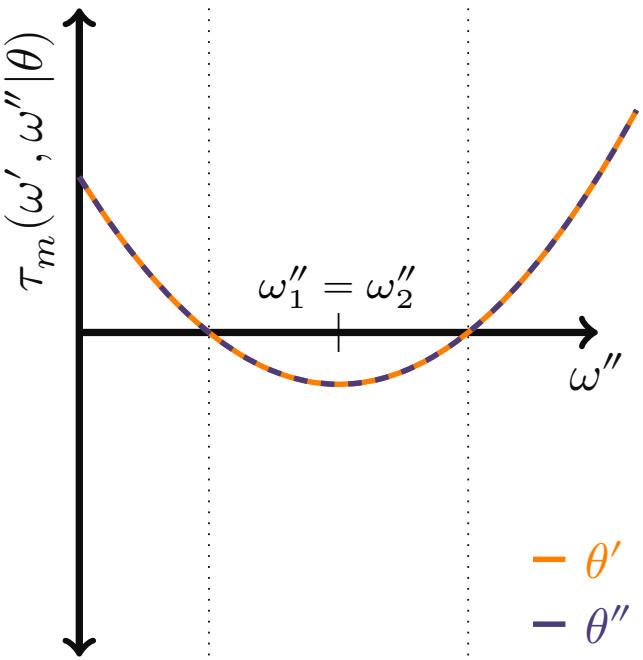
Under what conditions do we achieve **target equivalence** or **target congruence**?



Achieving target equivalence

Theorem

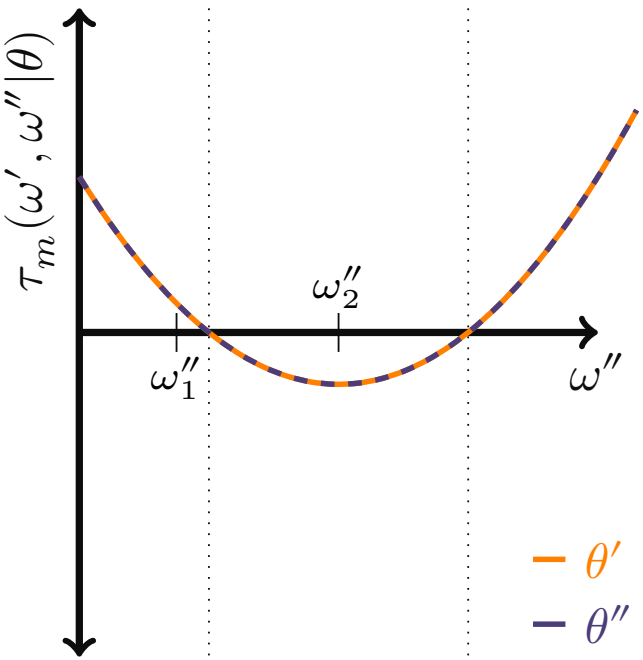
*Target-equivalence holds across a collection of studies if and only if the mechanism satisfies **external validity** and all studies are **harmonized** (almost everywhere).*



Achieving target equivalence

Theorem

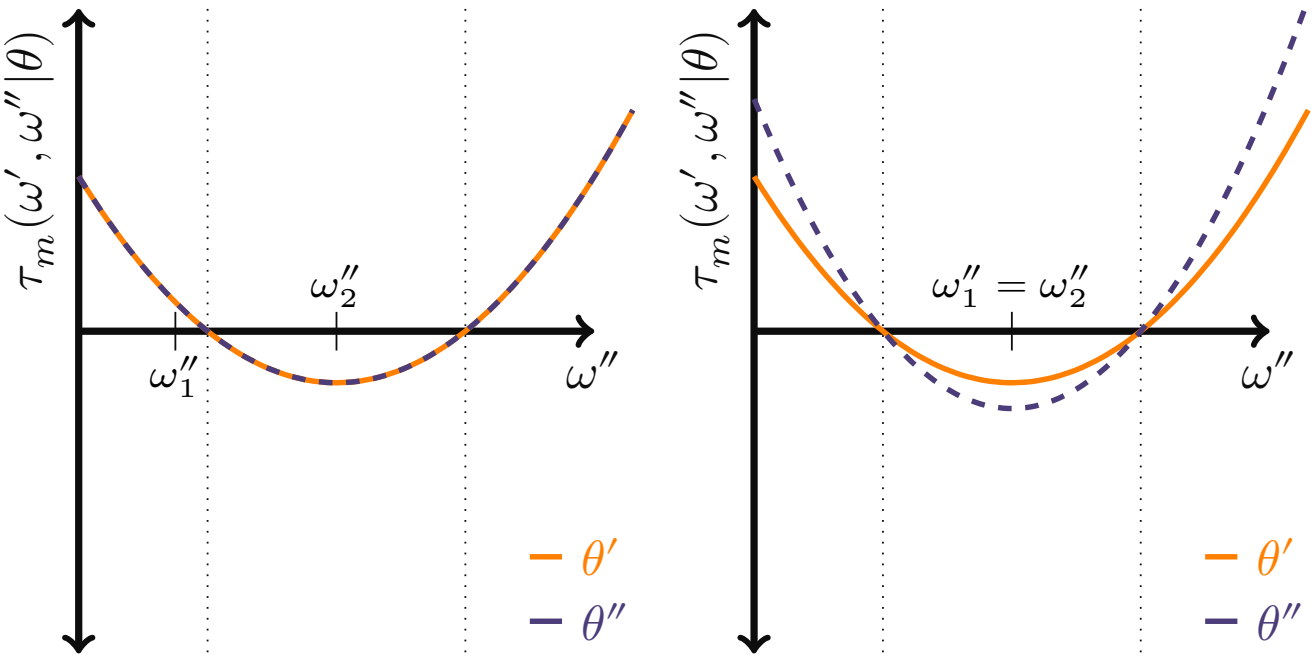
*Target-equivalence holds across a collection of studies if and only if the mechanism satisfies **external validity** and all studies are **harmonized** (almost everywhere).*



