External Validity for Social Inquiry

Michael Denly (Texas A&M)

with Michael Findley (UT Austin) and Kyosuke Kikuta (IDE—Japan)

July 19, 2024

The Reality of External Validity

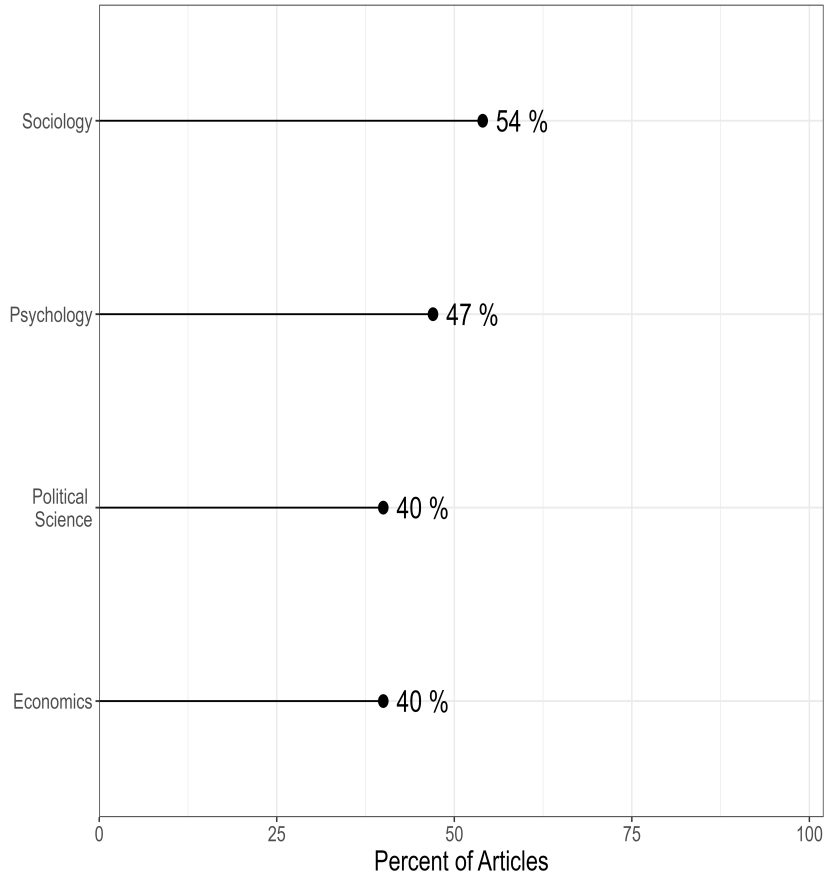
- External validity is both everywhere and nowhere, with a WEIRD bias...

The Reality of External Validity

- External validity is both everywhere and nowhere, with a WEIRD bias...

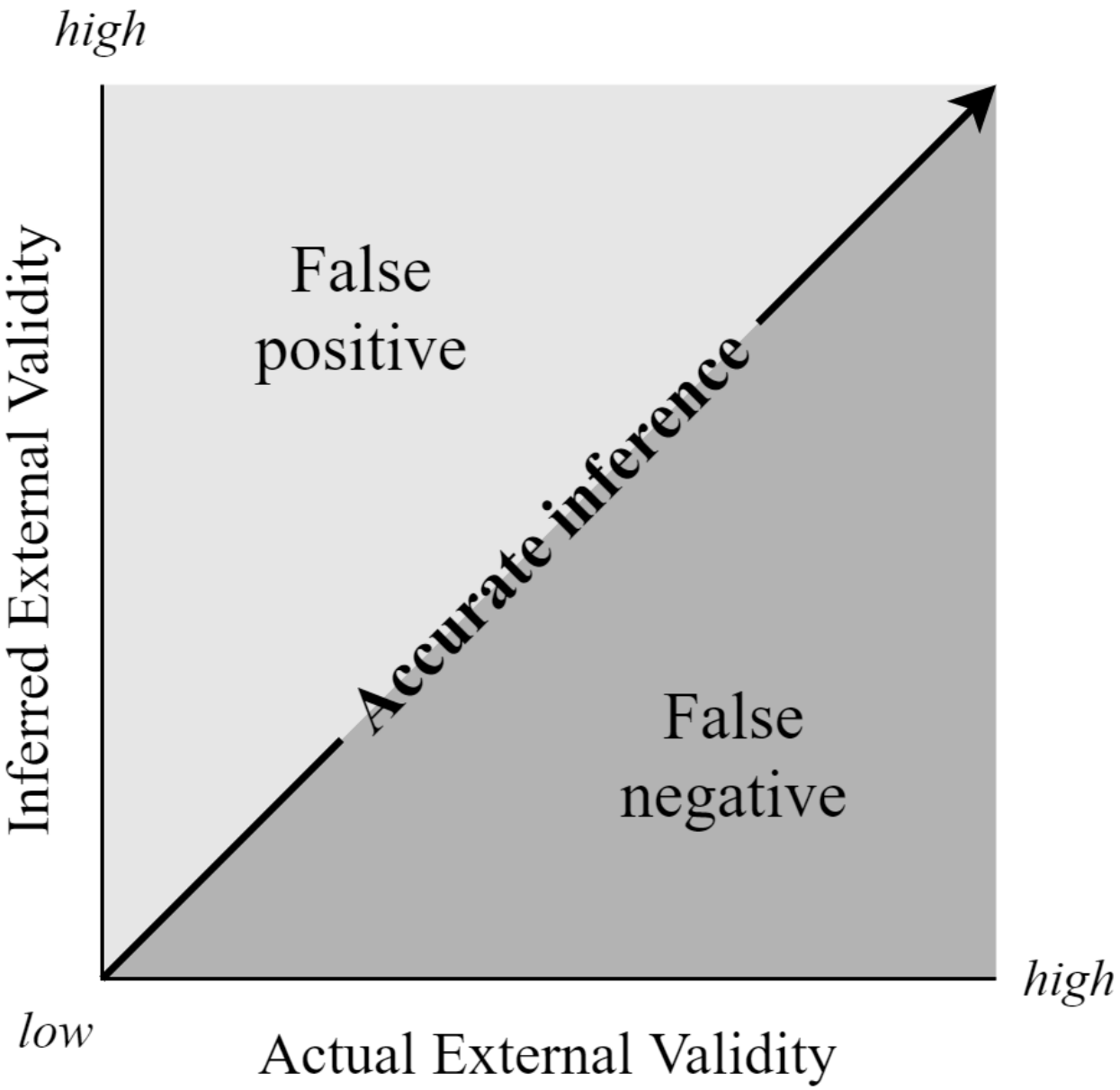
Figure: Scholarly Patterns in Making External Validity Inferences or References

(a) Inference/Reference Root Words



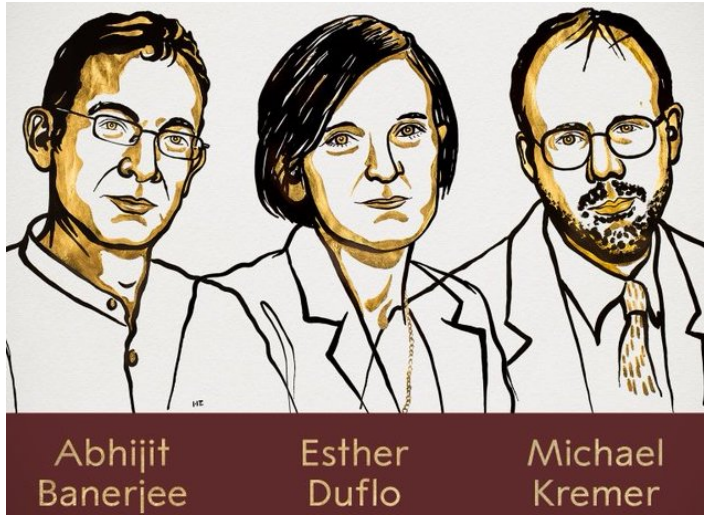
(b) Inferences/References by Field

The End Goal: Accurate Inference (EV Q-Q Plot)



Social Science's Dominant Paradigm: *All Else Equal*

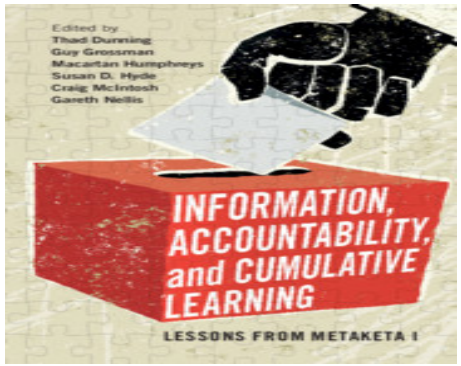
- Modern (social) science: X causes Y all else being equal
  - useful for internal validity and representing *sample* estimands (e.g., SATE, SATT, S-LATE)
- But social science is really interested in *population* estimands:
  - PATE, PATT, P-LATE
- Our PATE bias decomposition:
  - assignment selection bias (IV)
  - treatment effect heterogeneity (IV)
  - sample selection bias (EV)
  - variable selection bias (EV/CV)



Information and Accountability: Metaketa I

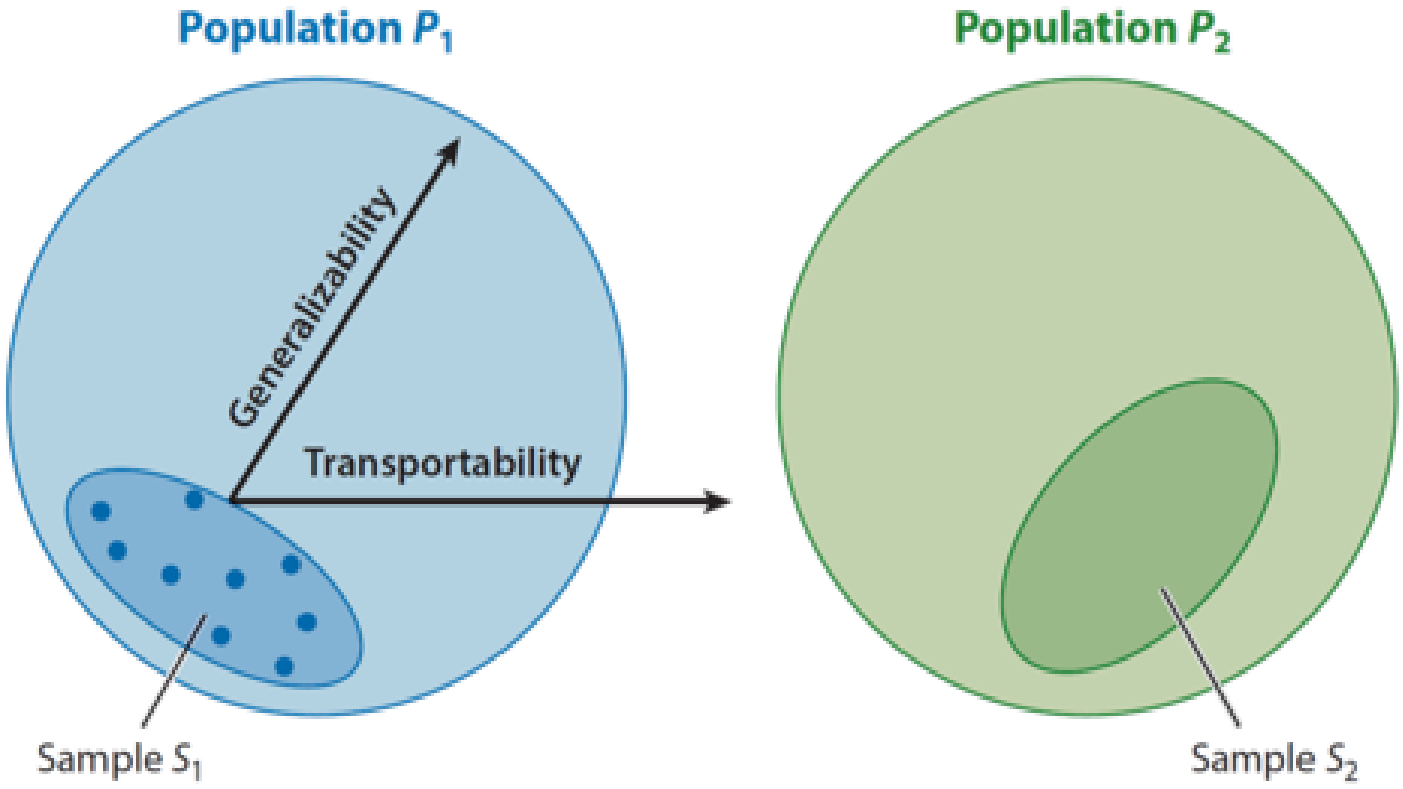
Voter information ⇒ political accountability?

- EGAP Metaketa I
  - EV Logic: Coordinating a common treatment arm and outcomes across different should reveal the extent to which a causal effect is externally valid
  - Result: Little evidence that information affects accountability

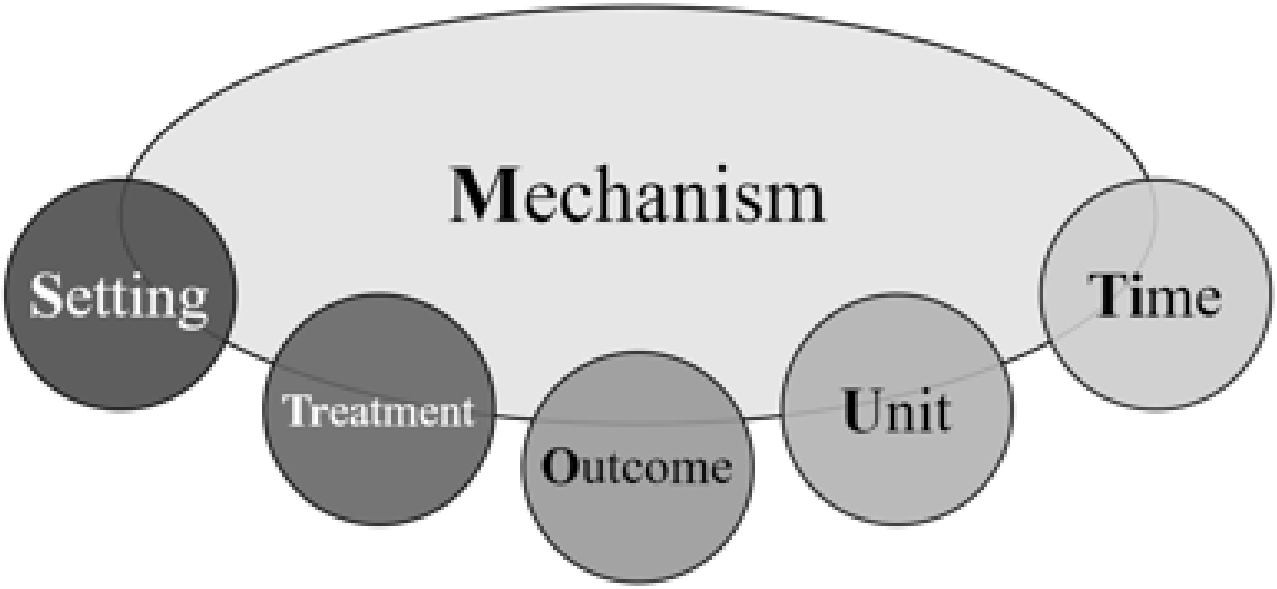


What Is External Validity? (Our Projectivist View)

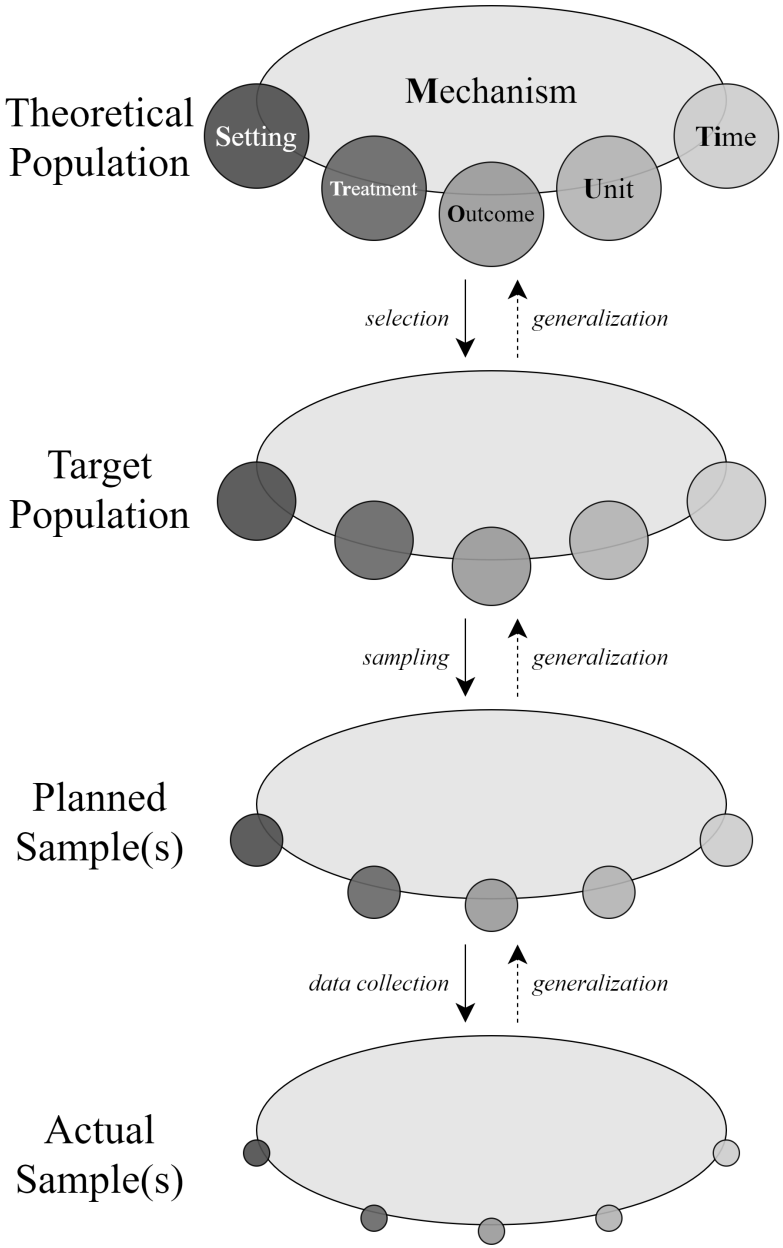
External validity captures the extent to which inferences from a given study's sample(s) apply to a broader population or other target populations.



