Sign-Congruence, External Validity, and Replication

Tara Slough & Scott A. Tyson

Replication and Accumulation

Most prominent critique of the credibility revolution: lack of external validity

Typical advice: *run more studies*

Then:

Compare results from multiple studies → **replication**
Combine results from multiple studies → **meta-analysis**

Replication and Accumulation

Most prominent critique of the credibility revolution: lack of external validity

Typical advice: *run more studies*

Then:

- Compare multiple studies → **replication** (Here!)
- Combine from multiple studies → **meta-analysis** (Slough & Tyson 2022)

Accumulation of Evidence

When we observe differences across experiments, kinds of explanations:

1. **Statistical noise**, e.g., sampling variability
2. Differences in **study design**, e.g., different outcome measures
3. Phenomenon is not **generalizable**

Accumulation of Evidence

When we observe differences across experiments, kinds of explanations:

1. **Statistical noise**, e.g., sampling variability
2. Differences in **study design**, e.g., different outcome measures
3. Phenomenon is not **generalizable**

Accumulating evidence requires addressing all three

What we do

Present a general framework to help think about evidence accumulation

Develop key concepts relevant for understanding replication

Formally link some replication approaches with external validity concepts

Advocate a *design-based approach to conceptual replication*

Outline

What's a study?

A **study** is a triple:

1. A **setting**, θ

   Contextual features, population, time, etc.

2. A **measurement strategy**, m

   Outcome choice and measurement components

3. A **contrast**, (ω', ω'')

   Comparison of interest (e.g., treatment/control)

What's a study?

A **study** is a triple:

1. A **setting**, θ

   Contextual features, population, time, etc.

2. A **measurement strategy**, m

   Outcome choice and measurement components

3. A **contrast**, (ω', ω'')

   Comparison of interest (e.g., treatment/control)

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

What's the Empirical Target?

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

- Smooth, derivative has full rank in measurement strategies and contrasts

What's the Empirical Target?

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

- Smooth, derivative has full rank in measurement strategies and contrasts

Measured effect:

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

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

Outline

Two key concepts for Accumulating Evidence

1. Two studies $\mathcal{E}_1 = \{m_1, (\omega_1', \omega_1''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2', \omega_2''), \theta_2\}$ are **target-equivalent** if

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

2. Two studies $\mathcal{E}_1 = \{m_1, (\omega_1', \omega_1''), \theta_1\}$ and $\mathcal{E}_2 = \{m_2, (\omega_2', \omega_2''), \theta_2\}$ 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 are (nonrandom) differences in the treatment effects (targets) that come from differences under the same design in different settings.

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 are (nonrandom) differences in the treatment effects (targets) that come from differences under the same design in different settings.

Target discrepancies are about departures from external validity

- Target-equivalence is when they're absent
- Target-congruence is when they take a particular form

Two notions of external validity

Two notions of external validity

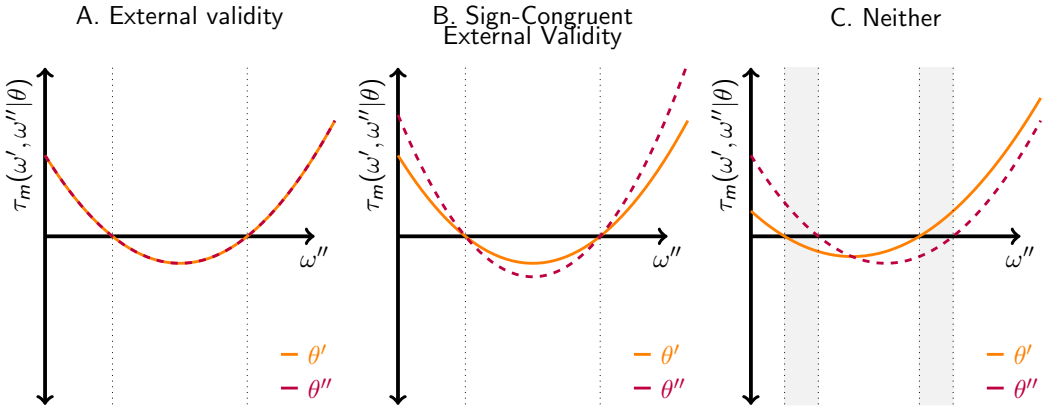
1. A mechanism has **external validity** from setting θ to setting θ' if for almost every measurement strategy and contrast,

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

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

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

Illustrating notions of external validity



Artifactual Discrepancies and Harmonization

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 are (nonrandom) differences in treatment effects produced by:

- Different measurement strategies
- Different contrasts

Artifactual Discrepancies and Harmonization

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 are (nonrandom) 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.