Achieving target equivalence

Theorem

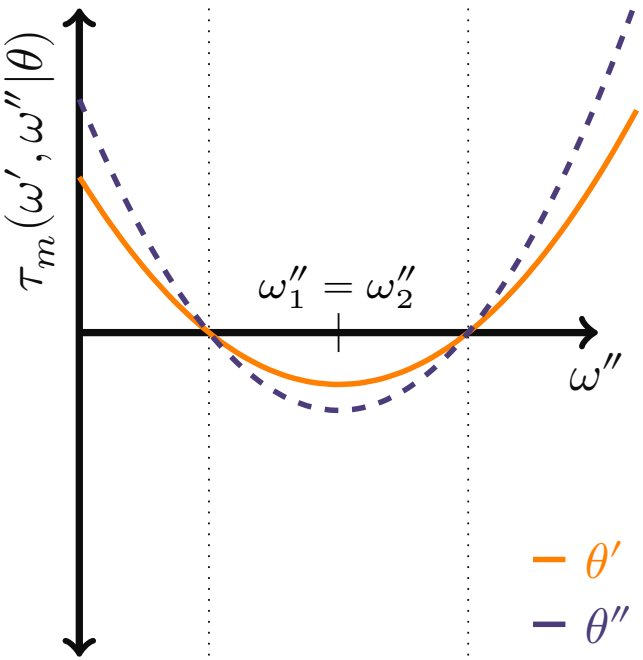
Target-equivalence holds across a collection of studies if and only if the mechanism satisfies ***external validity*** *and all studies are* ***harmonized*** *(almost everywhere).*



Achieving target congruence

Theorem

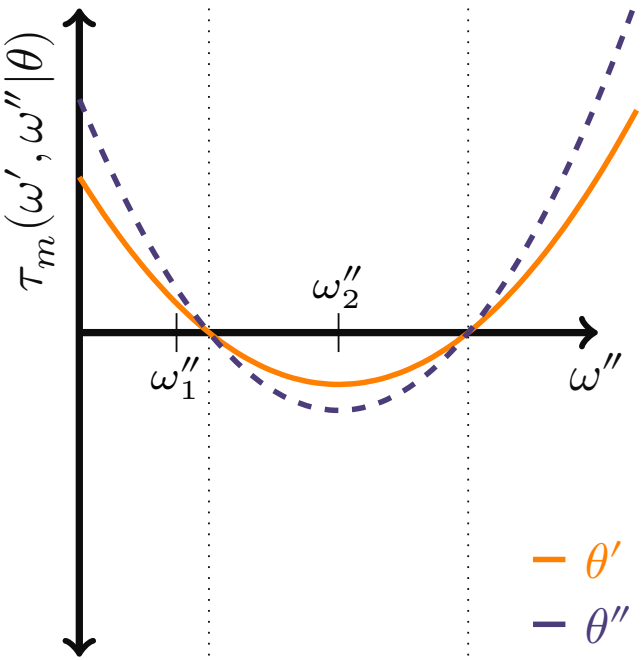
*Target-congruence holds across a collection of studies if and only if the mechanism satisfies **sign-congruent external validity** and all studies are **harmonized** (almost everywhere).*



Achieving target congruence

Theorem

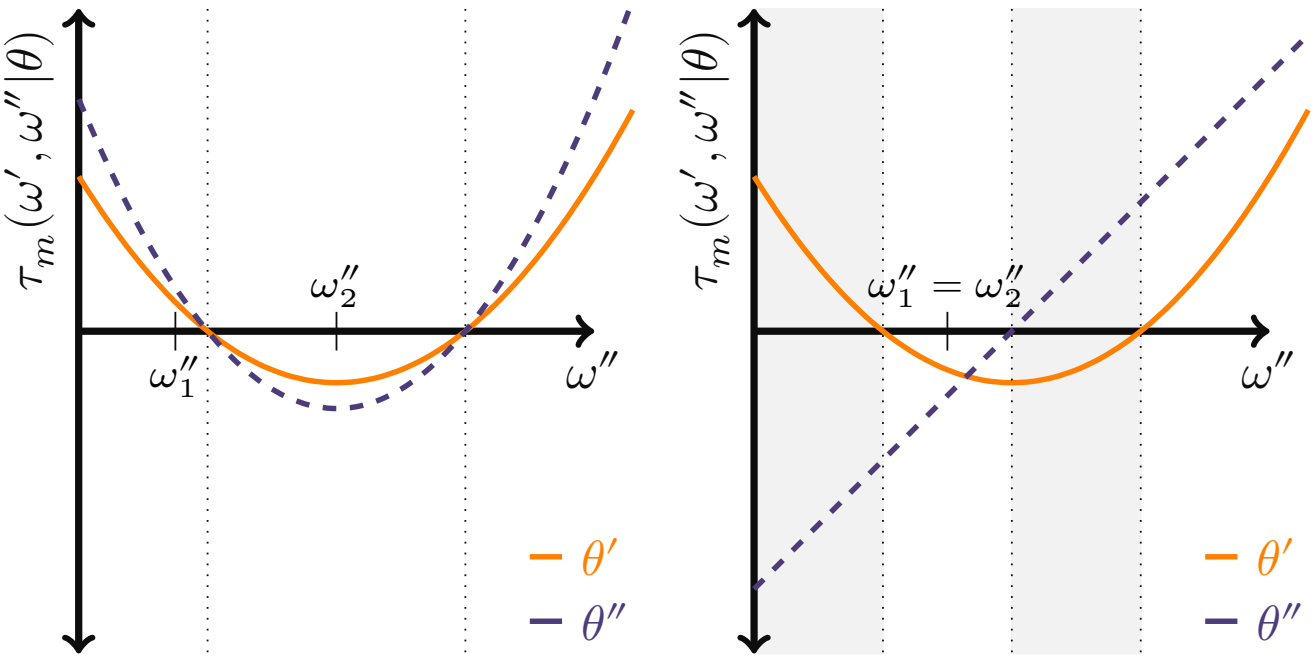
*Target-congruence holds across a collection of studies if and only if the mechanism satisfies **sign-congruent external validity** and all studies are **harmonized** (almost everywhere).*