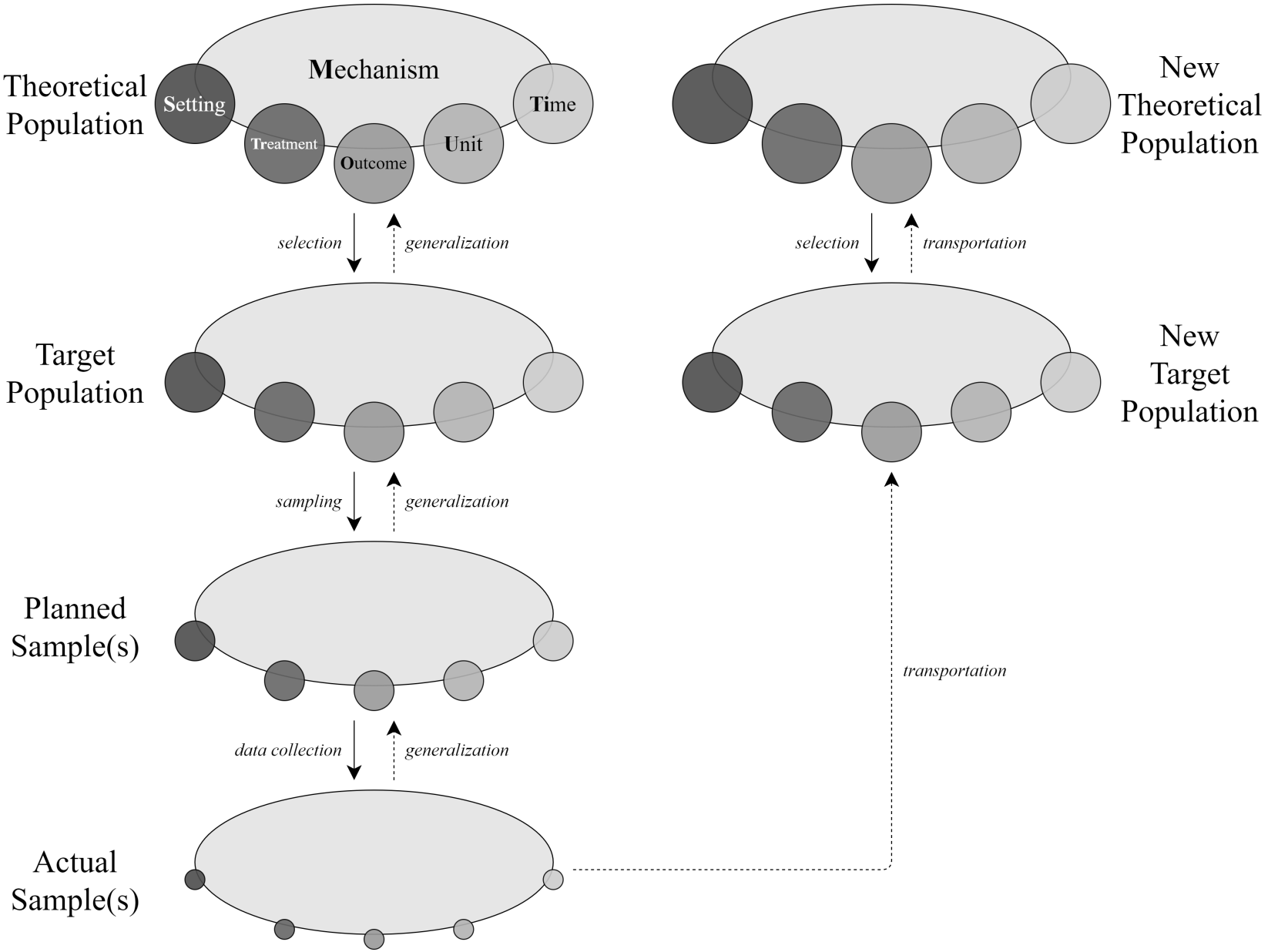
M-STOUT: The Dimensions of External Validity



M-STOUT: Positivity & Different Levels of Abstraction

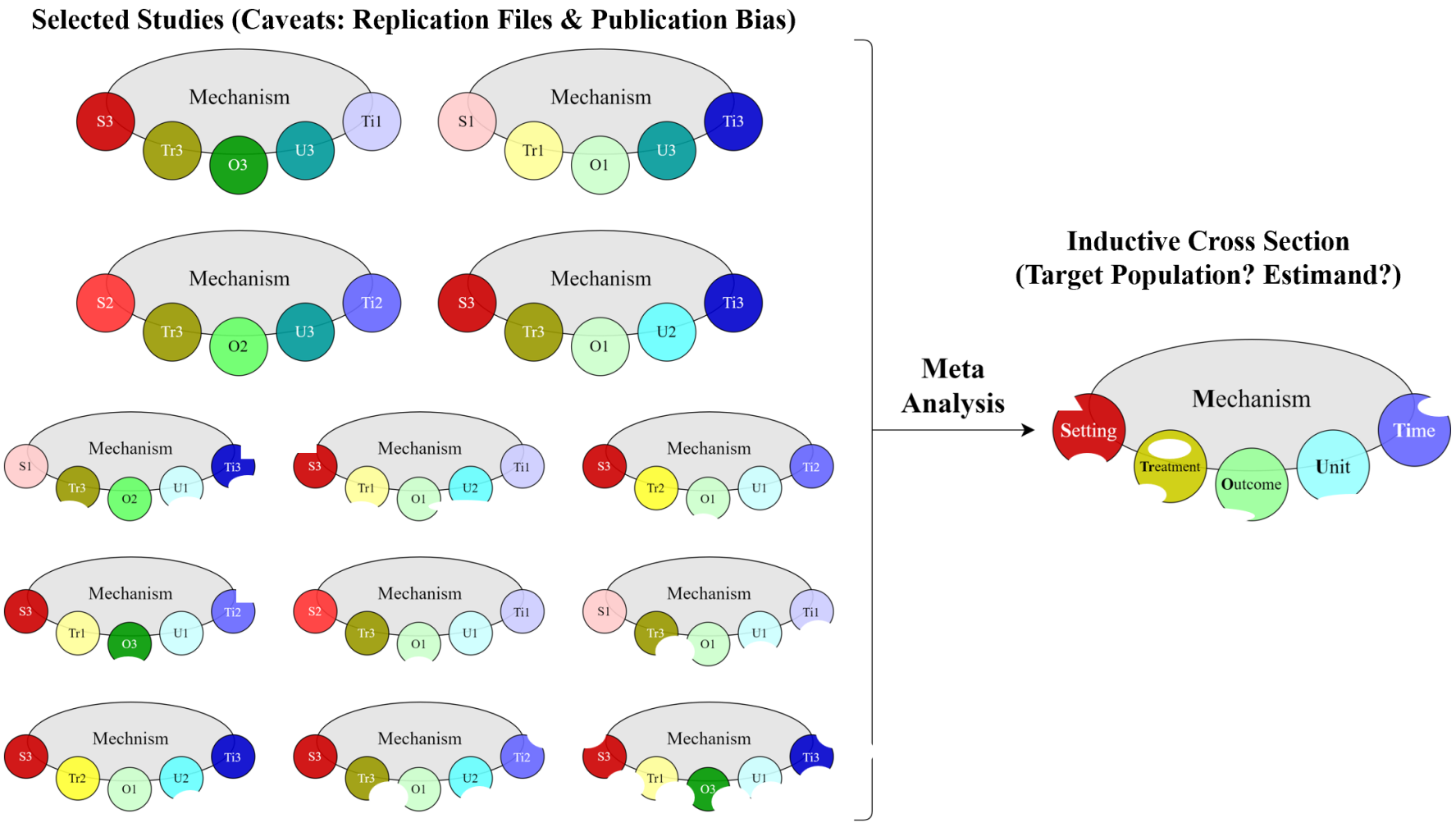


M-STOUT: Positivity & Different Levels of Abstraction

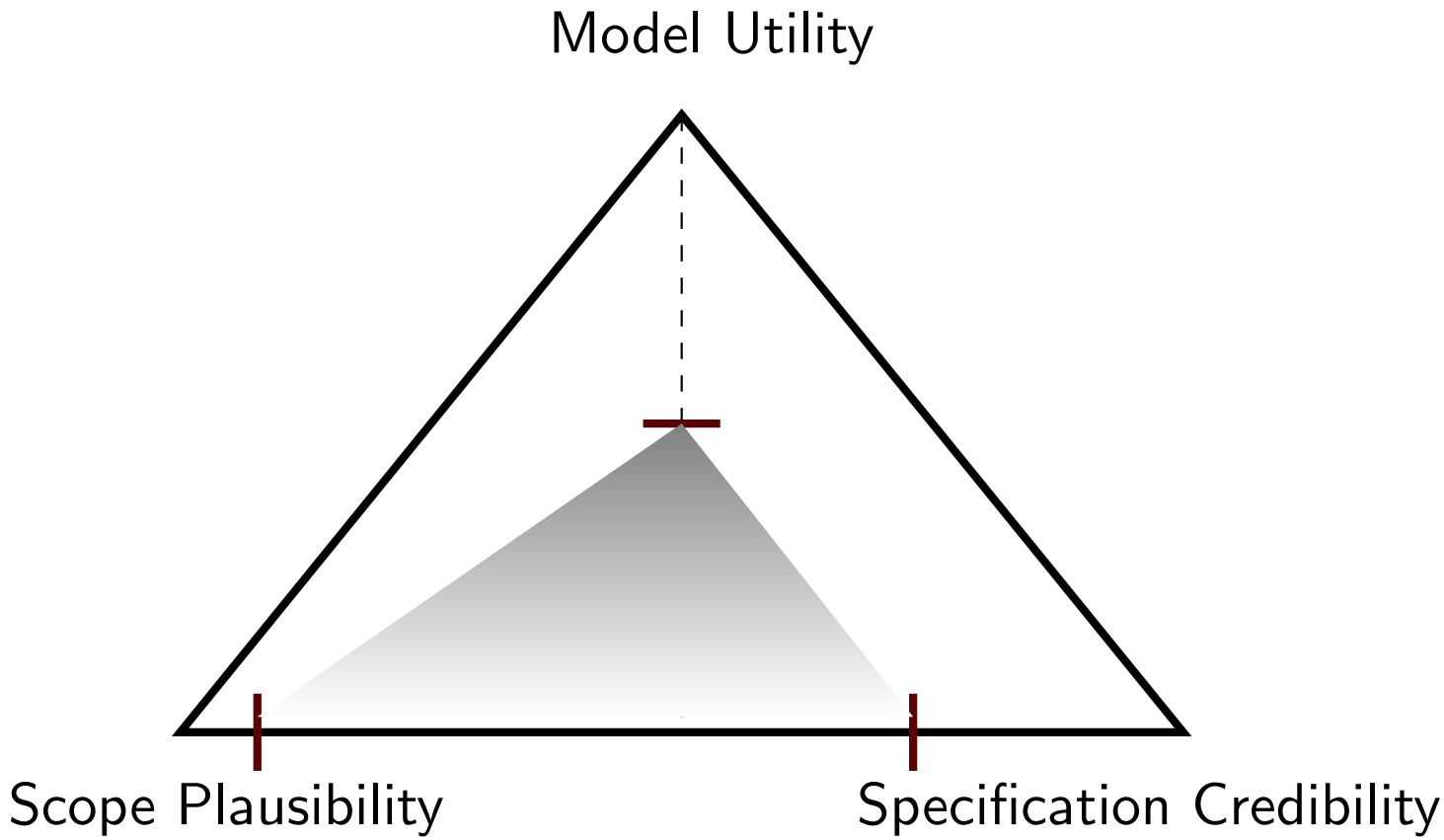


What Is External Validity? (The Cross-Sectional Take)

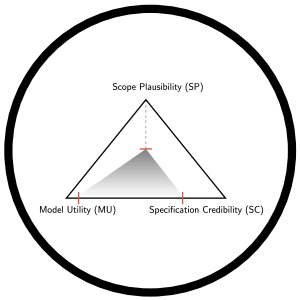
External validity is (an inductive) property of a collection of a cross-section of studies without an empirical destination (Slough & Tyson 2023)



Evaluative Criteria for Improving External Validity



Evaluative Criteria: Model Utility



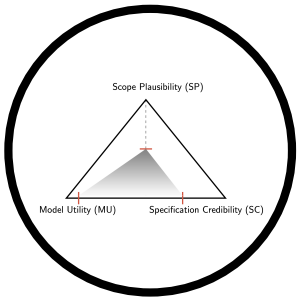
Components	**Strategies**
Research Goal/ Query	
Mechanism: Causal Structure	
Mechanism Representation	

△

A useful model refers to one with a feasible research query that can recognize heterogeneity, invoke mechanisms (including their support factors & constraints), and represent them with a clear estimand.

Evaluative Criteria: Model Utility

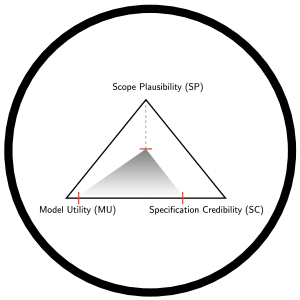
Components	**Strategies**
Research Goal/ Query	



△

Define a feasible research goal with a specific query about a desired quantity of interest and for which populations

Evaluative Criteria: Model Utility



Components	Strategies
Research Goal/ Query	- Feasible research question - Dunning et al (2019, 6-7): "Do info. campaigns...in the lead-up to elections influence voter behavior...and, if so, under what conditions?" - Heterogeneity is everywhere, so (mostly) middle-range theory and partial equilibrium

△

Define a feasible research goal with a specific query about a desired quantity of interest and for which populations

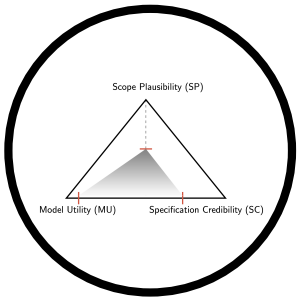
Evaluative Criteria: Model Utility

Components

Research Goal/
Query

Mechanism:
Causal Structure

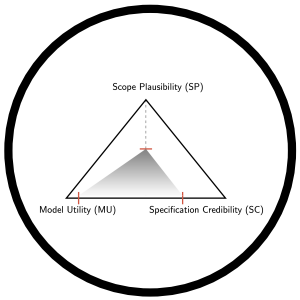
Strategies



△

Make external validity inferences primarily about mechanisms with a clear articulation of their support factors, constraints, and how they encode STOUT differences.

Evaluative Criteria: Model Utility

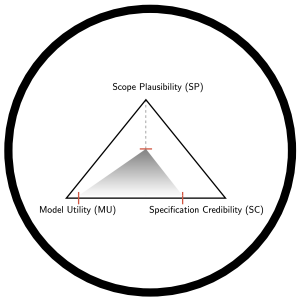


Components	Strategies
Research Goal/ Query	- Key: STOUT support factors ⇒ Assumptions
Mechanism: Causal Structure	- $P(Y \mid X, \mathbf{M})$: ⇒ Access to credible info sources; ⇒ Salient performance info (publicly) disseminated; ⇒ Ability to express political preferences; ⇒ Literate people to process info; ⇒ Free from poverty/clientelism;

△

Make external validity inferences primarily about mechanisms with a clear articulation of their support factors, constraints, and how they encode STOUT differences.

Evaluative Criteria: Model Utility



Components

Research Goal/ Query

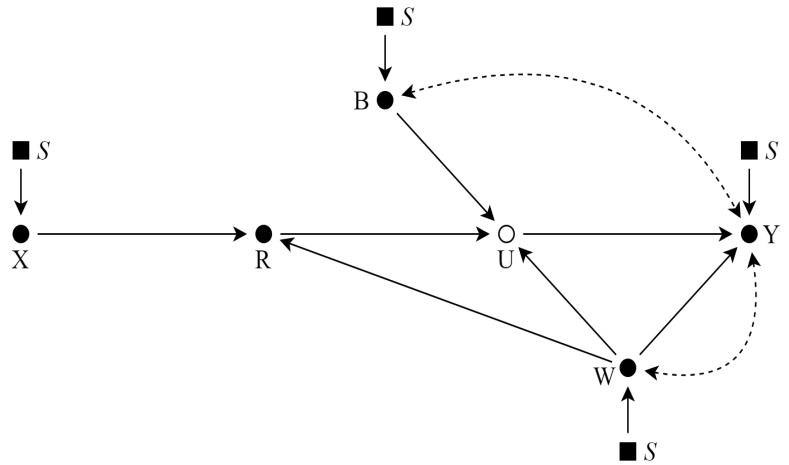
Mechanism: Causal Structure

Mechanism Representation

Strategies

Theoretical Estimand:

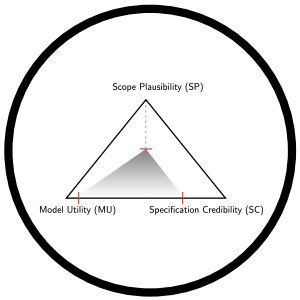
$$\tau = \frac{1}{n} \sum_{i=u_1,u_2 \ldots u_N | S=s, T=t}^{n} \Big(Y(\text{Info}, \mathbf{M}) - Y(\text{No Info}, \mathbf{M}) \Big)$$



△

A useful model representation captures the theoretical estimand and how it encodes mechanism support factors and form

Evaluative Criteria: Model Utility



Components

Research Goal/ Query

Mechanism: Causal Structure

Mechanism Representation

Strategies

"A mechanism seldom operates on its own. We sometimes pretend they do when trying to describe their 'natural' outcomes. But that makes no sense. Mechanisms don't operate in some kind of Platonic heaven 'all by themselves'. They operate in real settings. And in real settings other real things have influence as well." - Cartwright (2020)

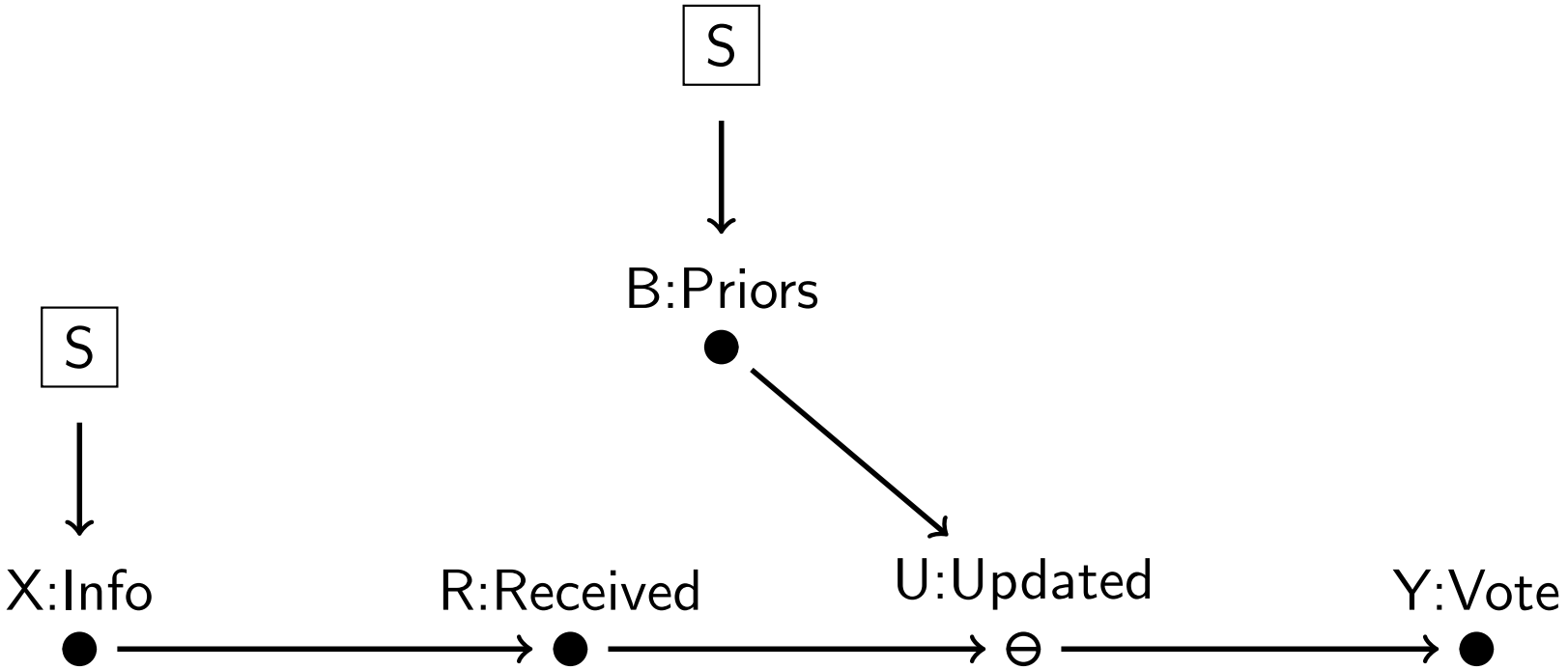
△

A useful model representation captures the theoretical estimand and how it encodes mechanism support factors and form