Outline

Main Results

For a collection of studies $\{\mathcal{E}_i = \{m_i, (\omega_i', \omega_i''), \theta_i\}_{i=1}^N$:

Theorem: Target-equivalence holds across i if and only if the mechanism satisfies external validity and all studies are harmonized (almost everywhere).

Main Results

For a collection of studies $\{\mathcal{E}_i = \{m_i, (\omega_i', \omega_i''), \theta_i\}_{i=1}^N$:

Theorem: Target-equivalence holds across i if and only if the mechanism satisfies external validity and all studies are harmonized (almost everywhere).

Theorem: Target-congruence holds across i if and only if the mechanism satisfies sign-congruent external validity and all studies are harmonized (a.e.).

Main Results

For a collection of studies $\{\mathcal{E}_i = \{m_i, (\omega_i', \omega_i''), \theta_i\}_{i=1}^N$:

Theorem: Target-equivalence holds across i if and only if the mechanism satisfies external validity and all studies are harmonized (almost everywhere).

Theorem: Target-congruence holds across i if and only if the mechanism satisfies sign-congruent external validity and all studies are harmonized (a.e.).

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

Outline

Comparing Signs

The **sign-comparison test** computes:

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

and tests the null hypothesis $H_0^z : \mathcal{Z} > 0$ against the alternative $H_a^z : \mathcal{Z} \leq 0$.

Comparing Signs

The **sign-comparison test** computes:

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

and tests the null hypothesis $H_0^z : \mathcal{Z} > 0$ against the alternative $H_a^z : \mathcal{Z} \leq 0$.

Proposition: If two studies $\mathcal{E}_1 = (m_1, (\omega_1^{'}, \omega_1^{''}), \theta_1)$ and $\mathcal{E}_2 = (m_2, (\omega_2^{'}, \omega_2^{''}), \theta_2)$ are harmonized, and estimation errors, $\varepsilon_1^{n_1}$ and $\varepsilon_2^{n_2}$, are unbiased and consistent, then the sign-comparison test assesses a null hypothesis of sign-congruent external validity.

Comparing Estimates

The **estimate-comparison test** computes:

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

and test the null hypothesis $H_0^w : \mathcal{W} = 0$ against the alternative $H_a^w : \mathcal{W} \neq 0$.

Comparing Estimates

The **estimate-comparison test** computes:

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

and test the null hypothesis $H_0^w : \mathcal{W} = 0$ against the alternative $H_a^w : \mathcal{W} \neq 0$.

Proposition: If two studies $\mathcal{E}_1 = (m_1, (\omega_1^{'}, \omega_1^{''}), \theta_1)$ and $\mathcal{E}_2 = (m_2, (\omega_2^{'}, \omega_2^{''}), \theta_2)$ have unbiased and consistent estimation errors, then

1. If studies 1 and 2 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 1 and 2 are harmonized.

Outline

Structural Approach

Posit a structural model of cross-study environment

- Constrain the kind of data the external world is permitted to supply

Structural Approach

Posit a structural model of cross-study environment

- Constrain the kind of data the external world is permitted to supply

Key strength: facilitates strong empirical conclusions from data,

Key drawback: cannot support causal interpretations some researchers may wish to impart to results from replication (or meta-analysis).

Design-based Approach

How to maintain causal interpretation in meta-studies?

Design-based Approach

How to maintain causal interpretation in meta-studies?

Focus on the importance of research design studies

- connected with credibility approaches to internal validity

Design-based Approach

How to maintain causal interpretation in meta-studies?

Focus on the importance of research design studies

- connected with credibility approaches to internal validity

Design-based approach to conceptual replication takes a *sequential* method that proceeds by admitting one discrepancy at a time

3-step approach:

Design-based approach to conceptual replication

Step	Description	Learning	Caveats/limitations
1.	Harmonized	External validity	Nothing about target discrepancies or external validity under different designs.
2.	Single-setting	How τ changes in design	Artifactual discrepancies may not be equivalent across settings
3.	Non-harmonized multi-study	With steps 1 and 2, see whether artifactual discrepancies vary in settings.	

Summary

General framework and concepts for evidence accumulation

Summary

General framework and concepts for evidence accumulation

- Key concepts: Target and Artifactual discrepancies

Summary

General framework and concepts for evidence accumulation

- Key concepts: Target and Artifactual discrepancies

Formally link some approaches to comparison and concepts of external validity

- External validity and sign-congruent external validity

Summary

General framework and concepts for evidence accumulation

- Key concepts: Target and Artifactual discrepancies

Formally link some approaches to comparison and concepts of external validity

- External validity and sign-congruent external validity

Advocate a *design-based approach to conceptual replication*

Sign-Congruence, External Validity, and Replication

Tara Slough & Scott A. Tyson

Thanks!!