Achieving target congruence

Theorem

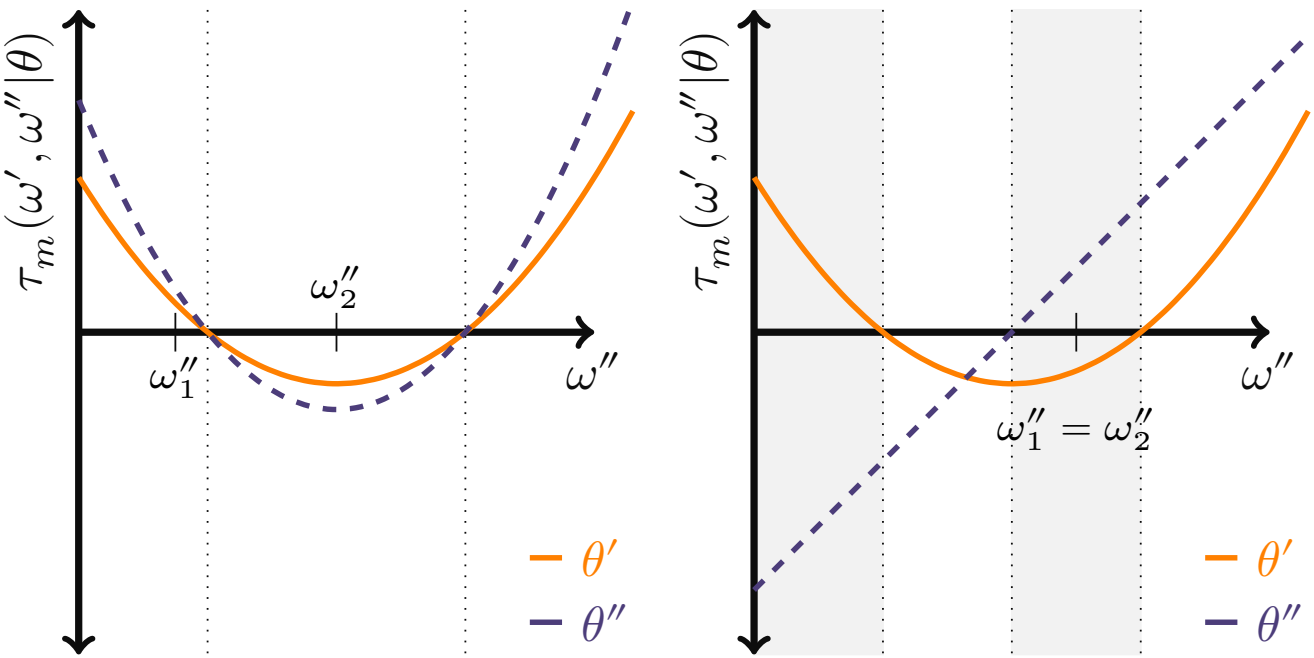
*Target-congruence holds across a collection of studies if and only if the mechanism satisfies **sign-congruent external validity** and all studies are **harmonized** (almost everywhere).*



Achieving target congruence

Theorem

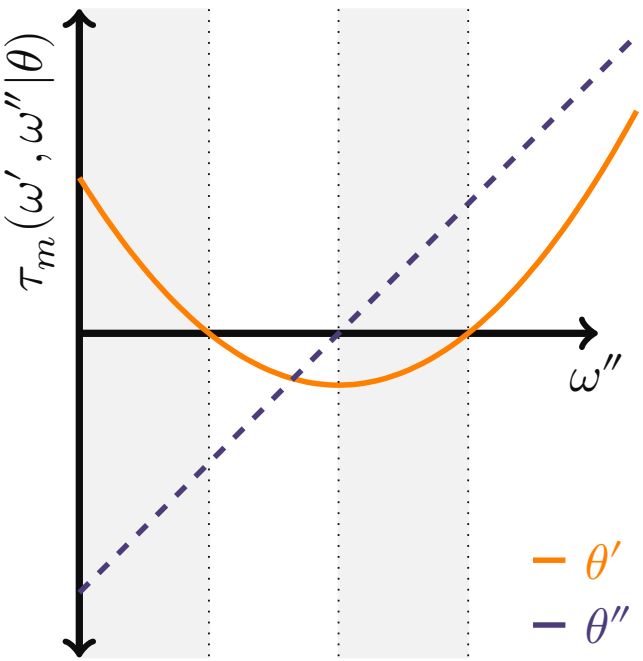
*Target-congruence holds across a collection of studies if and only if the mechanism satisfies **sign-congruent external validity** and all studies are **harmonized** (almost everywhere).*



Relationship to the number of studies

Theorem

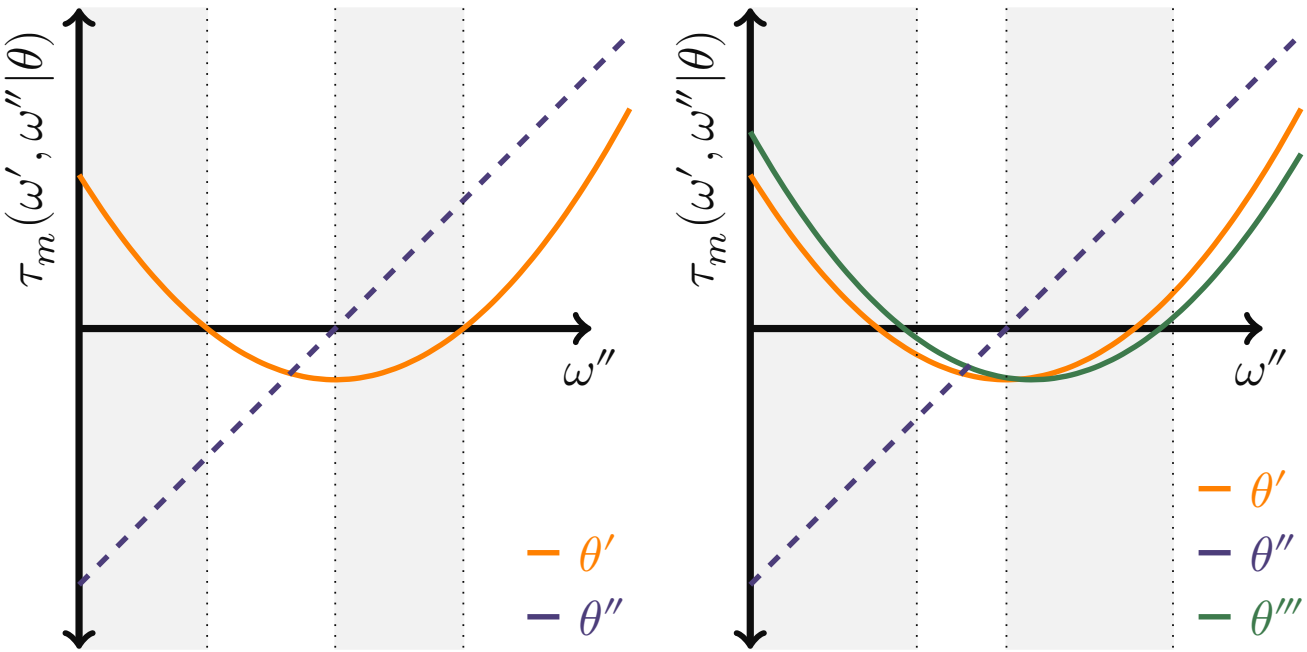
The set where the sign of empirical targets is different is nondecreasing (in the set inclusion order) in the ***number of studies*** N.