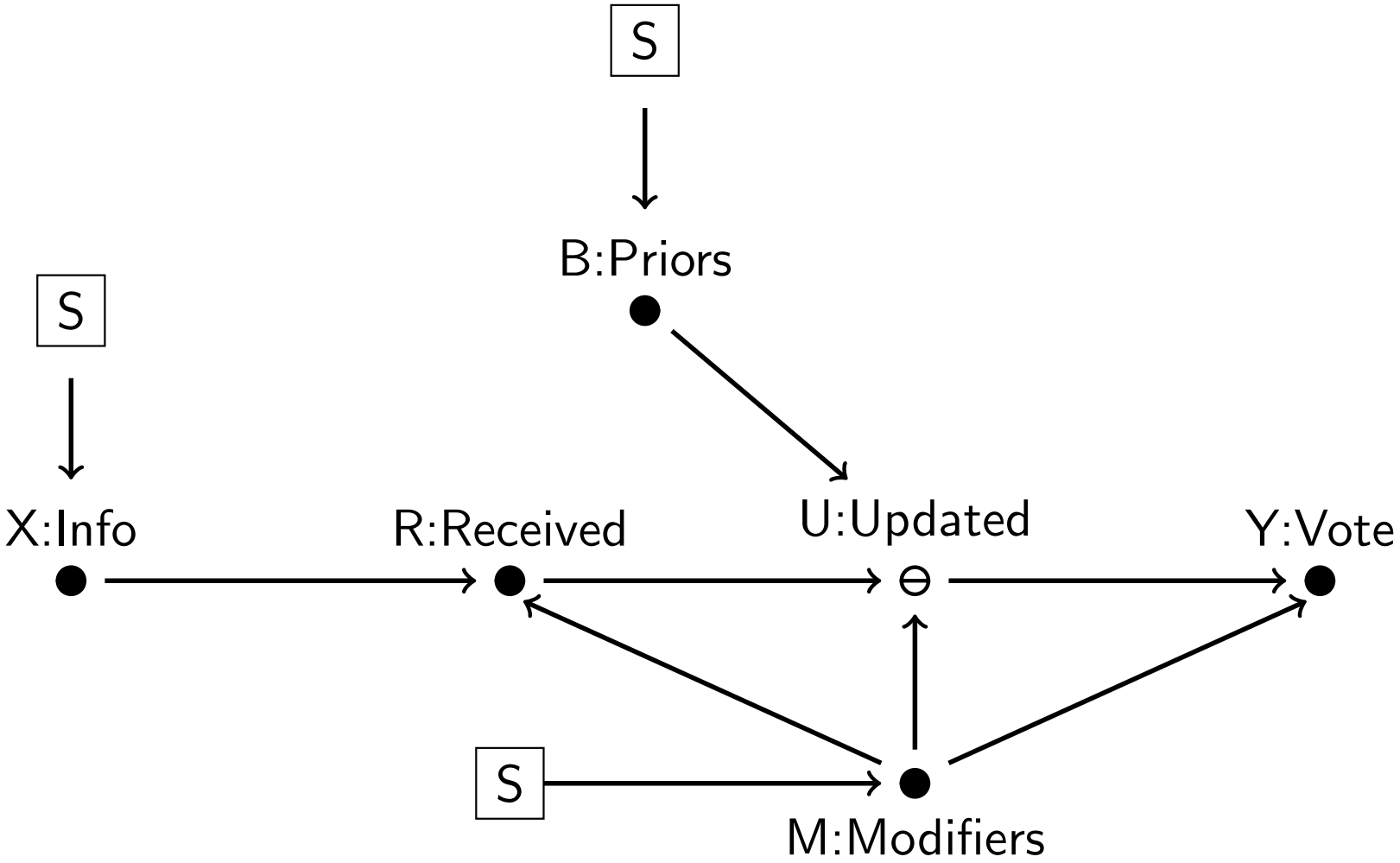
Metaketa I Causal Structure

Metaketa I Causal Structure

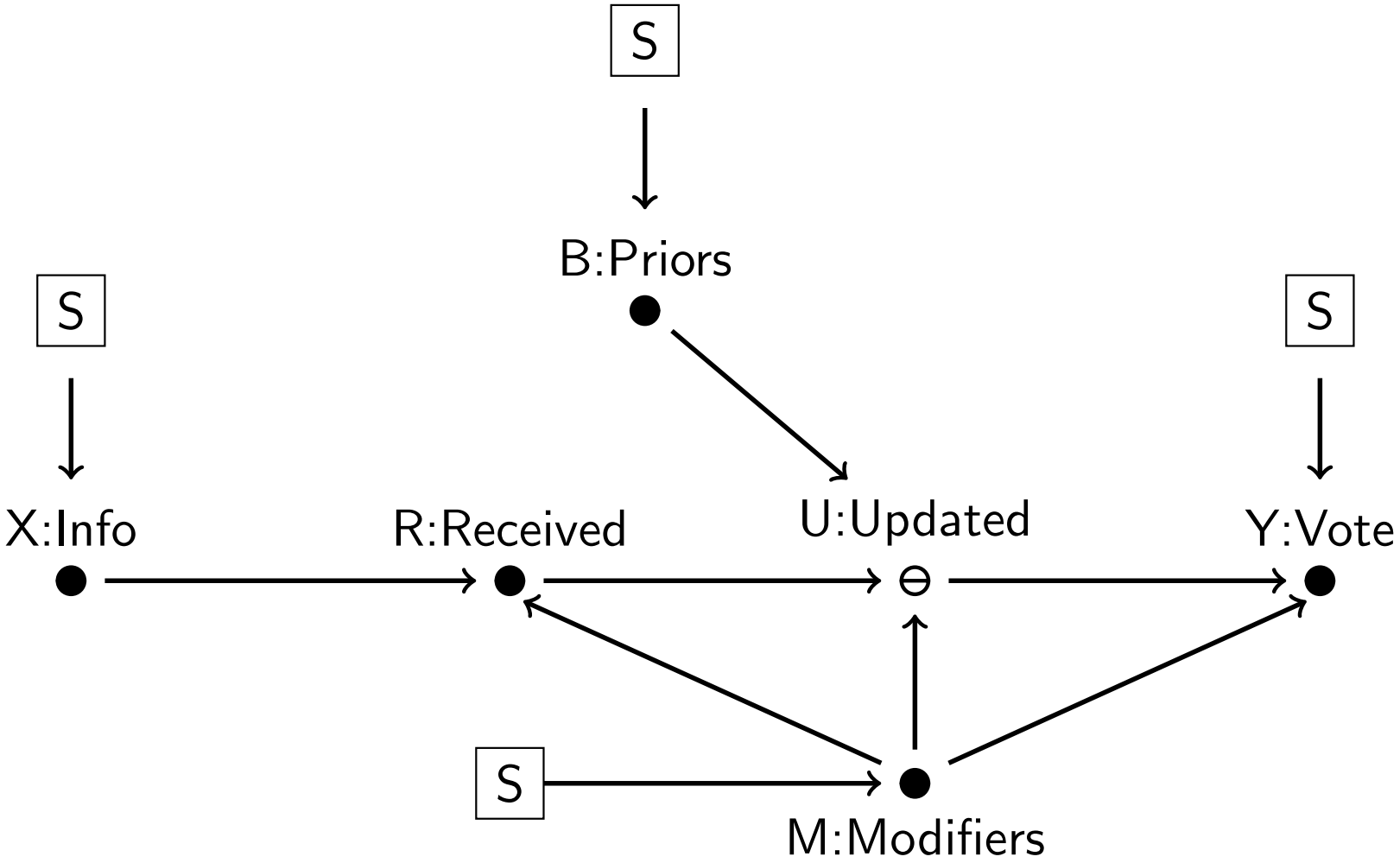
Metaketa I Causal Structure



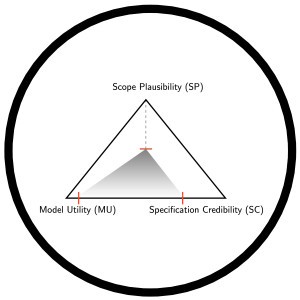
Metaketa I Causal Structure



Metaketa I Causal Structure



Evaluative Criteria: Scope Plausibility



Components	Strategies
Target Population	
Planned Sample	
Actual Sample & Empirical Estimand	

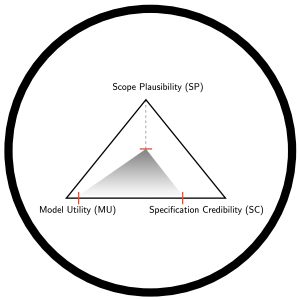
△

Design a feasible scope based on the model and positivity of the target population, planned sample, actual sample, taking into account the chosen estimand.

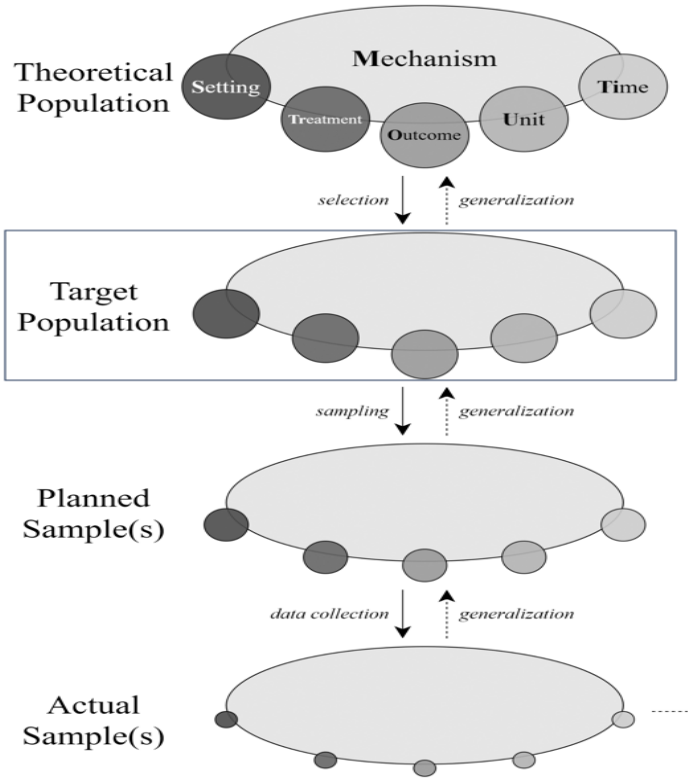
Evaluative Criteria: Scope Plausibility

Components

Target Population



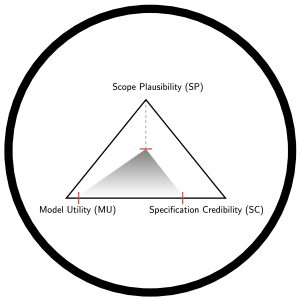
Strategies



△

Select target populations that maximize coverage of the theoretical populations and make empirical investigation of causal and sampling heterogeneity possible.

Evaluative Criteria: Scope Plausibility

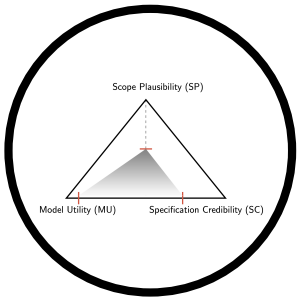


Components	**Strategies**
Target Population	- Select sampling frame reflecting theoretical population and mechanisms ⇒ Eligibility criteria ⇒ Avoid unrealistic targets - Ensure sampling frame has representation of key subgroups/strata ⇒ Causal heterogeneity ⇒ Sampling heterogeneity

△

Select target populations that maximize coverage of the theoretical populations and make empirical investigation of causal and sampling heterogeneity possible.

Evaluative Criteria: Scope Plausibility

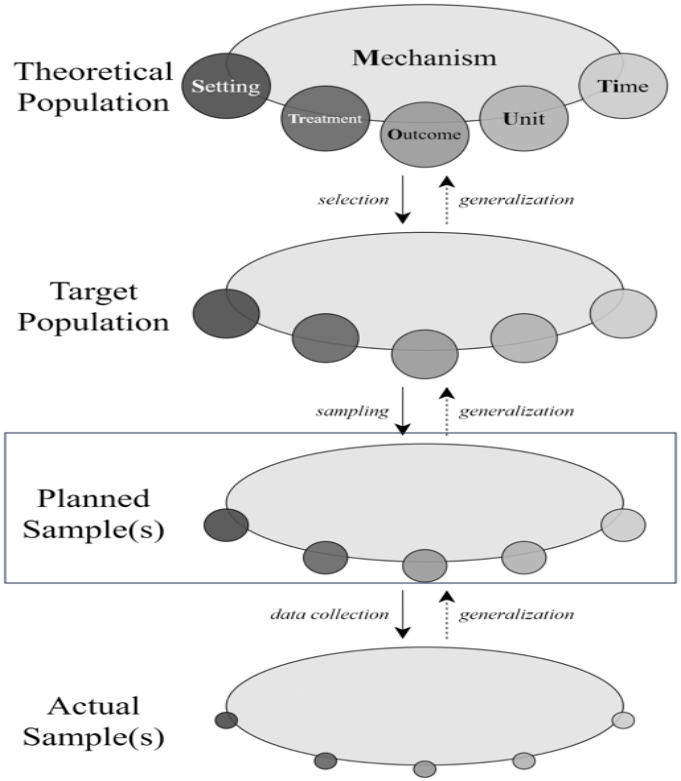


Components

Target Population

Planned Sample

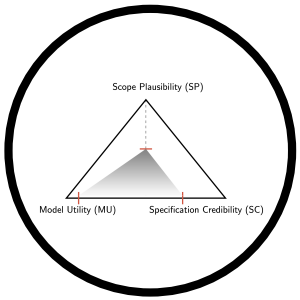
Strategies



△

Select planned STOUT samples with a precise mapping to the target population, carefully distinguishing causal and sampling heterogeneity, while paying attention to construct validity

Evaluative Criteria: Scope Plausibility

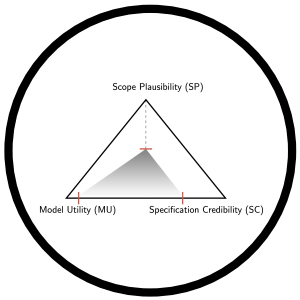


Components	Strategies
Target Population	- Sufficiently-powered sample that captures mechanism-specific strata ⇒ (Plausibly) stratified random sampling as a benchmark
Planned Sample	- Construct validity: ⇒ Abstraction and shielding

---

Select planned STOUT samples with a precise mapping △ to the target population, carefully distinguishing causal and sampling heterogeneity, while paying attention to construct validity

Evaluative Criteria: Scope Plausibility



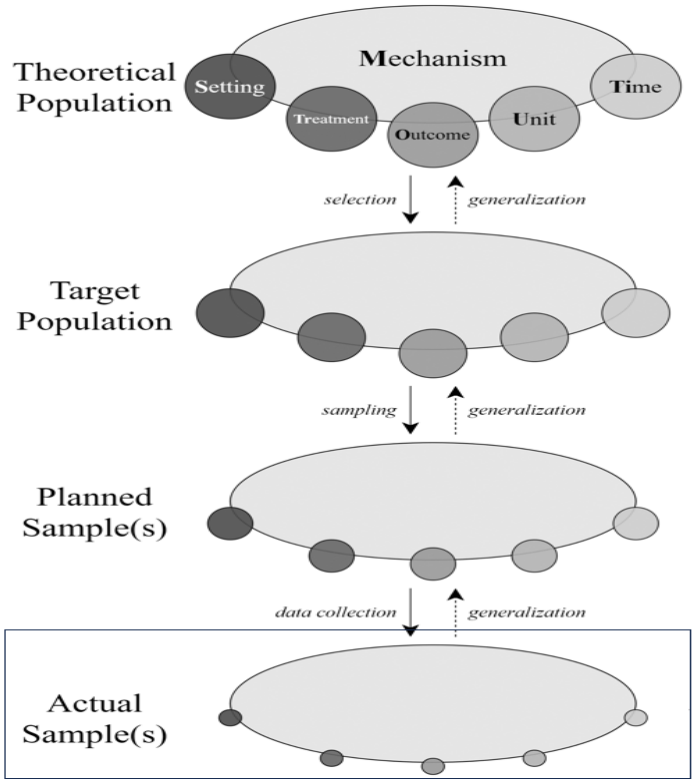
Components

Target Population

Planned Sample

Actual Sample &
Empirical Estimand

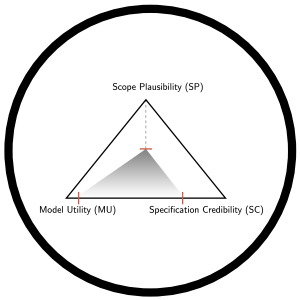
Strategies



△

Characterize the actual samples and empirical estimand that remain given the **chosen** method, samples, and inferential challenges.

Evaluative Criteria: Scope Plausibility



Components	Strategies
Target Population	- Actual samples may not correspond to the target population due to: ⇒ Attrition/missingness ⇒ Noncompliance ⇒ Spillover
Planned Sample	
Actual Sample & Empirical Estimand	- Empirical estimand may differ from theoretical estimand: ⇒ Solution: Empirical ID Tests

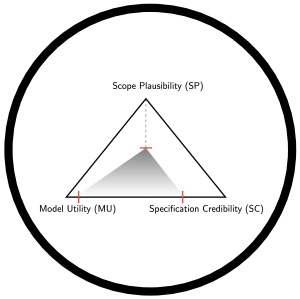
△

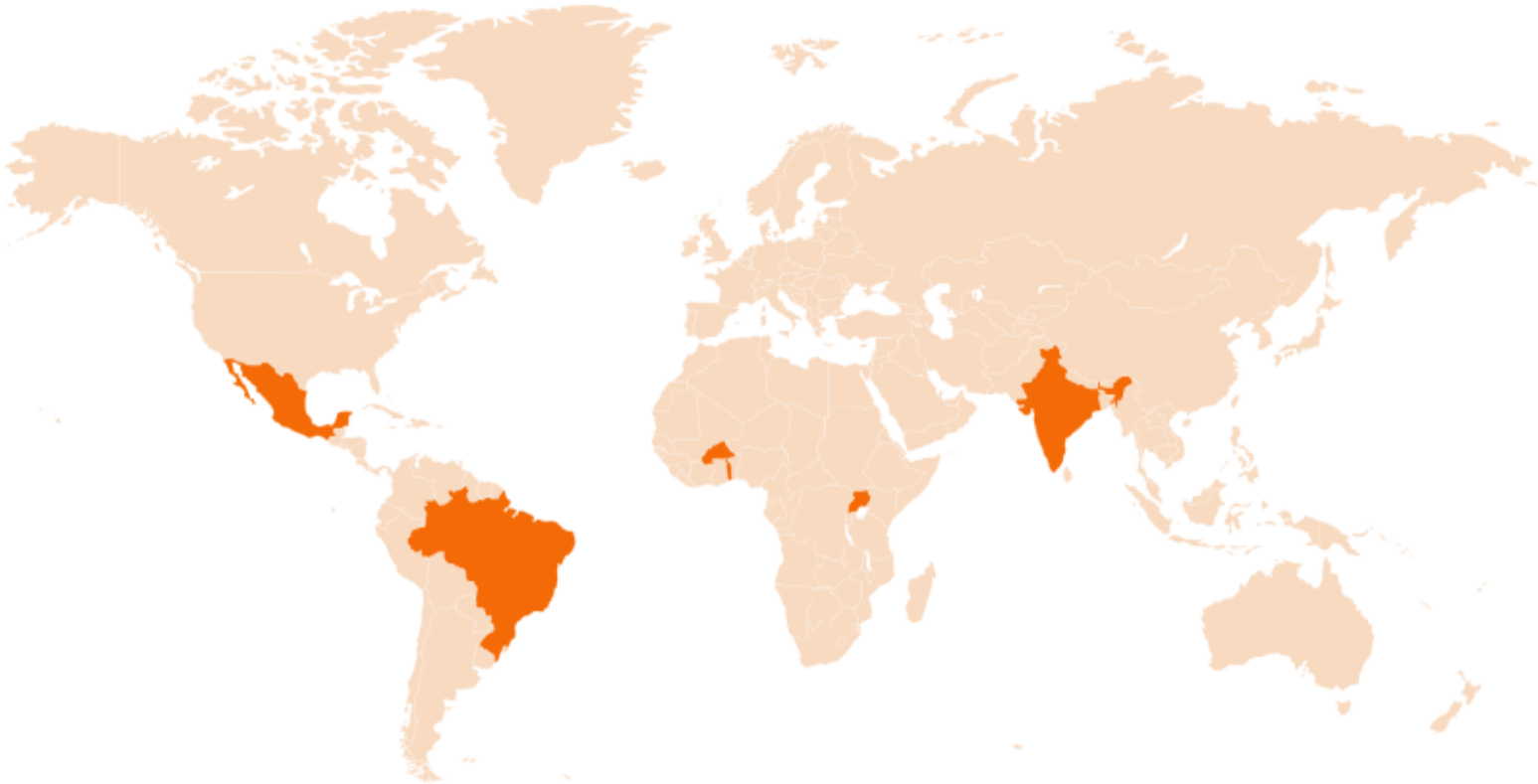
Characterize the actual samples and empirical estimand that remain given the **chosen** method, samples, and inferential challenges.

Evaluative Criteria: Scope Plausibility

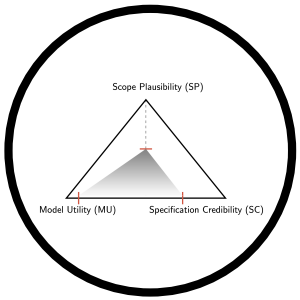
Scope Plausibility

Metaketa I



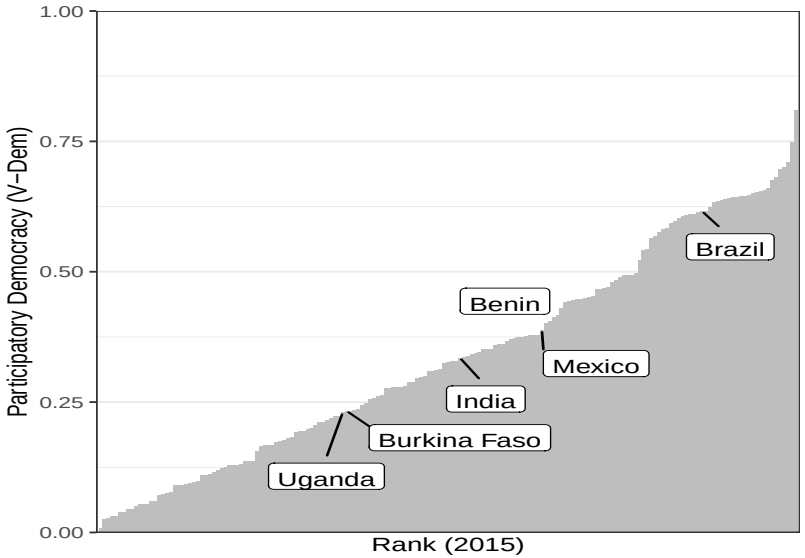


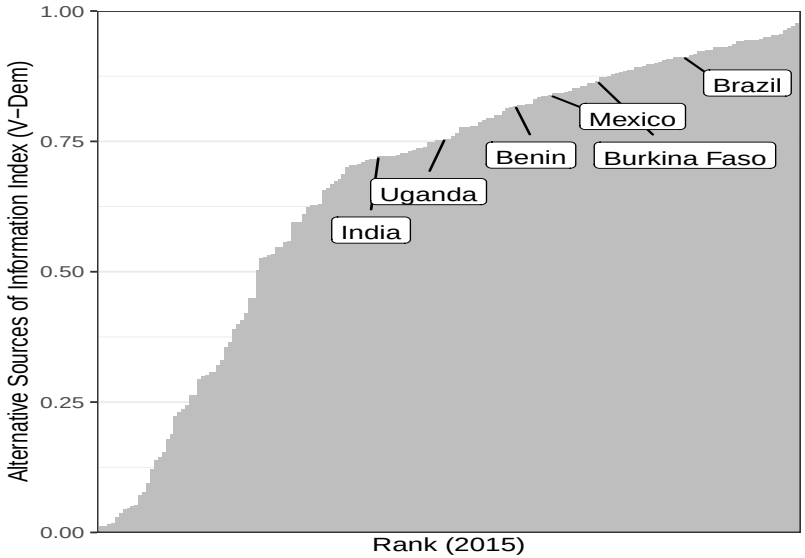
Evaluative Criteria: Scope Plausibility

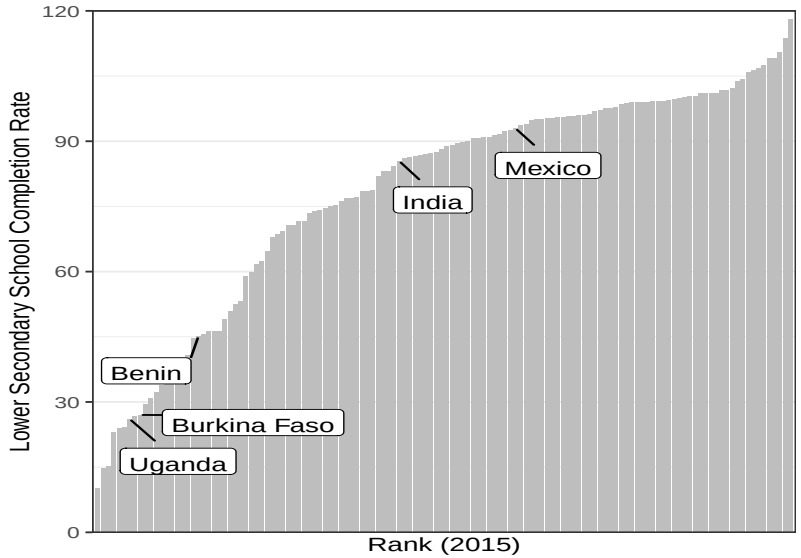


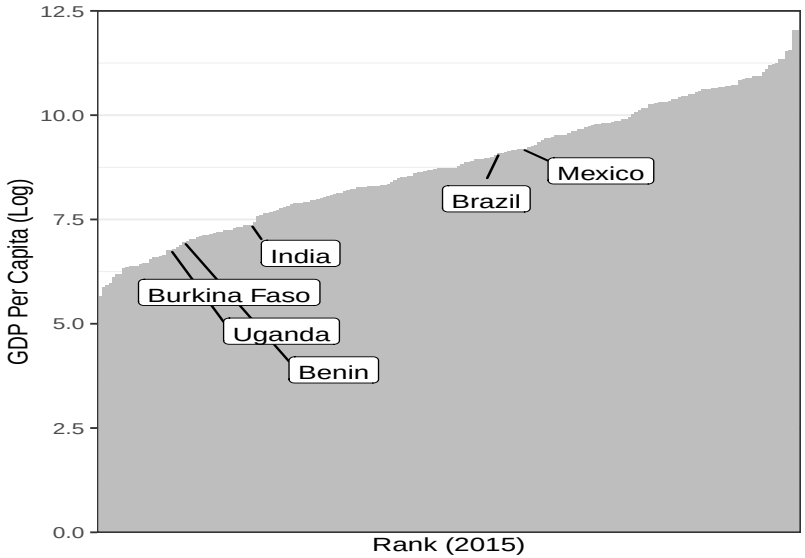
Scope Plausibility

Metaketa I









Evaluative Criteria: Specification Credibility

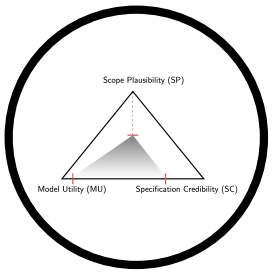
Components

Estimating the PATE/TATE

Assessing Mechanisms

Addressing Uncertainty

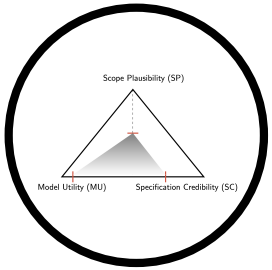
Strategies



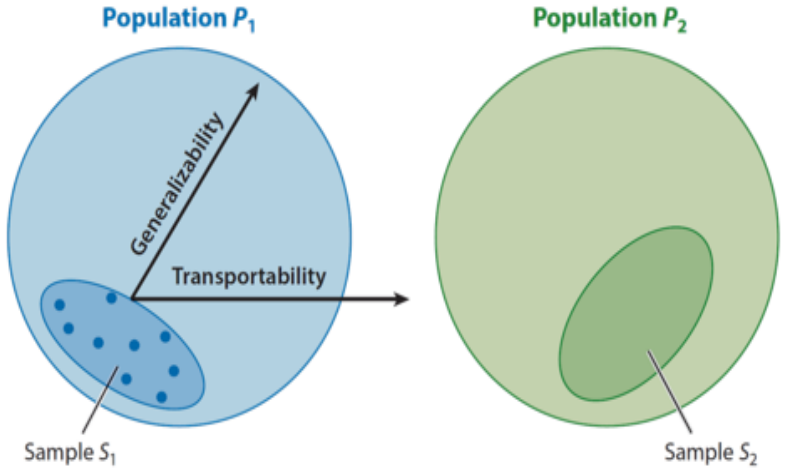
Evaluative Criteria: Specification Credibility

Components

Estimating the PATE/TATE



Strategies



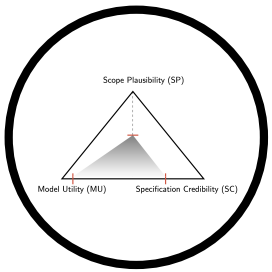
△

Use covariate- and model-based adjustment for units, relevant non-unit adjustment, and account for generalizability vs transportability to make falsiable inferences

Evaluative Criteria: Specification Credibility

Components

Estimating the PATE/TATE



Strategies

- Covariate-based adjustment
  - ⇒ Weighting
  - ⇒ Matching
  - ⇒ Subclassification
- Model-based adjustment
  - ⇒ Outcome regression
  - ⇒ Doubly-robust approaches

△

Use covariate- and model-based adjustment for units, relevant non-unit adjustment, and account generalizability and transportability to make falsiable inferences

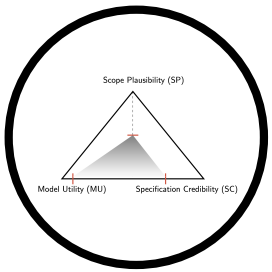
Evaluative Criteria: Specification Credibility

Components

Strategies

Estimating the PATE/TATE

Assessing Mechanisms

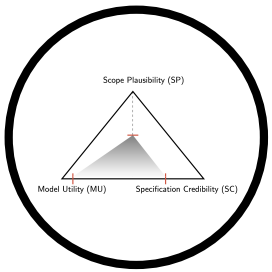


△

Isolate mechanisms in the study samples, develop discriminant criteria, evaluate mechanisms in the population, and assess mechanism regularity.

Evaluative Criteria: Specification Credibility

Components	Strategies
Estimating the PATE/TATE	
Assessing Mechanisms	- Establish mechanisms in the sample and population ⇒ Moderation/mediation ⇒ Intermediate outcomes? ⇒ Pathway analysis ⇒ Process tracing



△

Isolate mechanisms in the study samples, develop discriminant criteria, evaluate mechanisms in the population, and assess mechanism regularity.

Evaluative Criteria: Specification Credibility

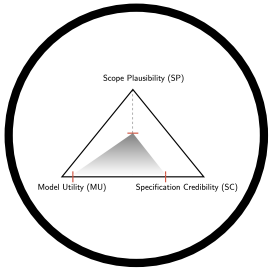
Components

Strategies

Estimating the PATE/TATE

Assessing Mechanisms

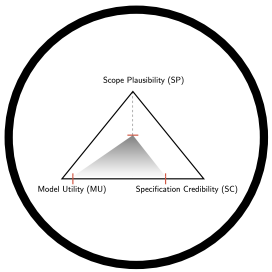
Addressing Uncertainty



△

Employ alternative estimands, conduct robustness tests, and synthesize within- and across-study evidence

Evaluative Criteria: Specification Credibility



Components

Estimating the PATE/TATE

Assessing Mechanisms

Addressing Uncertainty

Strategies

- Alternative estimands:
⇒ Different estimators
⇒ Bounding/interval/sensitivity analysis

- Robustness tests:
⇒ *STOUT* positivity
⇒ Counterfactual analyses

- Synthesis methods
⇒ Meta analysis
⇒ Replication

△

Employ alternative estimands, conduct robustness tests, and synthesize within- and across-study evidence

Evaluative Criteria (Summarized)

Model Utility

1. Research Goal/Query
2. Mechanisms: Causal Structure
3. Mechanism Representation

Scope Plausibility

1. Target Population
2. Planned Sample
3. Actual Sample & Empirical Estimand

Specification Credibility

1. Estimating the PATE/TATE
2. Assessing Mechanisms
3. Addressing Uncertainty

Methodological Applications of Evaluative Criteria

Part 1: Fundamentals

1. Introduction
2. Scope, Populations, and Samples
3. Inference Objectives

Part 2: Evaluative Criteria

4. Model Utility
5. Scope Plausibility
6. Specification Credibility

Part 3: Methodological Applications

7. **Experiments**
8. **Natural Experiments**
9. **Quantitative Observational**
10. **Qualitative Methods**
11. **Research Synthesis**

Part 4: Progress

12. Reporting & Conclusion

Summing Up/The Way Forward

- Every study should report on external validity as a matter of course:
  - Avoid false positives and false negatives!
- Let's not ditch the credibility revolution BUT EXTEND IT...
  - Pay closer attention to mechanisms, assumptions, and estimands
- We have proposed new Evaluative Criteria to improve external validity
  - Model Utility
  - Scope Plausibility
  - Specification Credibility

Decomposition of External Validity Bias

Terms

PATE = Population Average Treatment Effect (Generalizability)

TATE = Target Average Treatment Effect (Transportability)

$$Estimate = PATE\ (TATE) + Bias + Noise$$

$$Bias = Internal\ validity\ biases\ + External\ validity\ biases$$

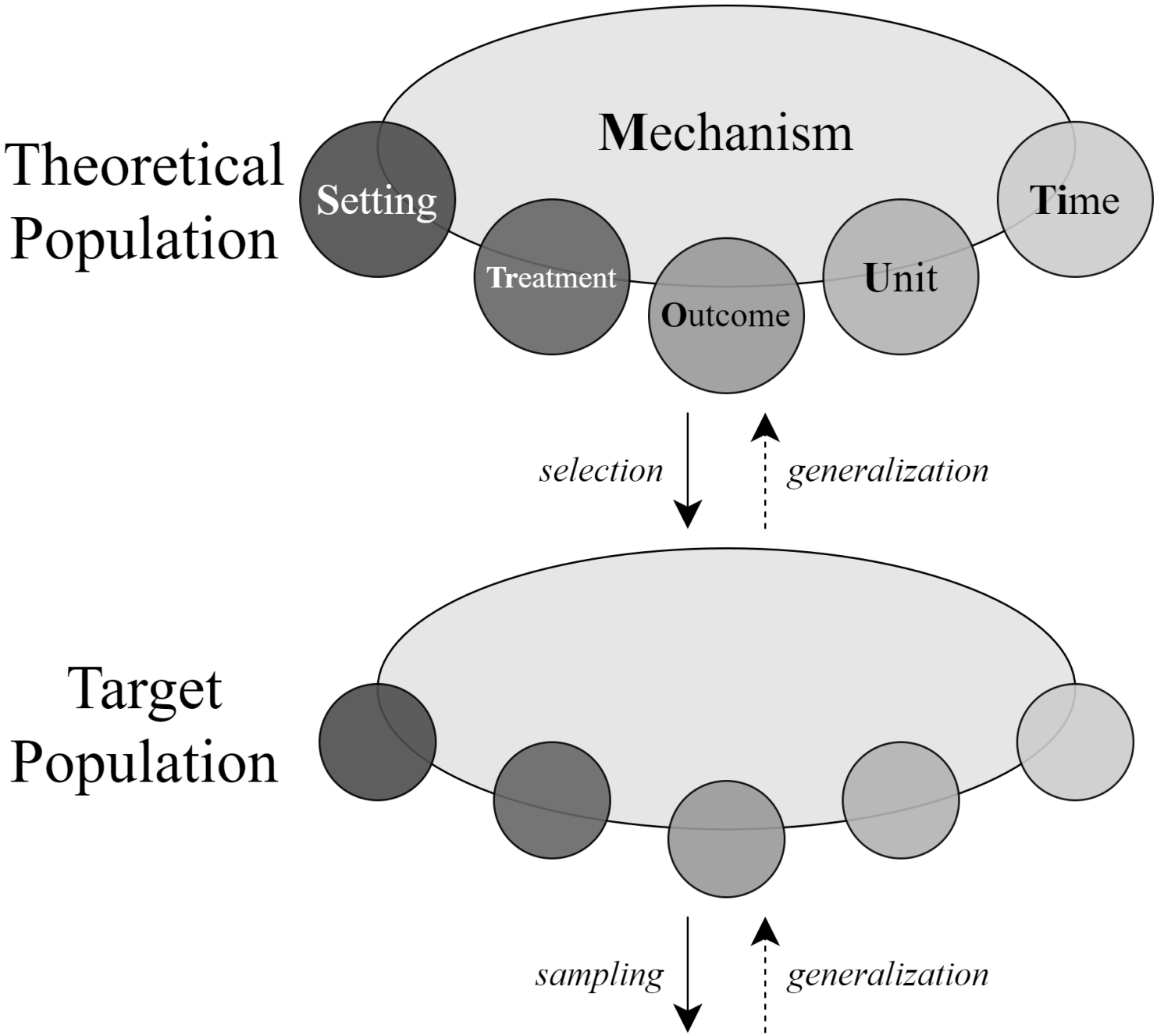
$$= \underbrace{assignment\ selection\ bias}_{Internal\ Validity} + \underbrace{treatment\ effect\ heterogeneity}_{Internal\ Validity}$$
$$+ \underbrace{sample\ selection\ bias}_{External\ Validity} + \underbrace{variable\ selection\ bias}_{External\ Validity}$$