Relationship to the number of studies

Theorem

The set where the sign of empirical targets is different is nondecreasing (in the set inclusion order) in the ***number of studies*** N.



Why it matters

Recall the P2P studies from Uganda:

- 2004-5 study (θ') $\rightarrow$ larger reduction in under-5 mortality
- 2014-15 study (θ'') $\rightarrow$ smaller reduction in under-5 mortality

Why it matters

Recall the P2P studies from Uganda:

- 2004-5 study (θ') $\rightarrow$ larger reduction in under-5 mortality
- 2014-15 study (θ'') $\rightarrow$ smaller reduction in under-5 mortality

Thought experiment: suppose these *were* the treatment effects.

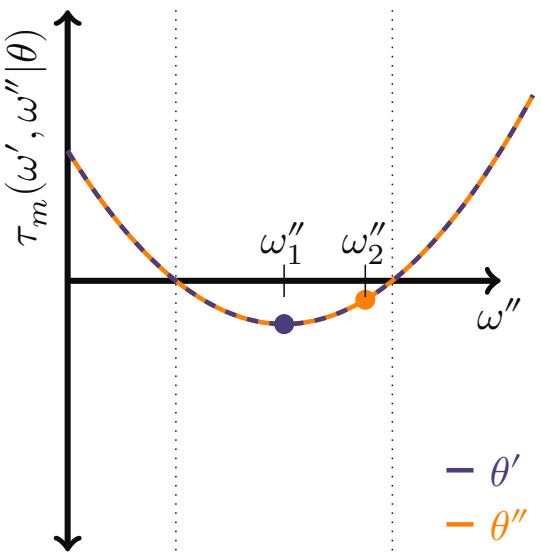
Why it matters

Recall the P2P studies from Uganda:

- 2004-5 study (θ') $\rightarrow$ larger reduction in under-5 mortality
- 2014-15 study (θ'') $\rightarrow$ smaller reduction in under-5 mortality

Thought experiment: suppose these *were* the treatment effects.

A. External validity

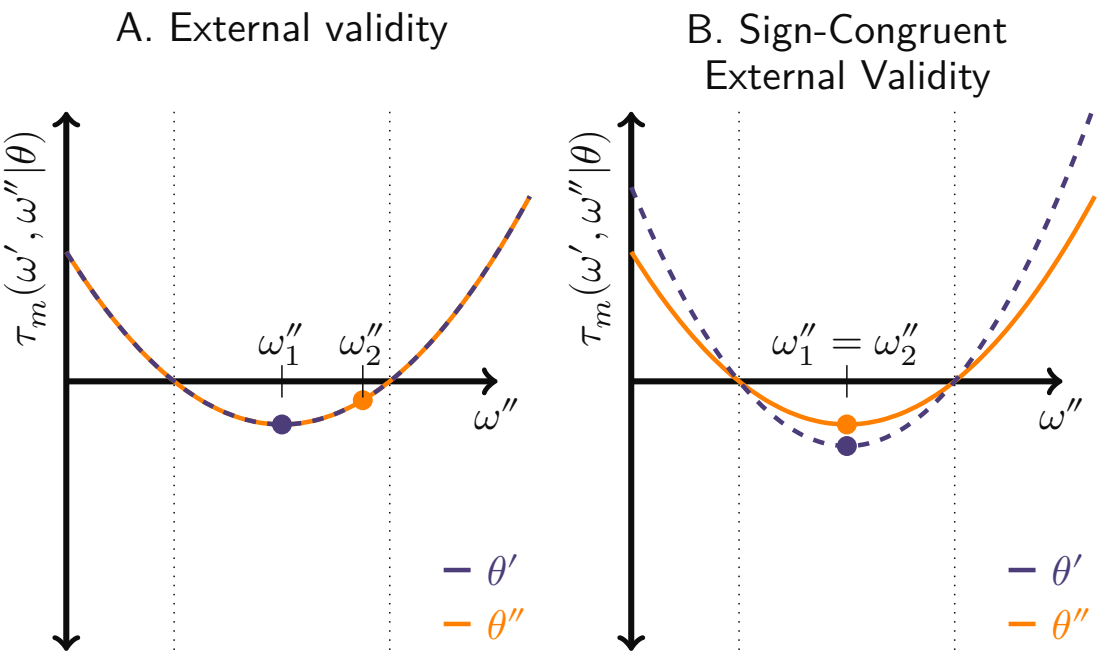


Why it matters

Recall the P2P studies from Uganda:

- 2004-5 study (θ') $\rightarrow$ larger reduction in under-5 mortality
- 2014-15 study (θ'') $\rightarrow$ smaller reduction in under-5 mortality

Thought experiment: suppose these *were* the treatment effects.