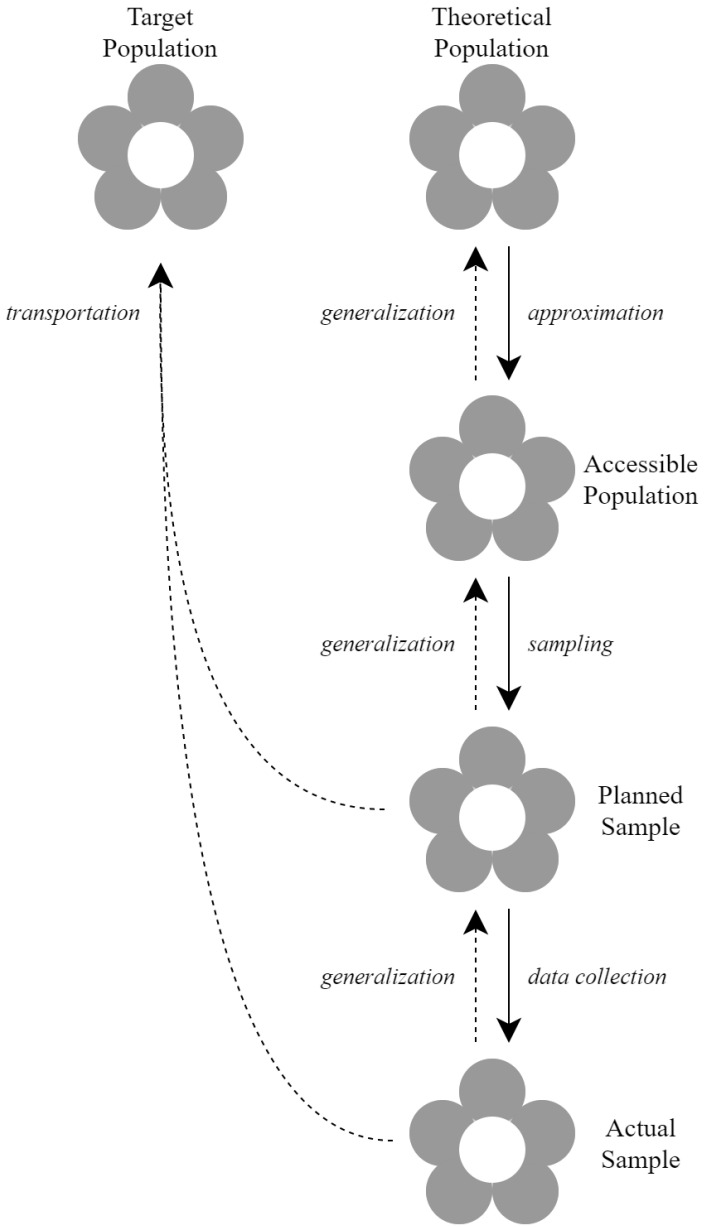
Appendix Slides

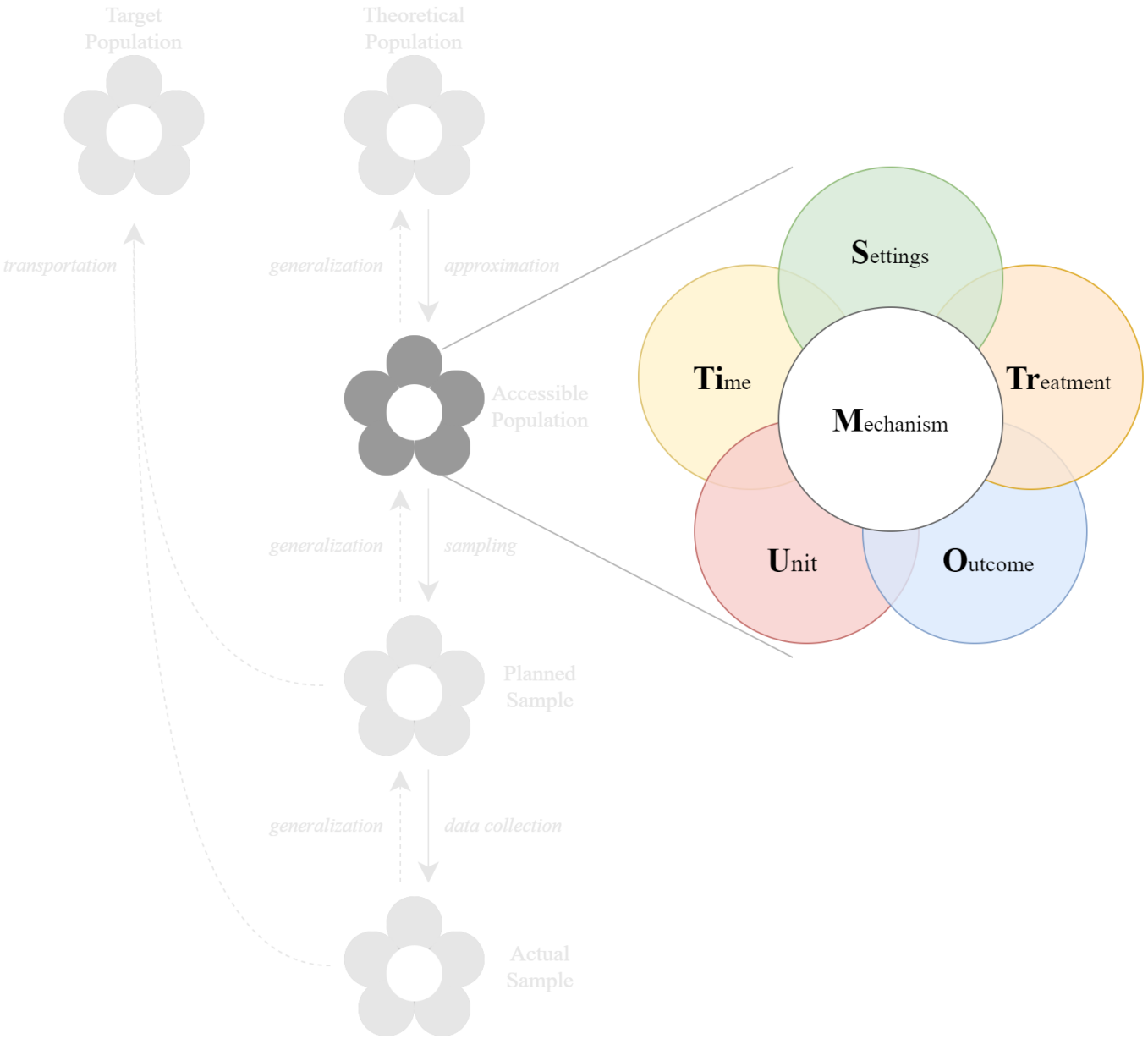
Specification pathologies undermine EV integrity

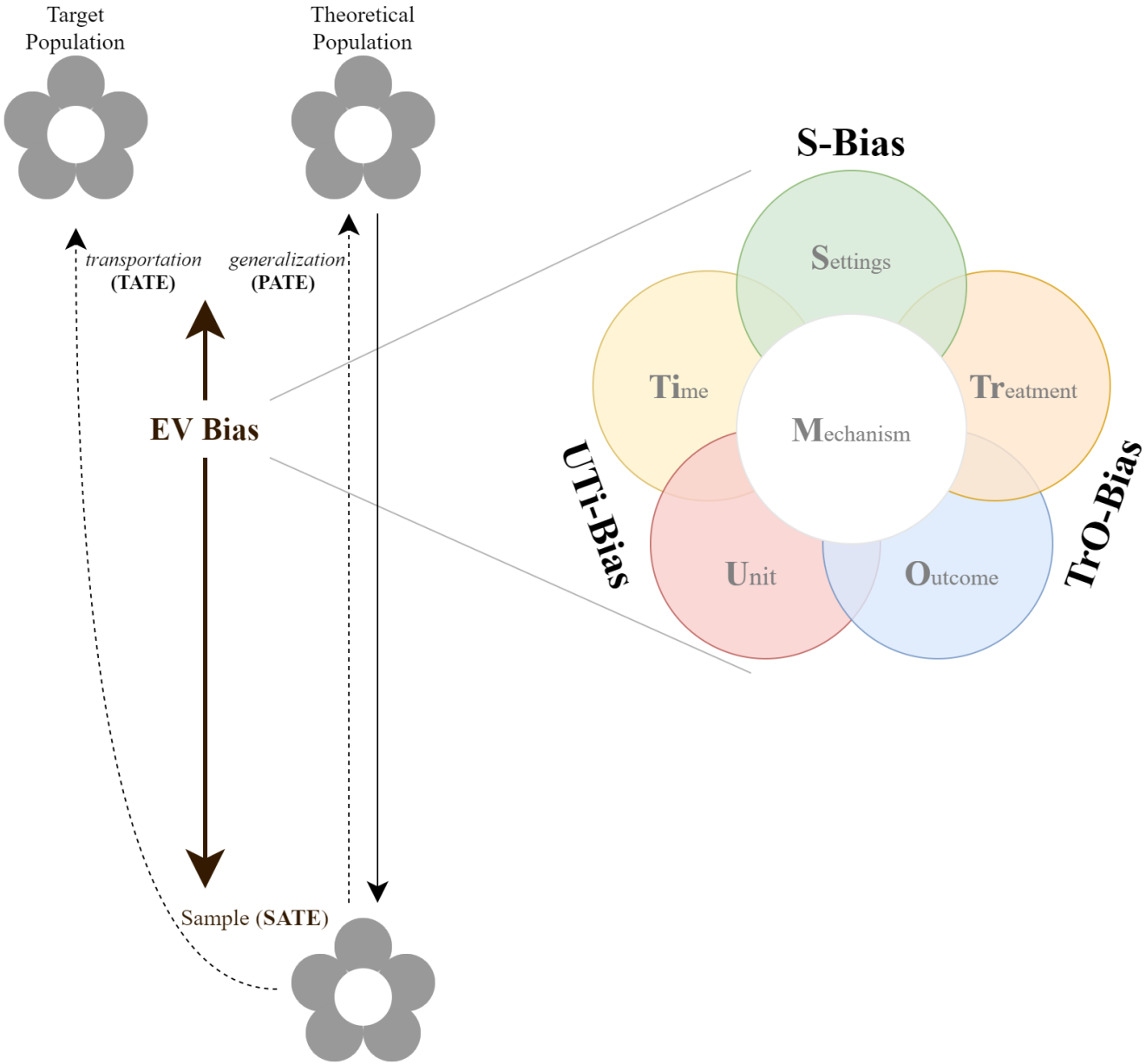
Table: Pathologies Mitigating All Else Equal and Specification Credibility in Empirics

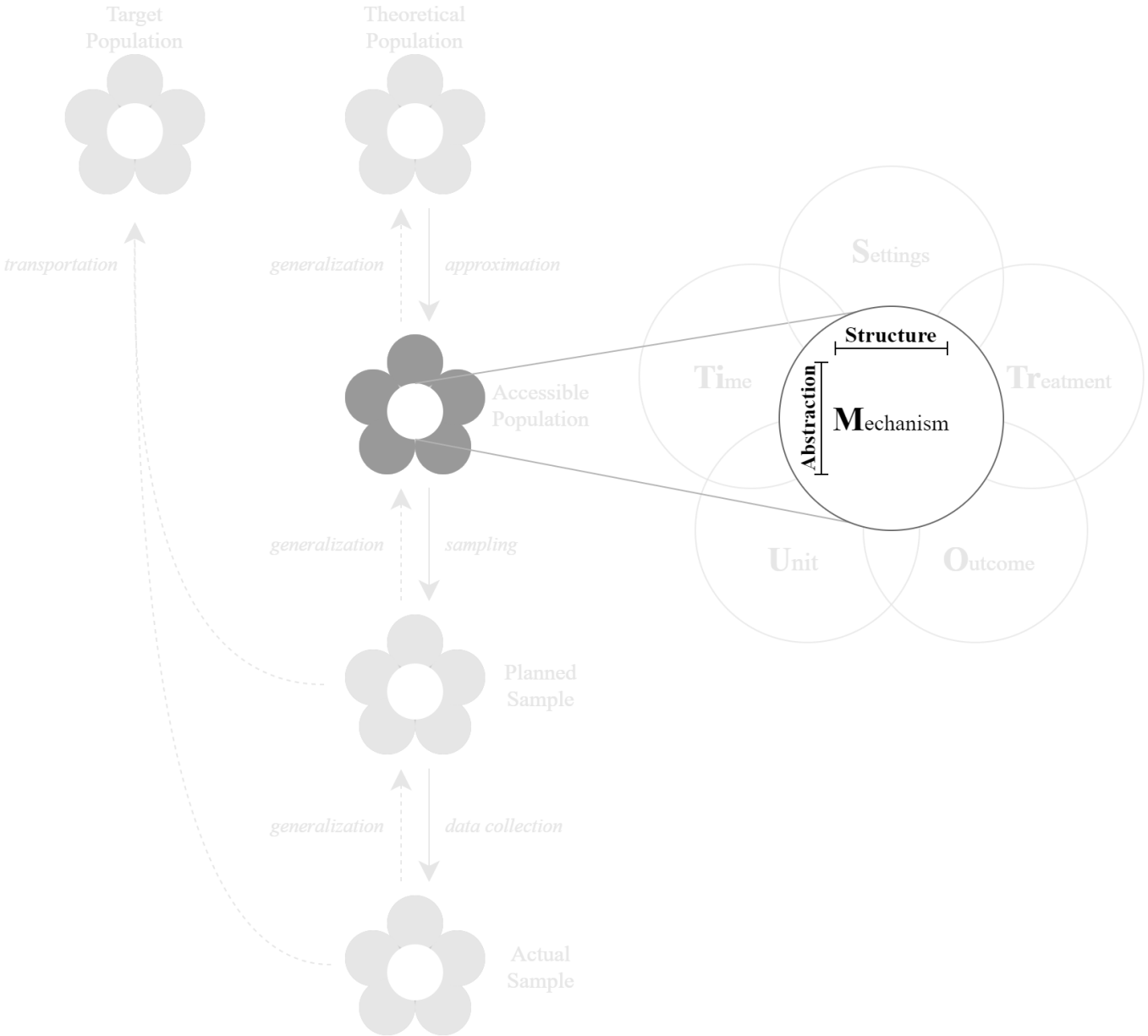
ALL EMPIRICAL METHODS
- Causal structure bias, construct validity bias, publication bias, WEIRD sample bias
Experiments
Field Experiments
- Attrition, noncompliance, site/case selection bias, spillover, weak treatments
Lab Experiments
- Attitude crystallization, ecological validity, experimental realism, Hawthorne effects, John Henry effects, mundane realism, noncompliance, social desirability bias, spillover, student samples
Survey Experiments
- Attrition, demand effects, ecological validity, experimental realism, Hawthorne effects, mundane realism, noncompliance, overly WEIRD samples, social desirability bias, spillover
Natural Experiments
Standard Natural Experiments
- Case-specific (*sui generis*)
Instrumental Variables
- Case-specific, compliers undefined (sometimes), exclusion restriction untestable
Regression Discontinuity
- Case-specific, confounding, estimand challenges (in RD variants), weighting techniques

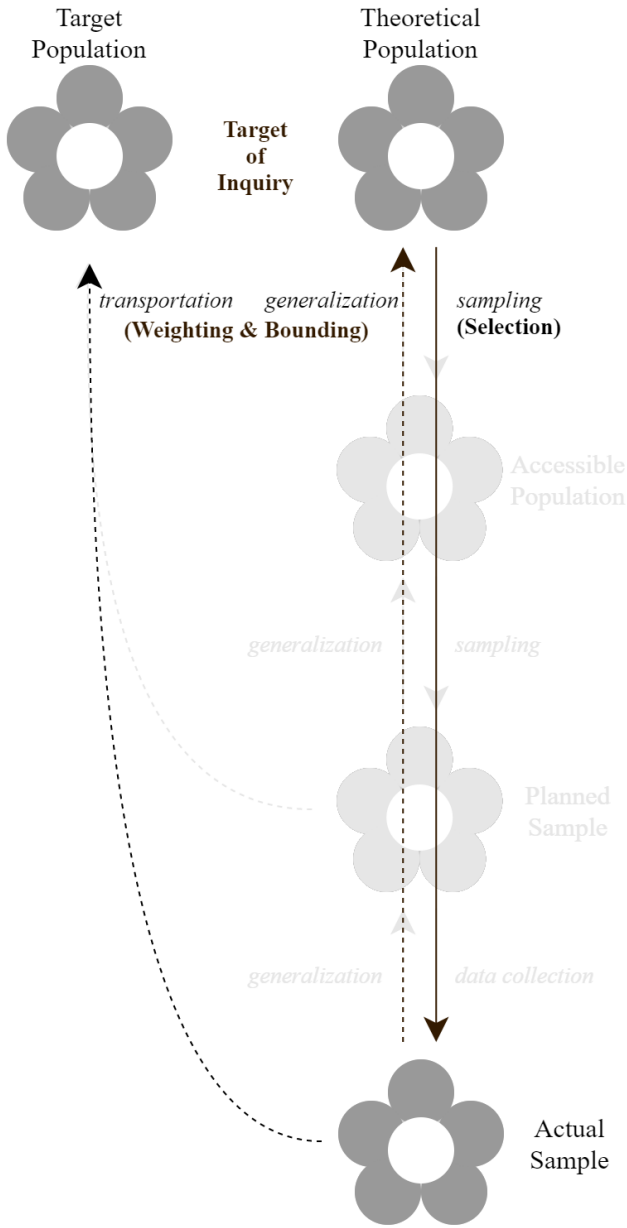

Theoretical Population
Mechanism
Setting
Treatment
Outcome
Unit
Time
selection
generalization
Target Population
sampling
generalization

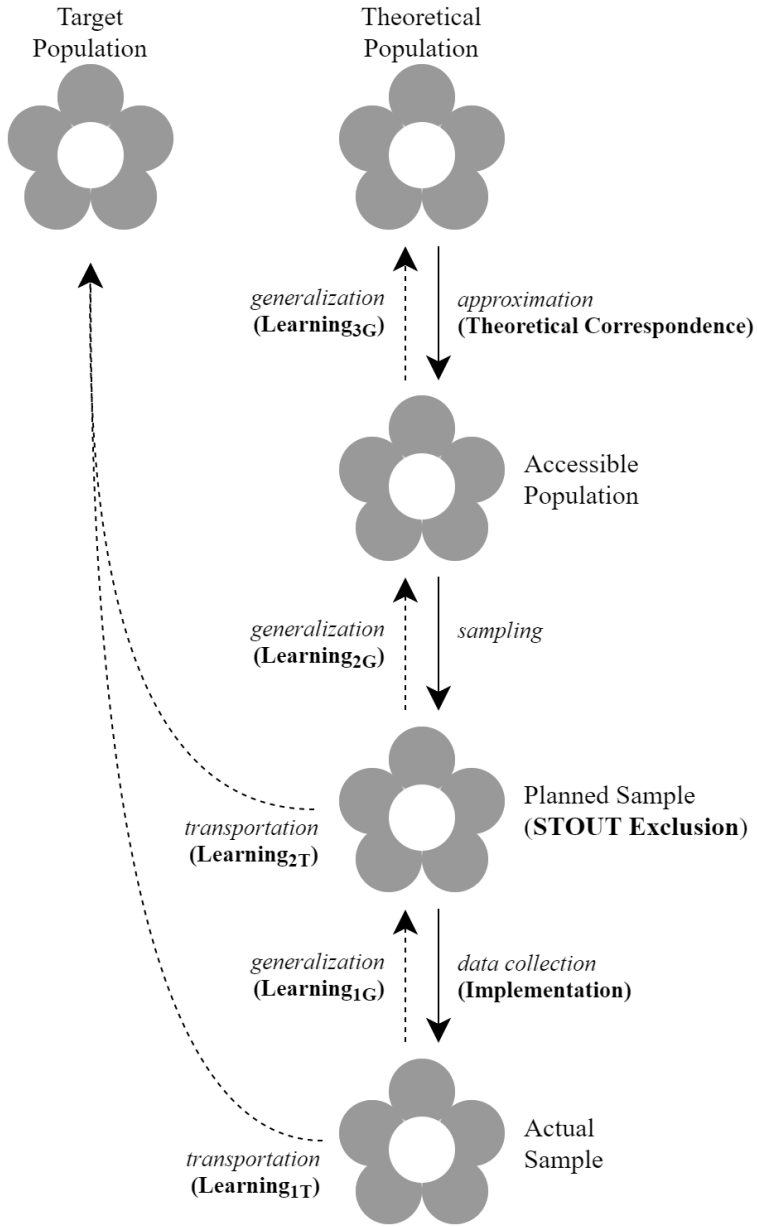

Target
Population
Theoretical
Population
transportation
generalization
approximation
Accessible
Population
generalization
sampling
Planned
Sample
generalization
data collection
Actual
Sample