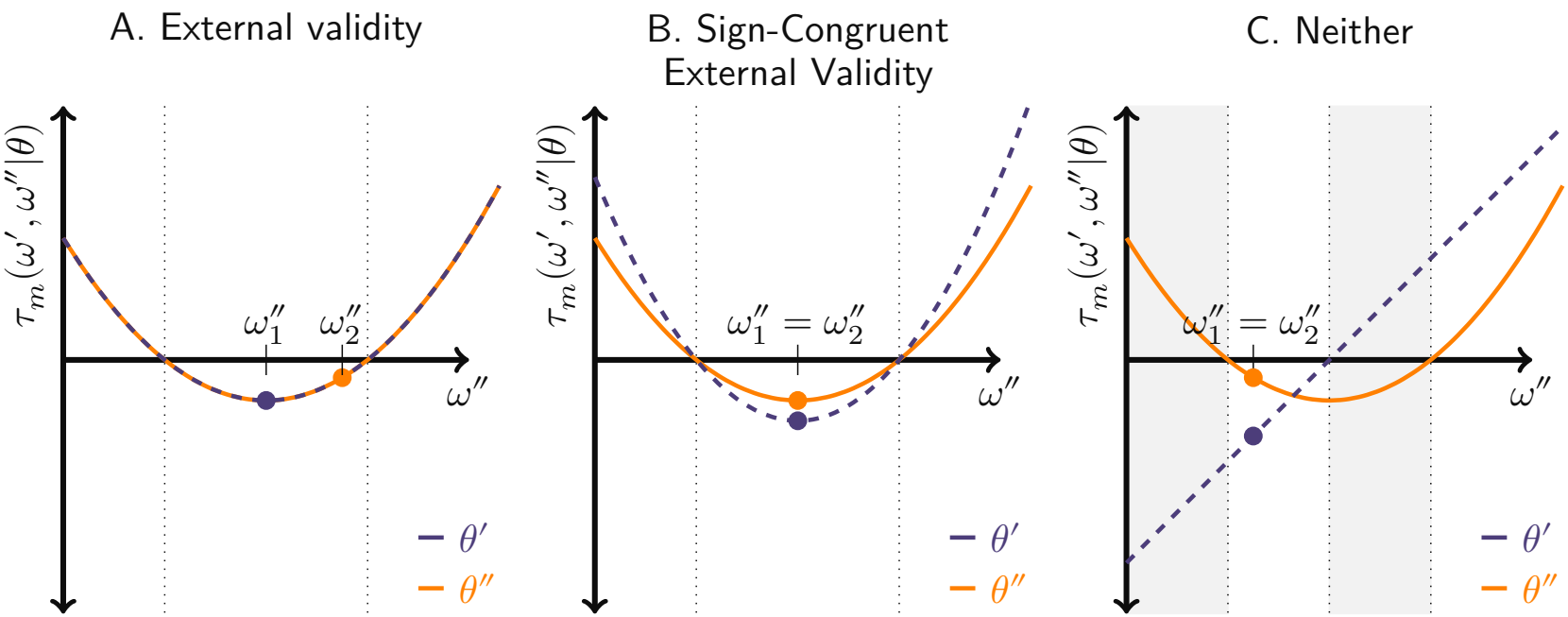


Why it matters

Recall the P2P studies from Uganda:

- 2004-5 study (θ') $\rightarrow$ larger reduction in under-5 mortality
- 2014-15 study (θ'') $\rightarrow$ smaller reduction in under-5 mortality

Thought experiment: suppose these *were* the treatment effects.



Taking Stock

When research designs are **harmonized**:

- External validity $\rightarrow$ **target equivalence**
- Sign-congruent external validity $\rightarrow$ **target congruence**

Taking Stock

When research designs are **harmonized**:

- External validity $\rightarrow$ **target equivalence**
- Sign-congruent external validity $\rightarrow$ **target congruence**

Without harmonization, relationship between empirical targets of studies is ambiguous, even when external validity holds.

Taking Stock

When research designs are **harmonized**:

- External validity $\rightarrow$ **target equivalence**
- Sign-congruent external validity $\rightarrow$ **target congruence**

Without harmonization, relationship between empirical targets of studies is ambiguous, even when external validity holds.

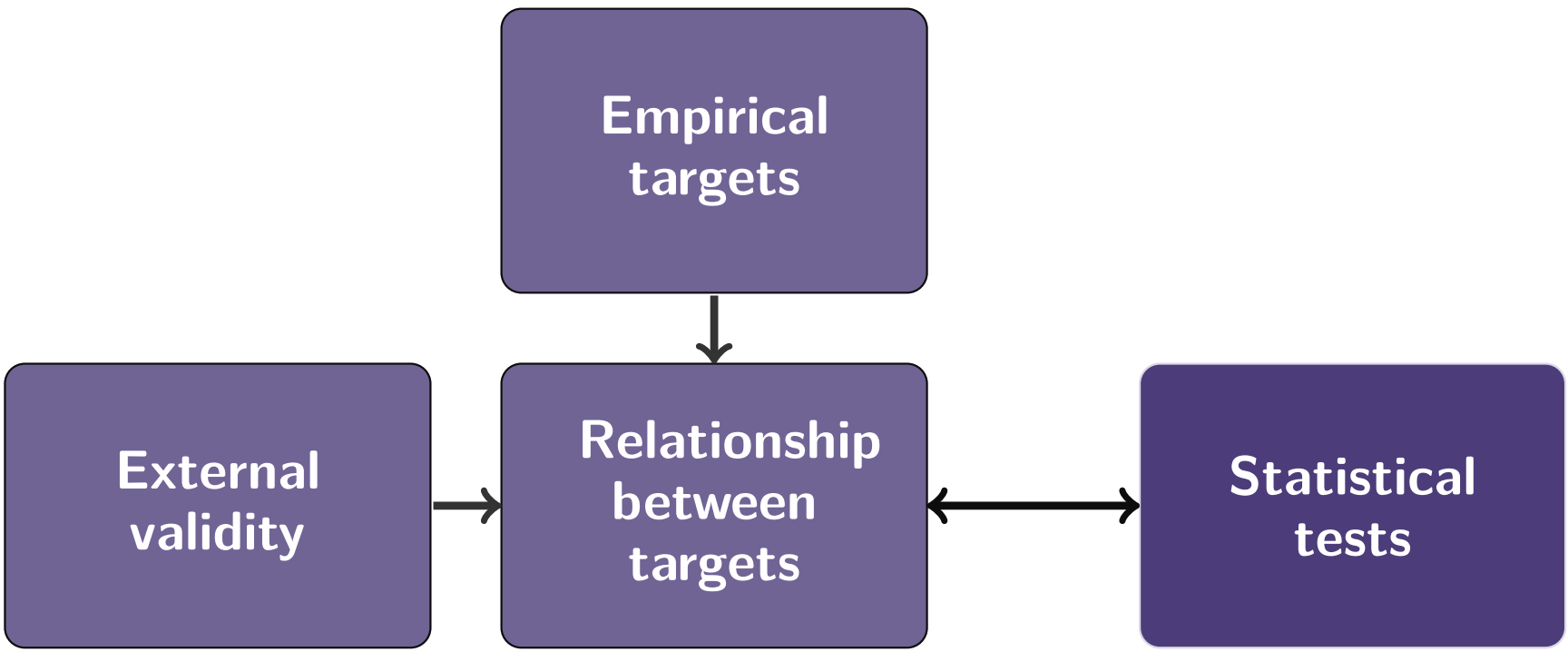
Doing more replications can only exacerbate these problems.