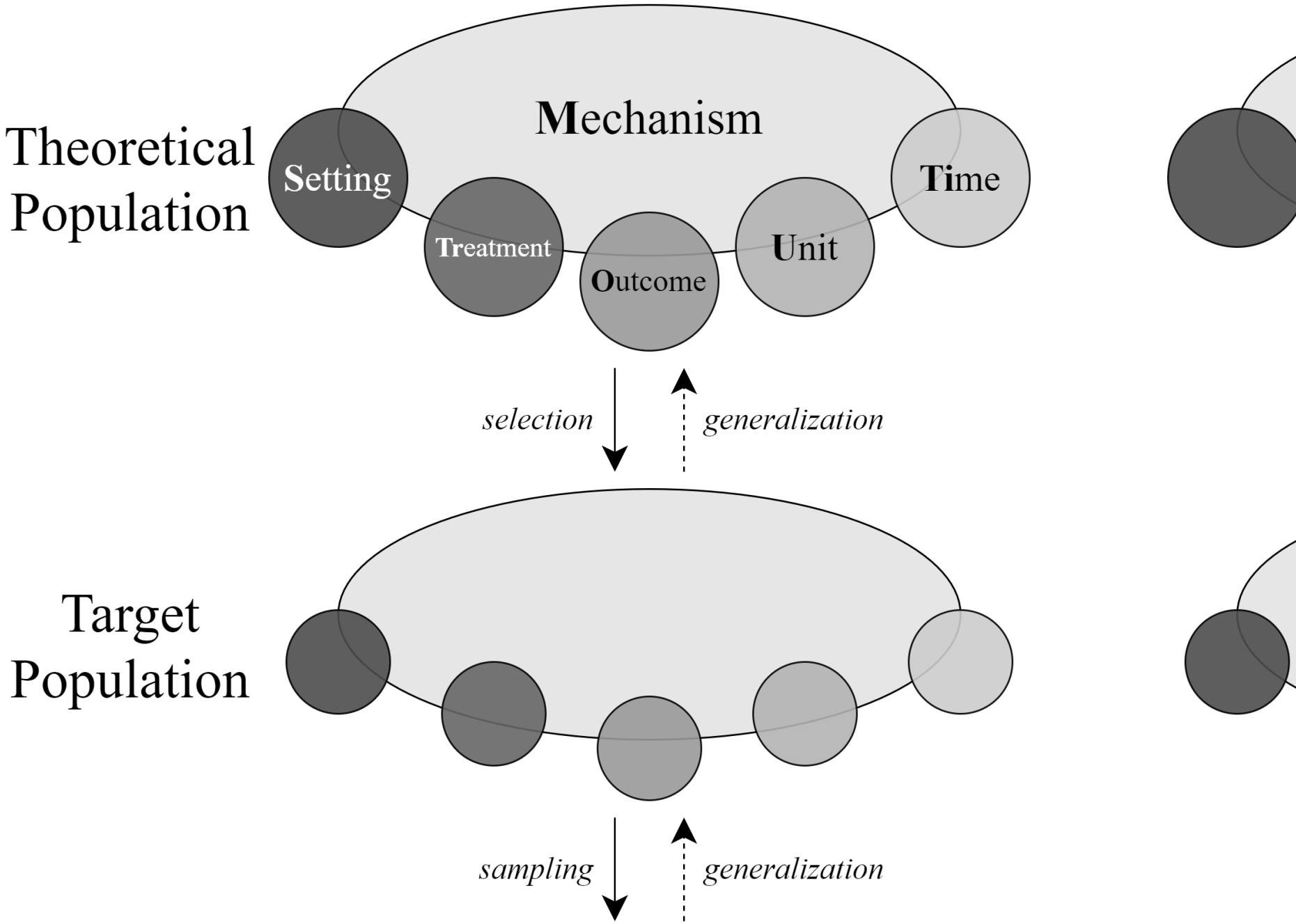

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Actual Sample
Settings
Time
Treatment
Mechanism
Unit
Outcome

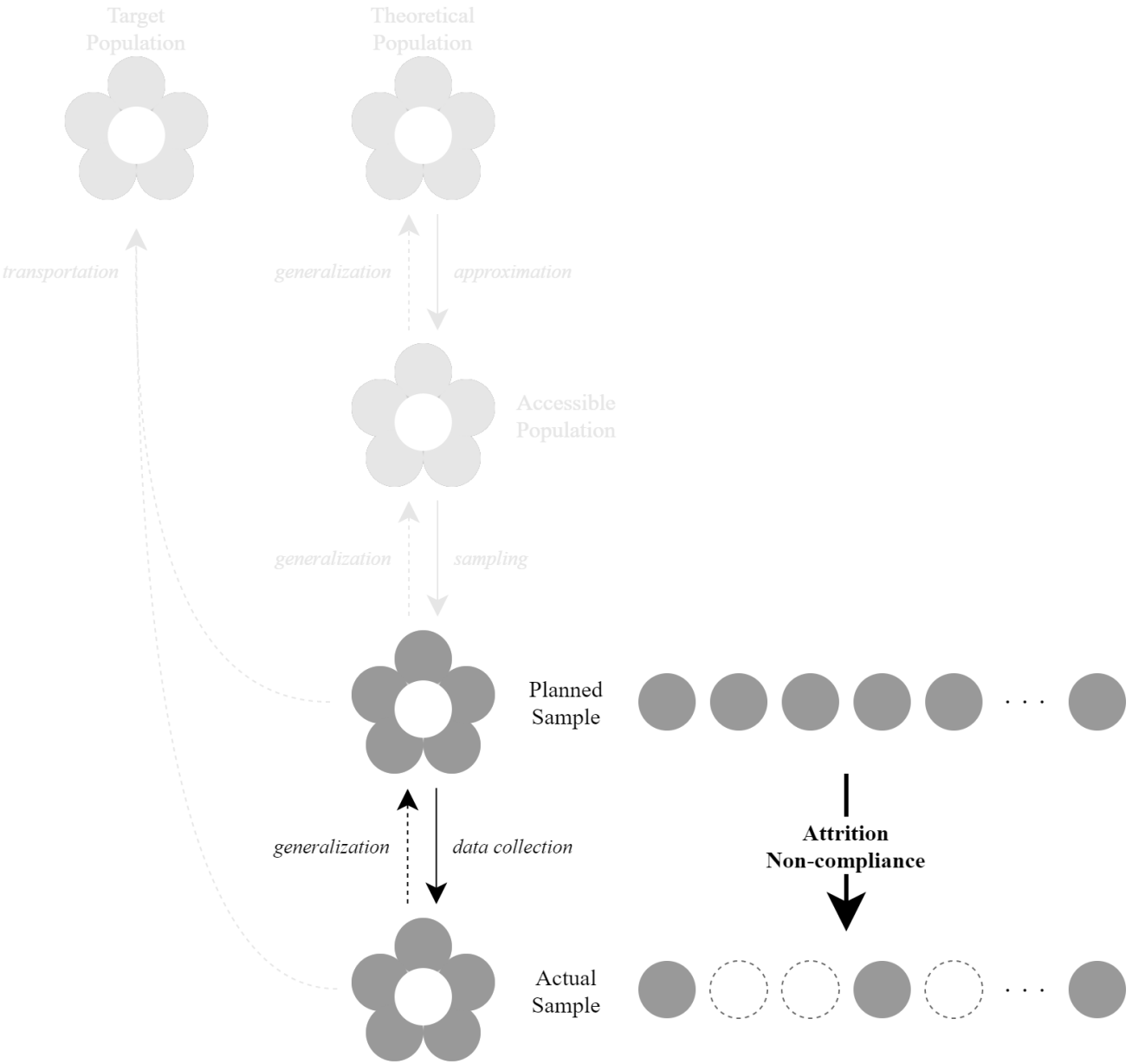

Target Population
Theoretical Population
transportation (TATE)
generalization (PATE)
EV Bias
Sample (SATE)
S-Bias
Settings
Time
Treatment
Mechanism
Unit
Outcome
UTi-Bias
TrO-Bias

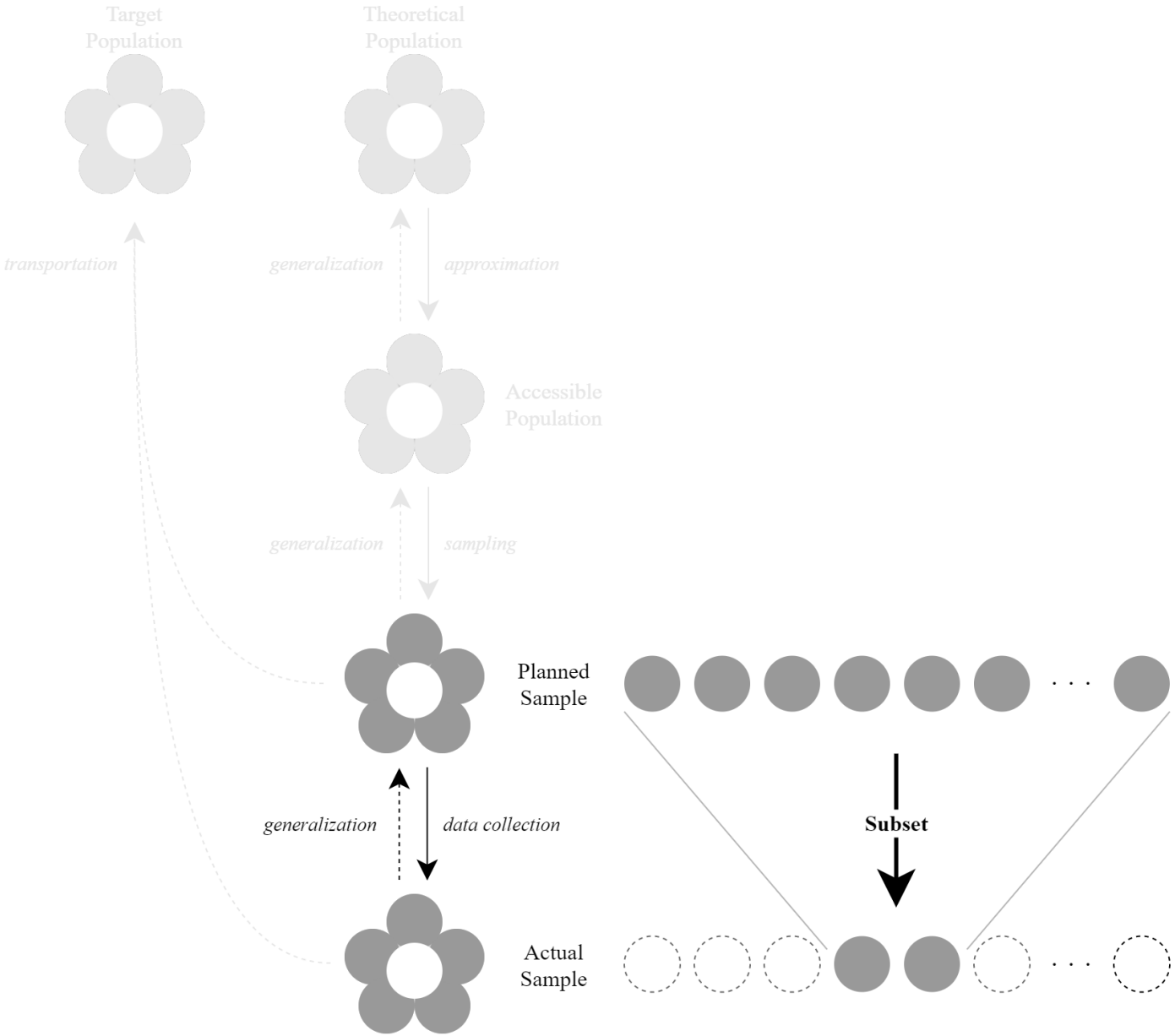

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
Settings
Time
Treatment
Structure
Abstraction
Mechanism
Unit
Outcome
generalization
sampling
Planned Sample
generalization
data collection
Actual Sample

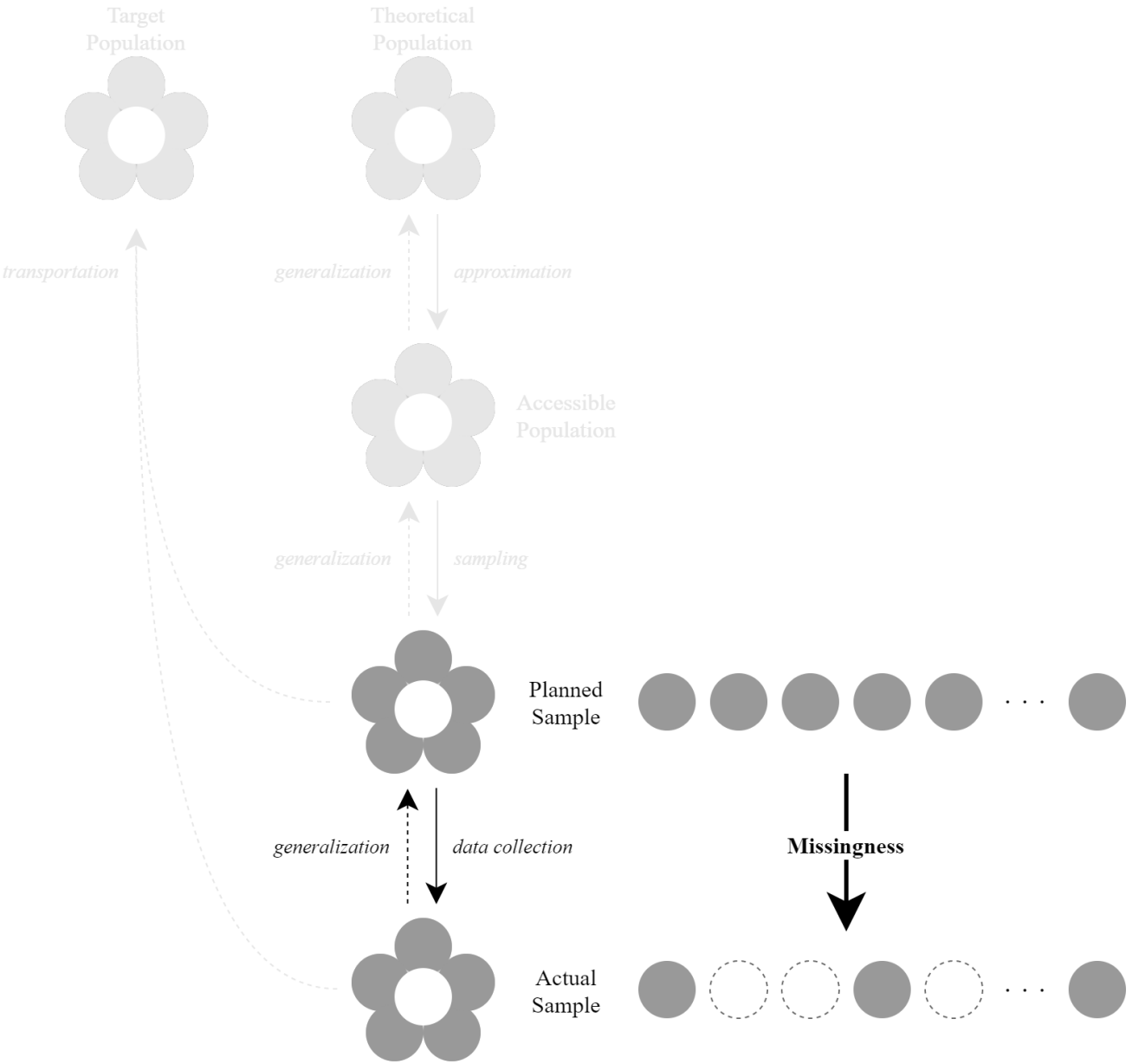

Target Population
Theoretical Population
Target of Inquiry
transportation
(Weighting & Bounding)
generalization
sampling
(Selection)
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Actual Sample

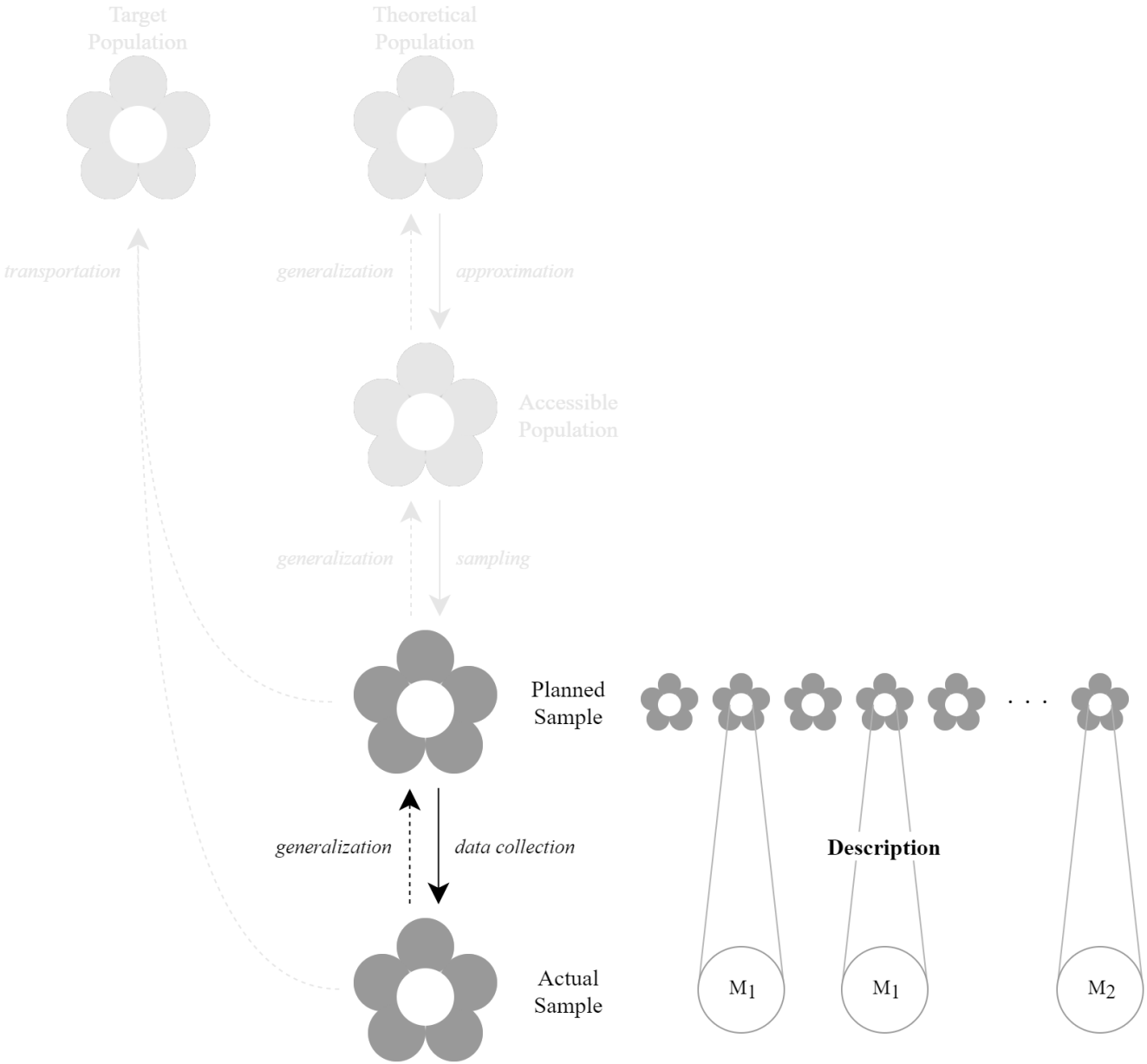

Target Population
Theoretical Population
generalization (Learning3G)
approximation (Theoretical Correspondence)
Accessible Population
generalization (Learning2G)
sampling
transportation (Learning2T)
Planned Sample (STOUT Exclusion)
generalization (Learning1G)
data collection (Implementation)
transportation (Learning1T)
Actual Sample