- These problems are non-statistical.

Comparing Estimates

Making Comparisons

- Two comparisons pursued in replications are:
  - Comparison of **point estimates** $\rightarrow$ target equivalence
  - Comparison of **estimate signs** $\rightarrow$ target congruence



Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies
- Target discrepancies

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies
- Target discrepancies

Examining the difference in treatment effects:

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies
- Target discrepancies

Examining the difference in treatment effects:

$$e_1 - e_2 = \tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) + \varepsilon_1^{n_1} - \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2) - \varepsilon_2^{n_2}$$

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies
- Target discrepancies

Examining the difference in treatment effects:

$$e_1 - e_2 = \tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) + \varepsilon_1^{n_1} - \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2) - \varepsilon_2^{n_2}$$

$$= \overbrace{\varepsilon_1^{n_1} - \varepsilon_2^{n_2}}^{\text{statistical discrepancy}} + \ldots$$

Three discrepancies

Estimated treatment effects in two studies will always differ:

- Statistical discrepancies
- Artifactual discrepancies
- Target discrepancies

Examining the difference in treatment effects:

$$e_1 - e_2 = \tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) + \varepsilon_1^{n_1} - \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2) - \varepsilon_2^{n_2}$$

$$= \overbrace{\varepsilon_1^{n_1} - \varepsilon_2^{n_2}}^{\text{statistical discrepancy}} + \underbrace{\Delta_{m_1,(\omega_1^{'},\omega_1^{''})}(\theta_1, \theta_2)}_{\text{target discrepancy}} - \overbrace{\mathcal{A}_{12}(\theta_2)}^{\text{artifactual discrepancy}}.$$

Two comparisons

1. The estimate-comparison test computes:

$$\mathcal{W} = e_1 - e_2,$$

and test the null hypothesis of **target equivalence**.

Two comparisons

1. The estimate-comparison test computes:

$$\mathcal{W} = e_1 - e_2,$$

and test the null hypothesis of **target equivalence**.

Two comparisons

1. The estimate-comparison test computes:

$$\mathcal{W} = e_1 - e_2,$$

and test the null hypothesis of **target equivalence**.

2. The sign-comparison test computes:

$$\mathcal{Z} = e_1 \cdot e_2,$$

and test the null hypothesis of **target congruence**, which occurs when $sign(\tau_{m_1}(\omega_1', \omega_1''|\theta_1)) \cdot sign(\tau_{m_2}(\omega_2', \omega_2''|\theta_2)) > 0$.

What does the estimate-comparison test evaluate?

Proposition

If two studies have unbiased and consistent estimation errors, then

1. *If the studies are harmonized, then the estimate-comparison test assesses a null hypothesis that the mechanism is* ***externally valid****;*
2. *If the mechanism has external validity, then the estimate-comparison test assesses a null hypothesis that the studies are* ***harmonized****.*

What does the estimate-comparison test evaluate?

Proposition

If two studies have unbiased and consistent estimation errors, then

1. *If the studies are harmonized, then the estimate-comparison test assesses a null hypothesis that the mechanism is* ***externally valid****;*
2. *If the mechanism has external validity, then the estimate-comparison test assesses a null hypothesis that the studies are* ***harmonized****.*

Key idea: learning about **external validity** is not automatic!

- We have to worry about cross-study design as well.
- Sometimes we prefer to learn about how effects vary in study design.

What does the sign-comparison test evaluate?

Proposition

If two studies are harmonized and have unbiased and consistent estimation errors, then the sign-comparison test assesses a null hypothesis of ***sign-congruent external validity****.*

What does the sign-comparison test evaluate?

Proposition

If two studies are harmonized and have unbiased and consistent estimation errors, then the sign-comparison test assesses a null hypothesis of ***sign-congruent external validity****.*

Key idea: learning about **sign-congruent external validity** is not automatic!

- We have to worry about cross-study design as well.
- A weaker concept of external validity limits what we could learn about artifactual discrepancies.

Two Approaches to Replication

Structural approach

Suppose you want to learn about external validity but cannot harmonize studies.

Structural approach

Suppose you want to learn about external validity but cannot harmonize studies.

Posit a **structural model** of cross-study environment:

- Specify how artifactual discrepancies vary in the design.
- (Conversely, specify how target discrepancies emerge.)
- Address discrepancies by assumption.

Structural approach

Suppose you want to learn about external validity but cannot harmonize studies.

Posit a **structural model** of cross-study environment:

- Specify how artifactual discrepancies vary in the design.
- (Conversely, specify how target discrepancies emerge.)
- Address discrepancies by assumption.

Strength: facilitates strong conclusions from data,

Structural approach

Suppose you want to learn about external validity but cannot harmonize studies.

Posit a **structural model** of cross-study environment:

- Specify how artifactual discrepancies vary in the design.
- (Conversely, specify how target discrepancies emerge.)
- Address discrepancies by assumption.

Strength: facilitates strong conclusions from data,

Drawback: inconsistent with notions of causality invoked within-study.

Design-based alternative

How can we maintain a causal interpretation in meta-studies?

Design-based alternative

How can we maintain a causal interpretation in meta-studies?

Focus on the importance of research design

- Connected with credibility approaches to internal validity.

Design-based alternative

How can we maintain a causal interpretation in meta-studies?

Focus on the importance of research design

- Connected with credibility approaches to internal validity.

Design-based approach to conceptual replication is **sequential**

1. Design-**harmonized** replications
   → measure target discrepancies (under one design).
2. **Single-setting** replicatons varying design
   → measure artifactual discrepancies (in one setting).
3. **Non-harmonized multi-setting** design
   → With steps #1 and #2, evaluate whether artifactual discrepancies vary in settings.

Limits to design-based approach to conceptual replication

Sequential nature requires that effects of mechanisms are stable over time.

- More likely for some interventions, settings than others.

Problems in the organization of research:

- Limited researcher incentives for replication.
- In principle, we favor replication by independent teams.
  - Requires more transparent communication of precise design, link to constructs than is current practice.

Conclusion

Conclusion

It is essential to learn about **external validity** for:

- Use of evidence in policymaking
- Our understanding of the generality of phenomena.

Conclusion

It is essential to learn about **external validity** for:

- Use of evidence in policymaking
- Our understanding of the generality of phenomena.

Replication permits learning about how the effects of mechanisms manifest across settings:

- Strength: does not assume mechanism across contexts.
- ... but not every comparison is informative.
- Cross-study design affects what we learn from comparison.

Conclusion

It is essential to learn about **external validity** for:

- Use of evidence in policymaking
- Our understanding of the generality of phenomena.

Replication permits learning about how the effects of mechanisms manifest across settings:

- Strength: does not assume mechanism across contexts.
- ... but not every comparison is informative.
- Cross-study design affects what we learn from comparison.

Formal **conceptual frameworks** as a necessary complement to advances in estimation.

Backstory

Backstory

The methods questions that I answer come from my applied work.

Backstory

The methods questions that I answer come from my applied work.

From 2017-21, I led a Metaketa on community monitoring of common pool resources:

- 6 coordinated experiments
- 18 researchers

Backstory

The methods questions that I answer come from my applied work.

From 2017-21, I led a Metaketa on community monitoring of common pool resources:

- 6 coordinated experiments
- 18 researchers

Multi-site experiment a response to the "on-going crisis of **external validity**," with goals of:

- Informing policy
- Contributing general causal knowledge

Backstory

The methods questions that I answer come from my applied work.

From 2017-21, I led a Metaketa on community monitoring of common pool resources:

- 6 coordinated experiments
- 18 researchers

Multi-site experiment a response to the "on-going crisis of **external validity**," with goals of:

- Informing policy
- Contributing general causal knowledge

Nagging question: What could we have learned?

Applied $\rightarrow$ methodological work: three directions

Metaketas $\rightarrow$ external validity and evidence accumulation:

- "External Validity and Meta-Analysis" (forthcoming *AJPS*)
- *External Validity and Evidence Accumulation* (under contract, Cambridge)

Applied $\rightarrow$ methodological work: three directions

Metaketas $\rightarrow$ external validity and evidence accumulation:

- "External Validity and Meta-Analysis" (forthcoming *AJPS*)
- *External Validity and Evidence Accumulation* (under contract, Cambridge)

Theory/empirics papers $\rightarrow$ theory of the reduced form:

- "Phantom Counterfactuals" (forthcoming in *AJPS*)

Applied $\rightarrow$ methodological work: three directions

Metaketas $\rightarrow$ external validity and evidence accumulation:

- "External Validity and Meta-Analysis" (forthcoming *AJPS*)
- *External Validity and Evidence Accumulation* (under contract, Cambridge)

Theory/empirics papers $\rightarrow$ theory of the reduced form:

- "Phantom Counterfactuals" (forthcoming in *AJPS*)

Designing field experiments $\rightarrow$ formal treatments of ethics:

- "The Ethics of Electoral Experiments: Design-Based Recommendations" (working paper)

Thank you!

www.taraslough.com

Supplementary Information

Contents

Treatment effects and potential outcomes framework

Define potential outcomes:

$$Y_i^m(\omega''|\theta) \text{ and } Y_i^m(\omega'|\theta)$$

The treatment effect function is given by:

$$\tau_m(\omega', \omega''|\theta) = f_{\mathcal{D}}(Y_i^m(\omega''|\theta) - Y_i^m(\omega'|\theta)),$$

where:

- $f(\cdot)$ is a function or operator.
- $\mathcal{D}$ is a set of units for whom treatment effects are estimated

C-validity

Egami and Hartman (2022) formalize C-validity as:

$$Y_i(T = 1, c) - Y_i(T = 0, c) = Y_i(T = 1, c^*) - Y_i(T = 0, c^*)$$

with our (potential outcome) notation, this can be written:

$$Y_i^m(\omega''|\theta) - Y_i^m(\omega'|\theta) = Y_i^m(\omega''|\theta') - Y_i^m(\omega'|\theta')$$

Recall that:

$$\tau_m(\omega', \omega''|\theta) = f_{\mathcal{D}}(Y_i^m(\omega''|\theta) - Y_i^m(\omega'|\theta)).$$

As such, C-validity implies that:

$$\tau_m(\omega', \omega''|\theta) = \tau_m(\omega', \omega''|\theta'),$$

which is the definition of **external validity**. But external validity does not imply that C-validity holds.

T-validity and Y-validity

T-validity holds that:

$$\mathbb{E}_{\mathcal{P}}[Y_i(T_i = 1, c) - Y_i(T_i = 0, c)] = \mathbb{E}_{\mathcal{P}}[Y_i(T_i^* = 1, c) - Y_i(T_i^* = 0, c)],$$

- Ruled out by symmetry assumption (i.e., $T_i^* = 1 - T_i$), except for case when $\mathbb{E}_{\mathcal{P}}[Y_i(T_i = 1, c) - Y_i(T_i = 0, c)] = 0 \forall T_i$, which is ruled out by full-rank assumption.

Y-validity holds that:

$$\mathbb{E}_{\mathcal{P}}[Y_i(T_i = 1, c) - Y_i(T_i = 0, c)] = \mathbb{E}_{\mathcal{P}}[Y_i^*(T_i = 1, c) - Y_i^*(T_i = 0, c)],$$

- Ruled out by full-rank assumption.