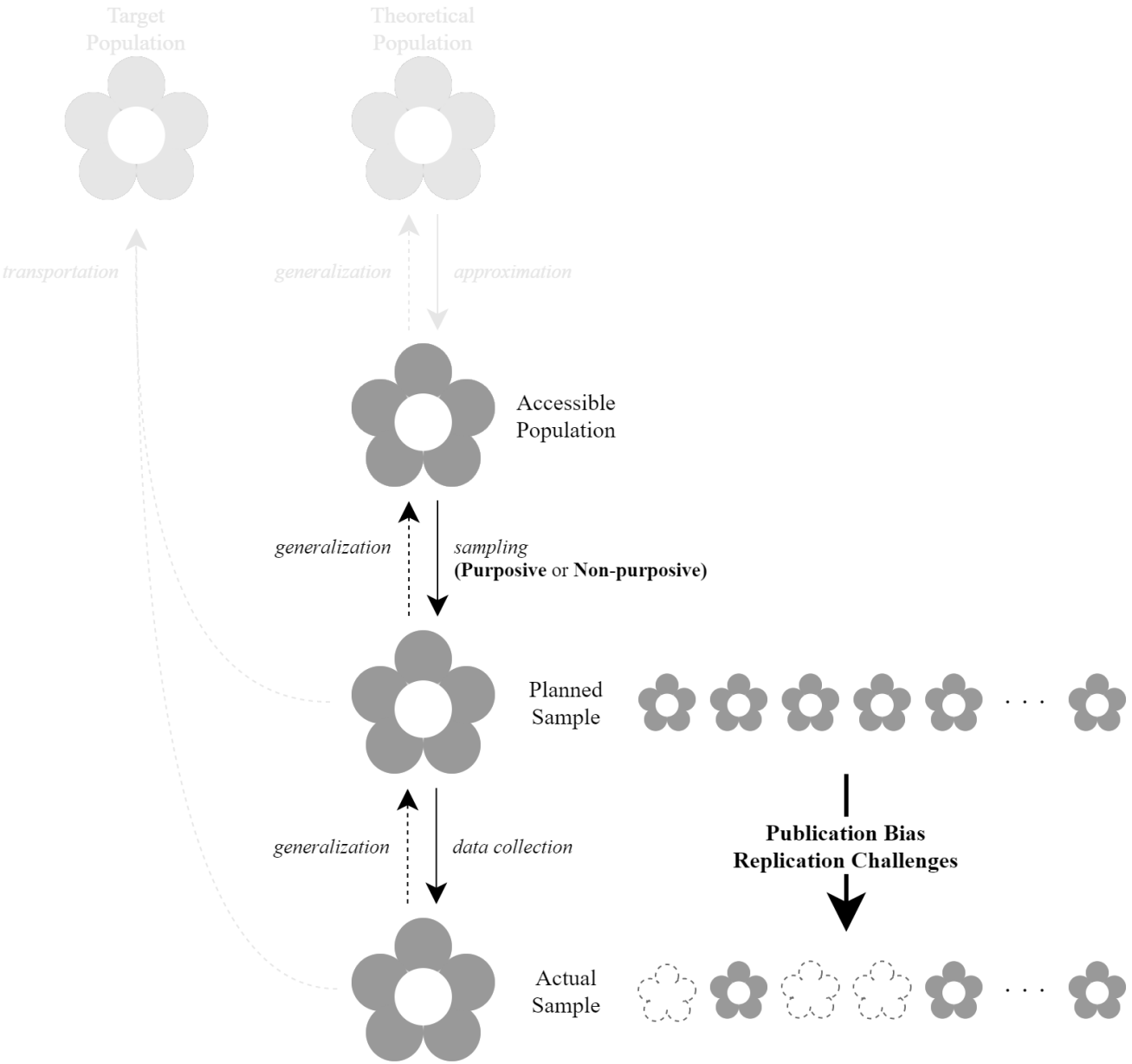

Theoretical Population
Mechanism
Setting
Treatment
Outcome
Unit
Time
selection
generalization
Target Population
sampling
generalization

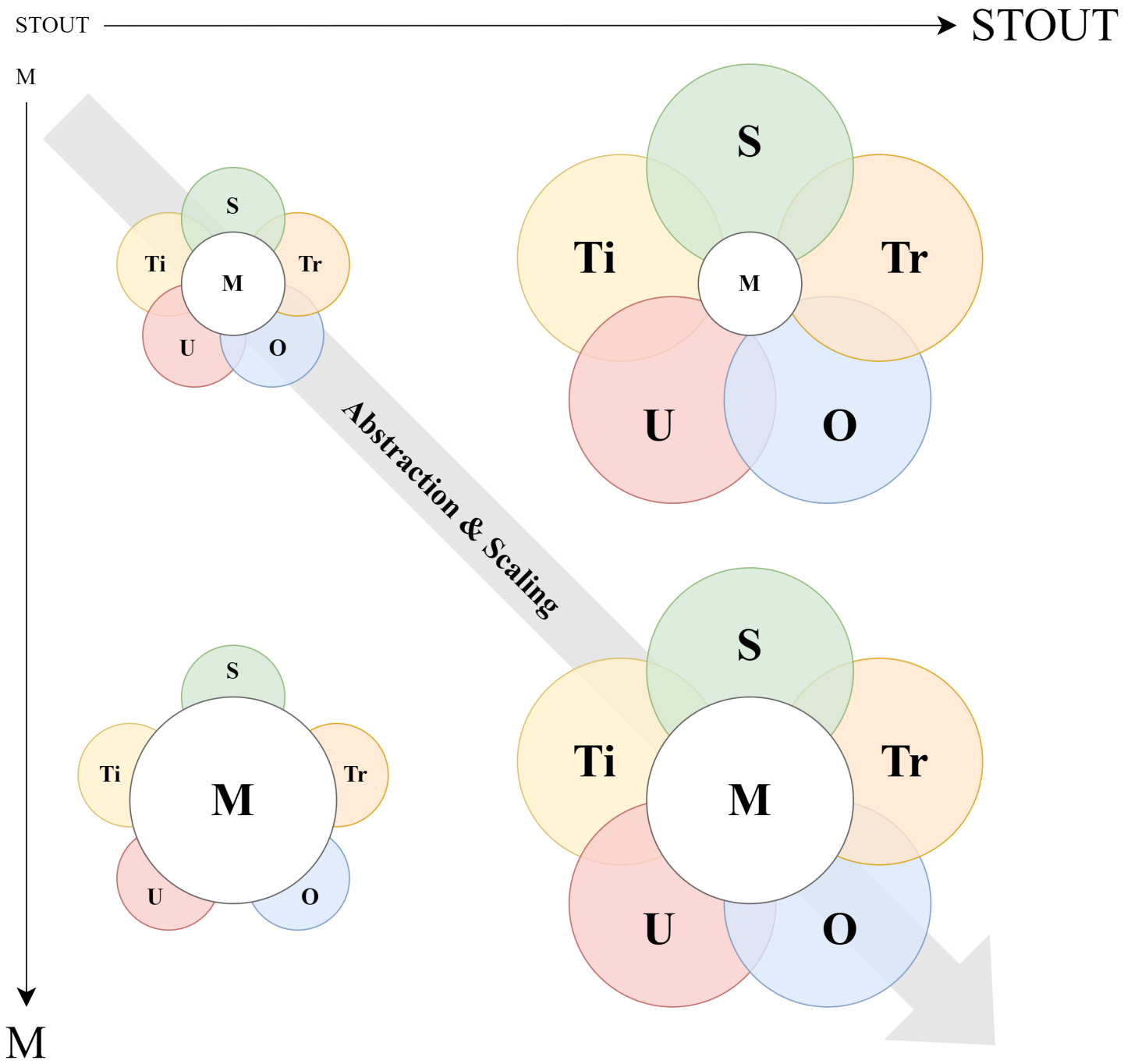

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Attrition
Non-compliance
Actual Sample

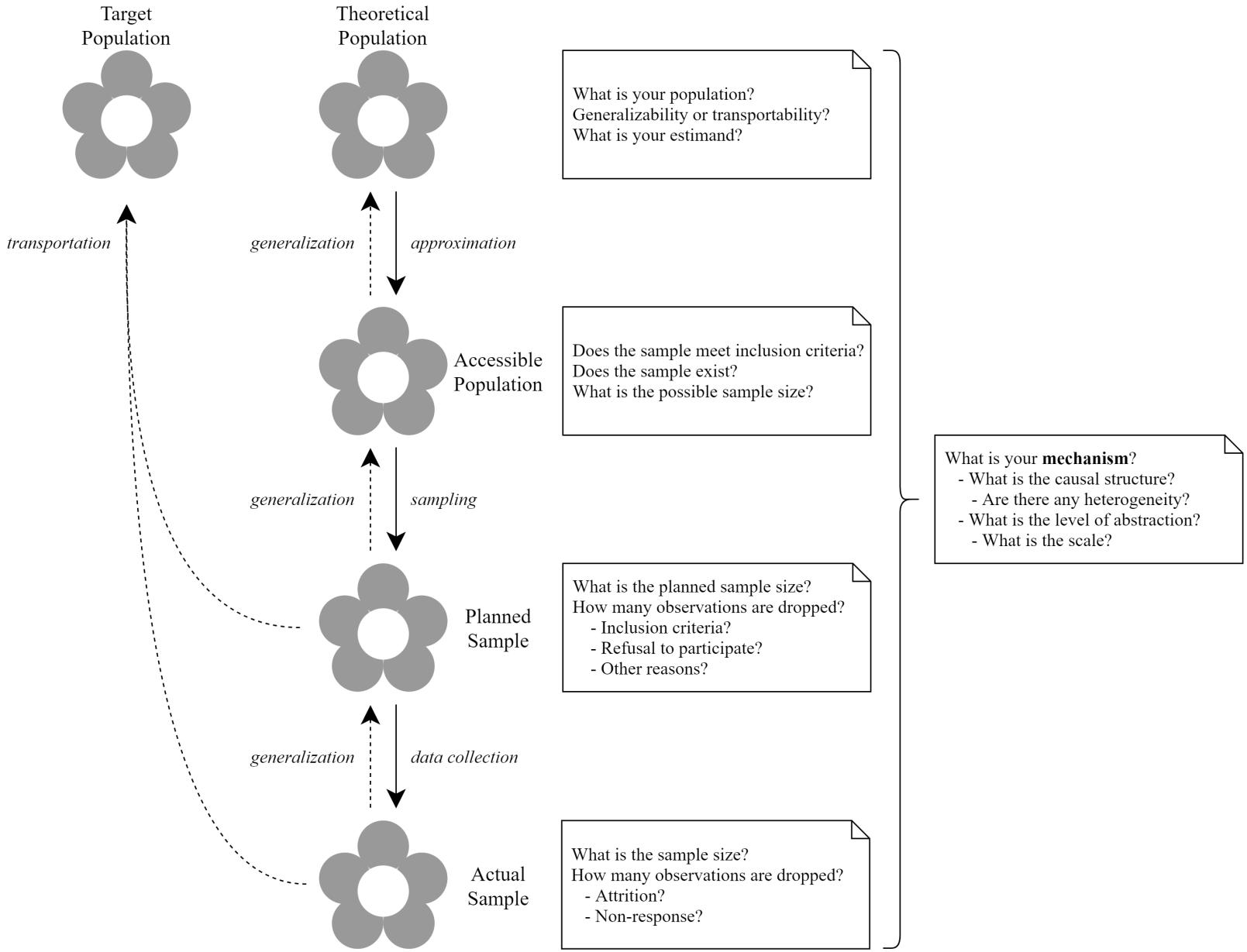

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Subset
Actual Sample

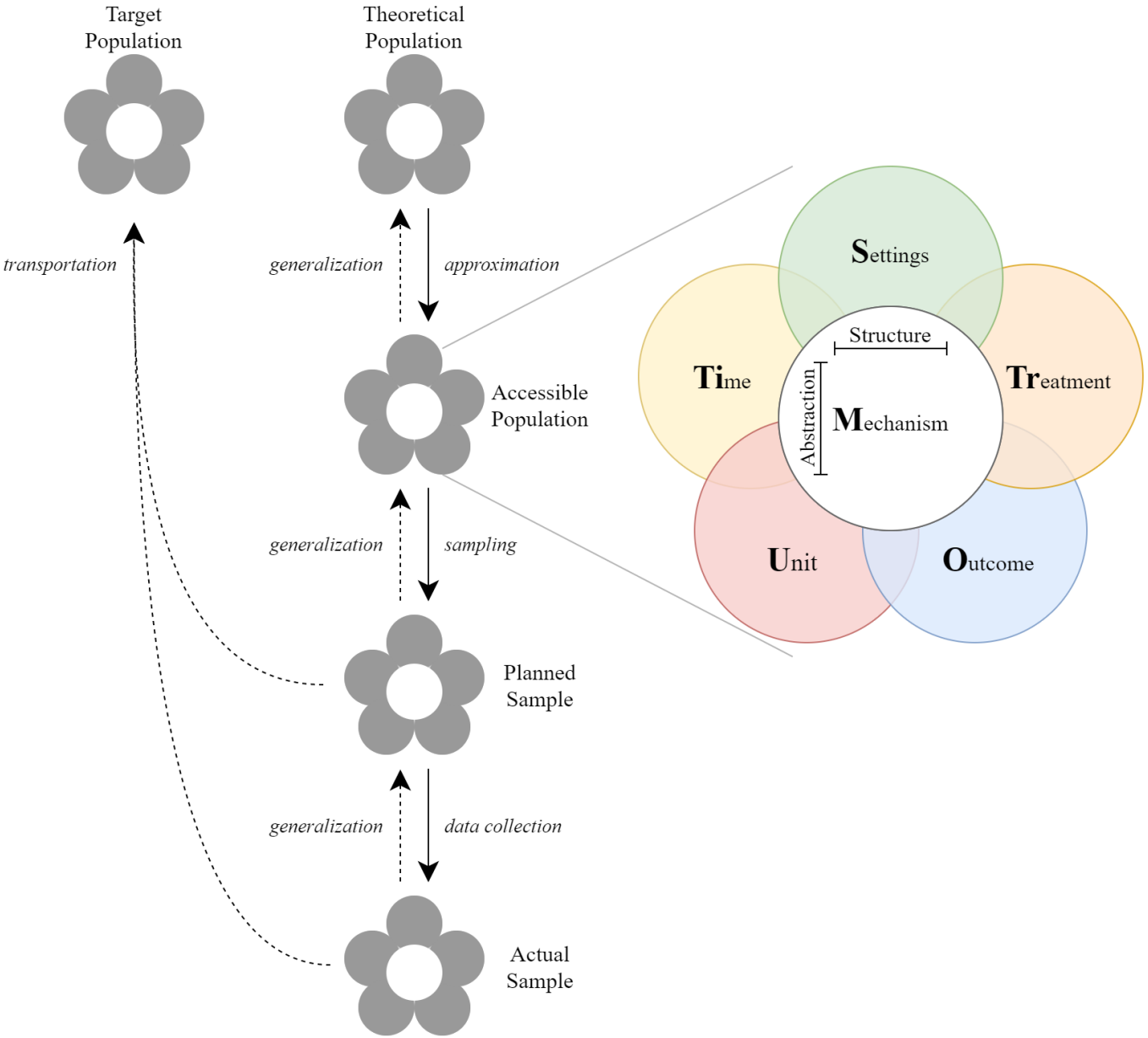

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Missingness
Actual Sample

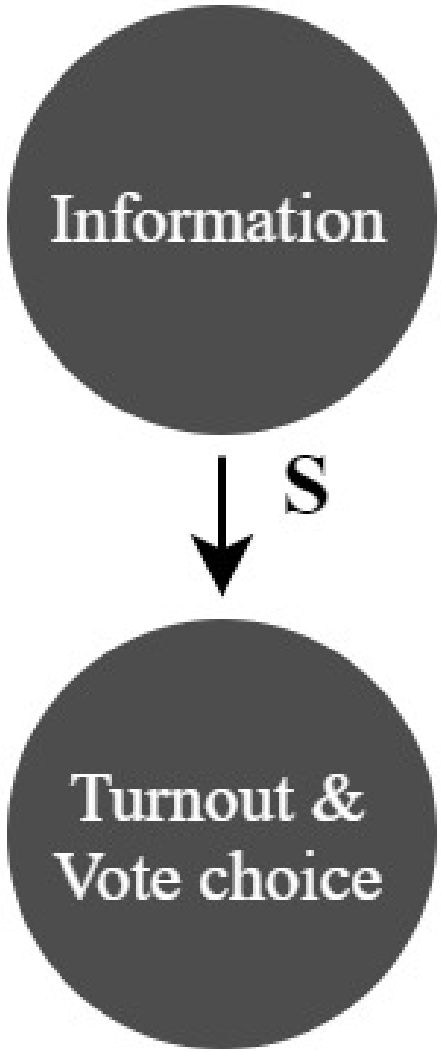

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Actual Sample
Description
M1
M1
M2

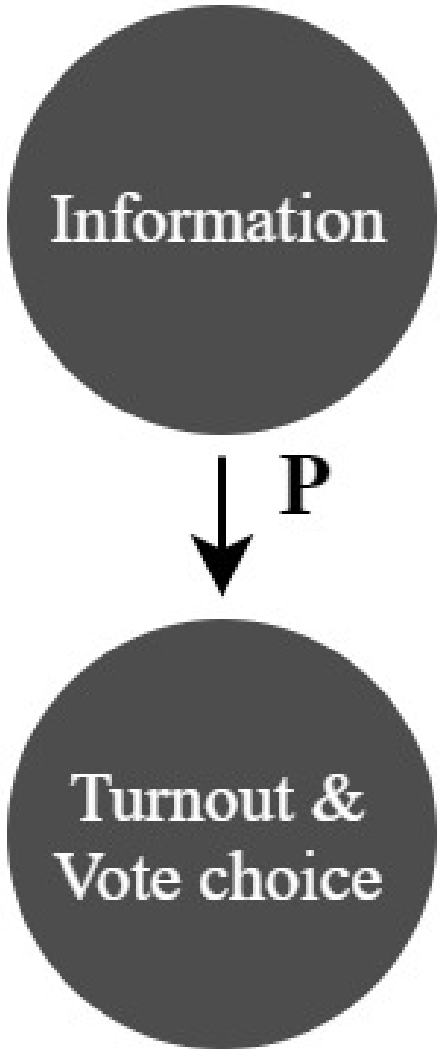

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
(Purposive or Non-purposive)
Planned Sample
generalization
data collection
Publication Bias
Replication Challenges
Actual Sample

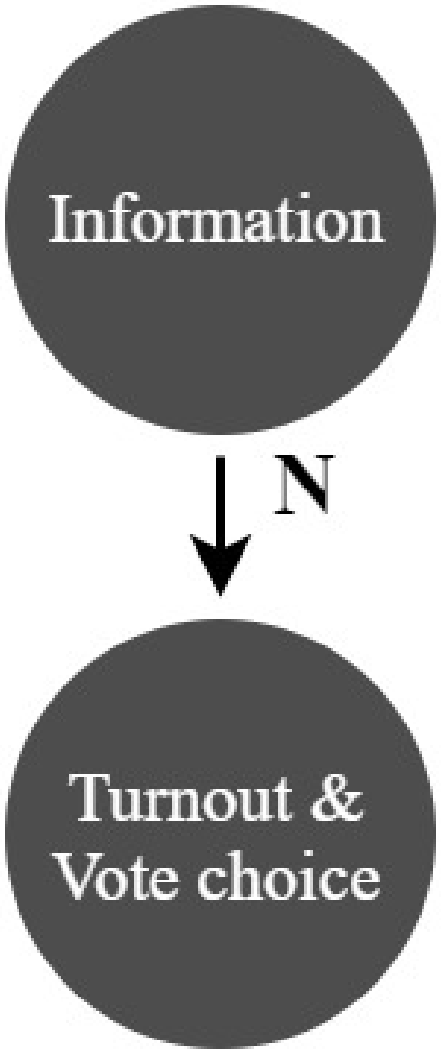

STOUT
STOUT
M
M
Abstraction & Scaling
S
Ti
Tr
M
U
O
S
Ti
Tr
M
U
O
S
Ti
Tr
M
U
O
S
Ti
Tr
M
U
O


Target Population
Theoretical Population
What is your population?
Generalizability or transportability?
What is your estimand?
transportation
generalization
approximation
Accessible Population
Does the sample meet inclusion criteria?
Does the sample exist?
What is the possible sample size?
generalization
sampling
Planned Sample
What is the planned sample size?
How many observations are dropped?
- Inclusion criteria?
- Refusal to participate?
- Other reasons?
generalization
data collection
Actual Sample
What is the sample size?
How many observations are dropped?
- Attrition?
- Non-response?
What is your **mechanism**?
- What is the causal structure?
- Are there any heterogeneity?
- What is the level of abstraction?
- What is the scale?