Fisher: Concepts in (applied) statistics

> *"...the obscurity which envelops the theoretical bases of statistical methods may perhaps be ascribed to two considerations. In the first place, it appears to be widely thought, or rather felt, that in a subject in which all results are liable to greater or smaller errors, precise definitions of ideas or concepts is, if not impossible, at least not a practical necessity. In the second place...it is customary to apply the same name...to both the true value we would like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation; so in applying the term probable error, writers sometimes would appear to suggest that the former quantity, and not merely the latter, is subject to error."*
>
> R.A. Fisher (1922, p. 311)

Sign-comparison test: inference

Let $\varepsilon_i^{n_i}$ be normally distributed with mean 0 and let the standard error of e_i be se_i.

- The p-value of the null hypothesis of sign-congruence is:

$$\begin{aligned} p &= \Pr(e_1 > 0)\Pr(e_2 > 0) + \Pr(e_1 < 0)\Pr(e_2 < 0) \\ &= \Phi(\frac{e_1}{se_1})\Phi(\frac{e_2}{se_2}) + (1 - \Phi(\frac{e_1}{se_1}))(1 - \Phi(\frac{e_2}{se_2})), \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

Rejection regions

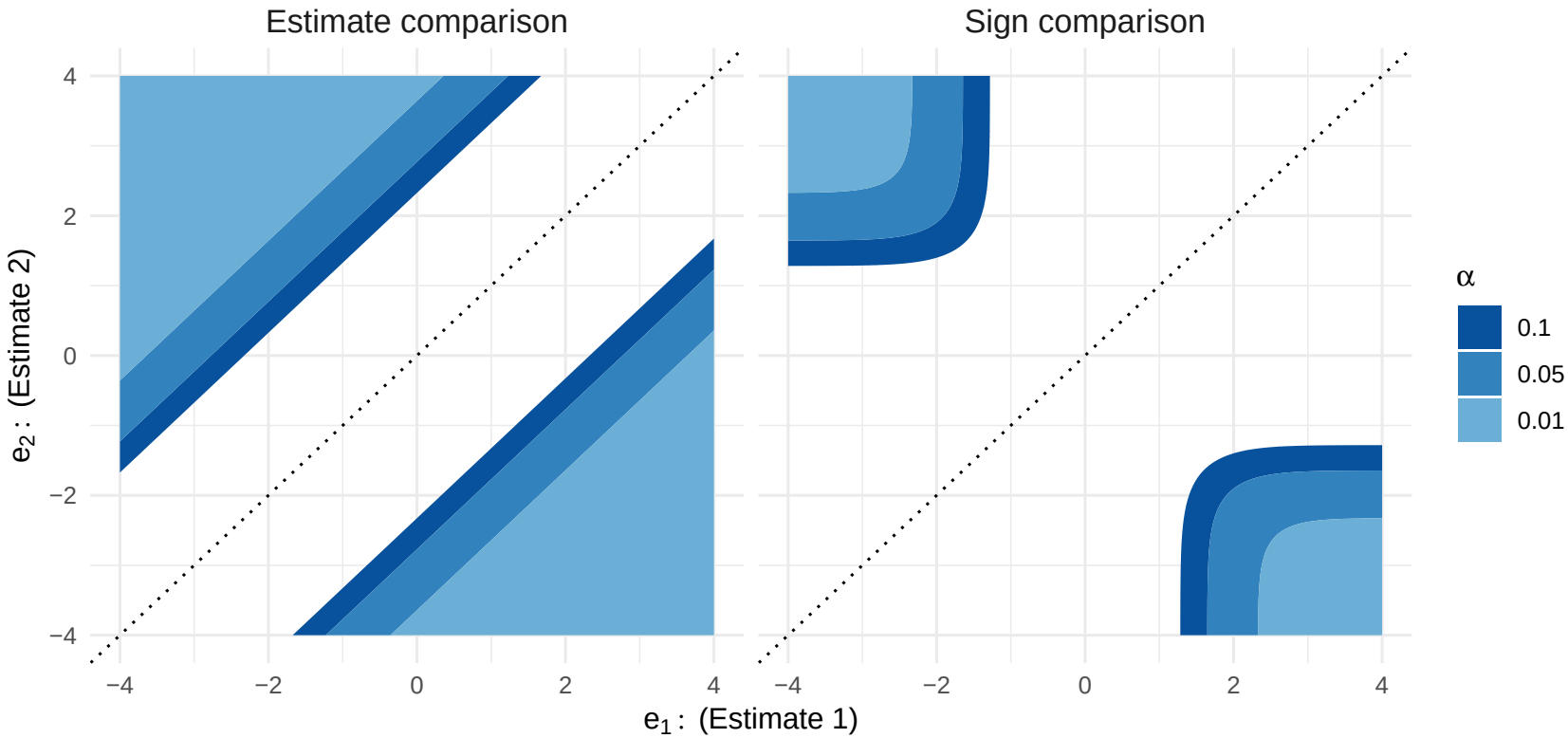


Figure: Rejection regions of the estimate- and sign-comparison approaches for Type-I error rates, $\alpha \in \{0.01, 0.05, 0.1\}$. Both plots fix $se_1 = se_2 = 1$ in order to visualize these regions in two dimensions.

Classification of replication designs

Class	Sub-class	Studies differ in... Samples	Settings	Design
Exact		–	–	–
Direct		✓	–	–
Conceptual	Harmonized	✓	✓	–
Conceptual	Single-setting	✓	–	✓
Conceptual	Non-harmonized, multi-setting	✓	✓	✓

Table: Mapping between conventional classification of replication studies and our framework. Note that the disaggregation of conceptual replications into sub-classes is non-standard in existing literature.

Additional estimands, tests in OSF (2015)

Suppose we have $N > 1$ replication studies, where each study consists of a pair of estimates.

OSF (2015) additionally compute:

- Share of replications with $p < 0.05$ in the same direction
- Effect size difference
- Meta-analytic estimate
- Original effect within replication 95% CI
- Subjective "yes" did it replicate