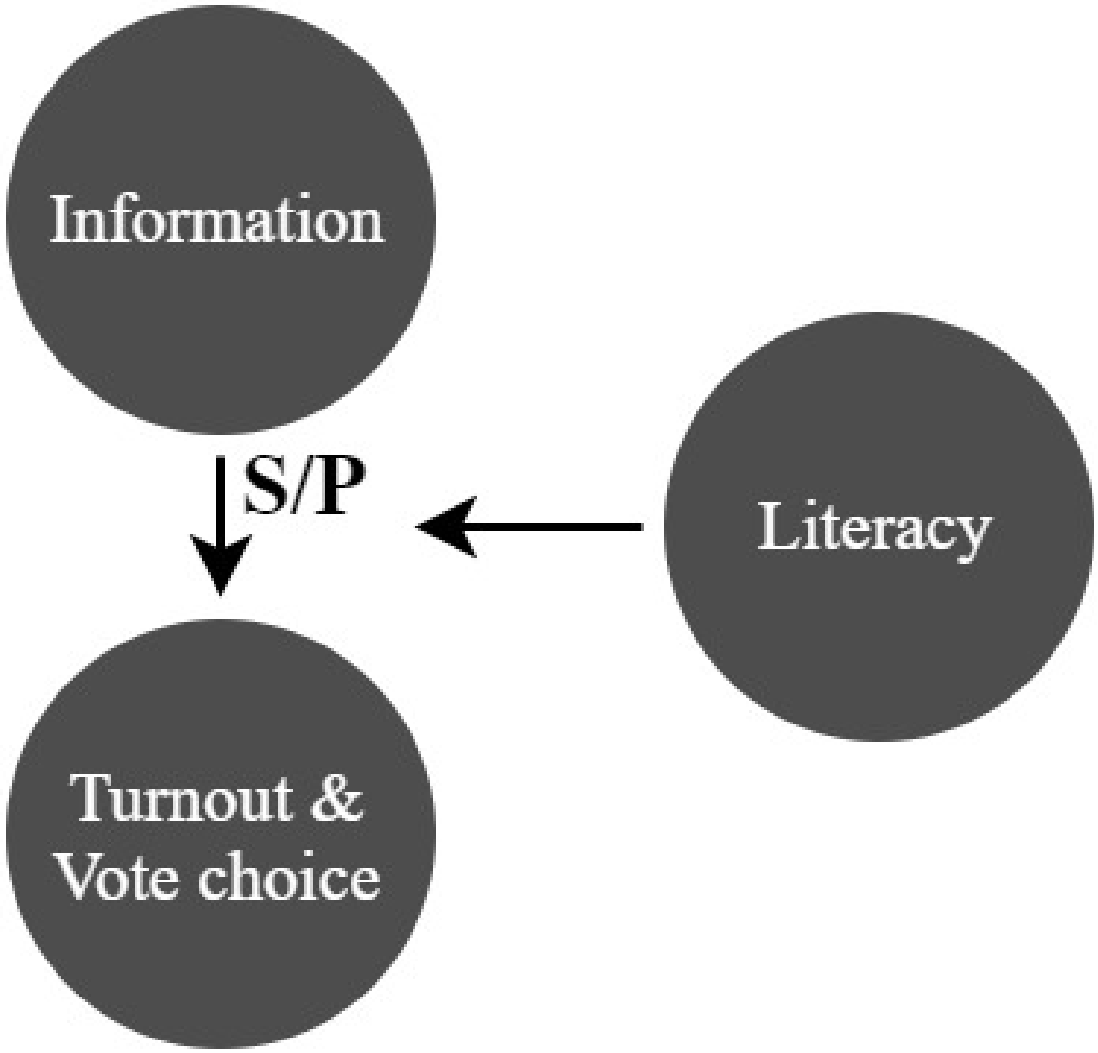

Target Population
Theoretical Population
transportation
generalization
approximation
Accessible Population
generalization
sampling
Planned Sample
generalization
data collection
Actual Sample
Settings
Time
Treatment
Structure
Abstraction
Mechanism
Unit
Outcome

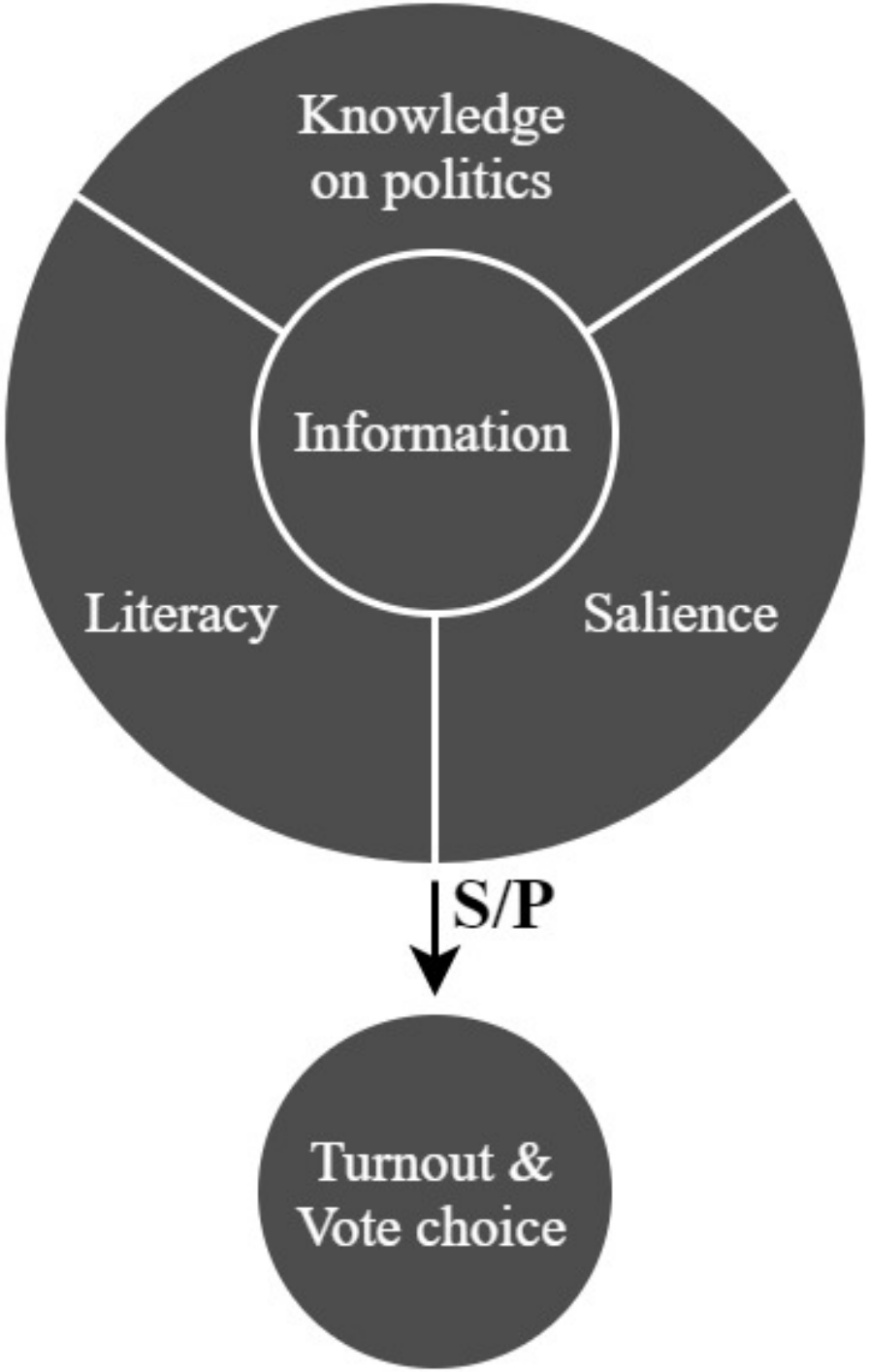

Information
S
Turnout &
Vote choice

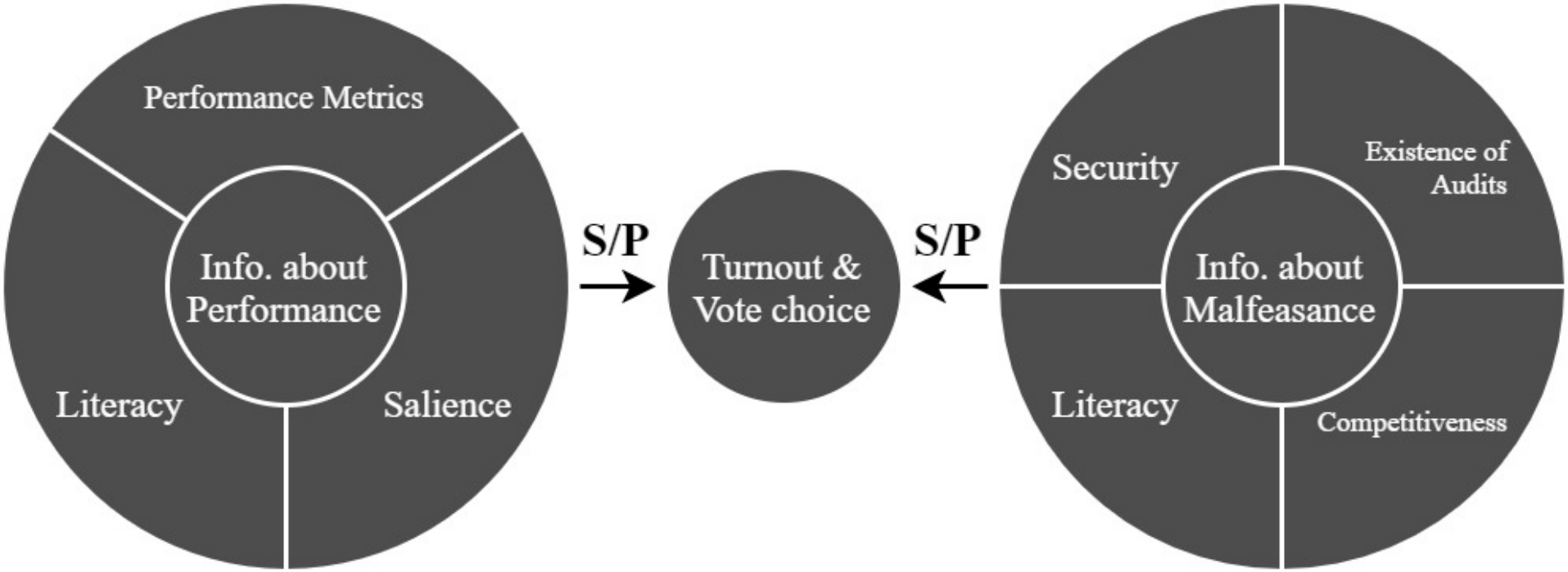

Information
P
Turnout &
Vote choice

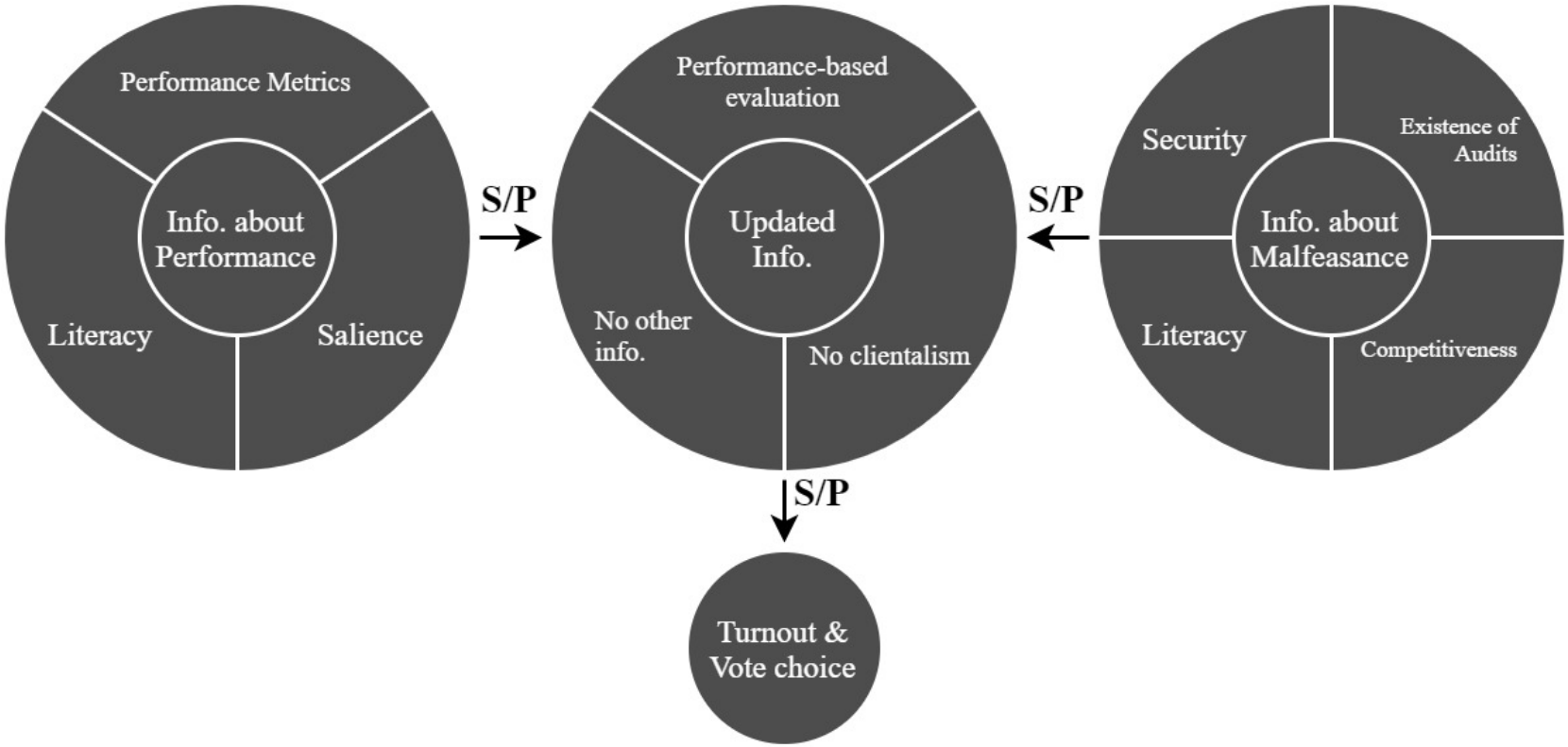

Information
N
Turnout &
Vote choice

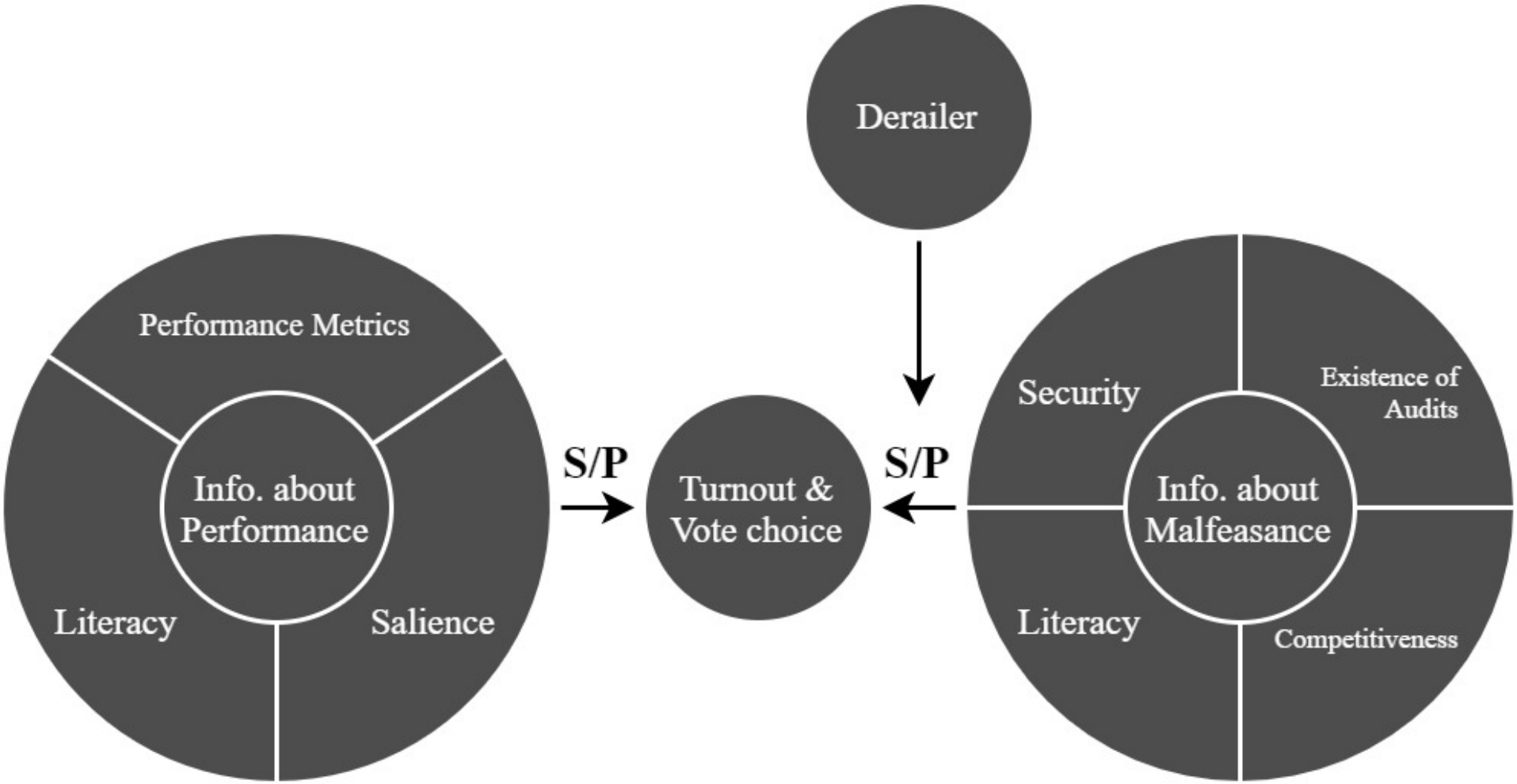

Assigned Info.
S/P
Updated Info.
S/P
Turnout & Vote choice

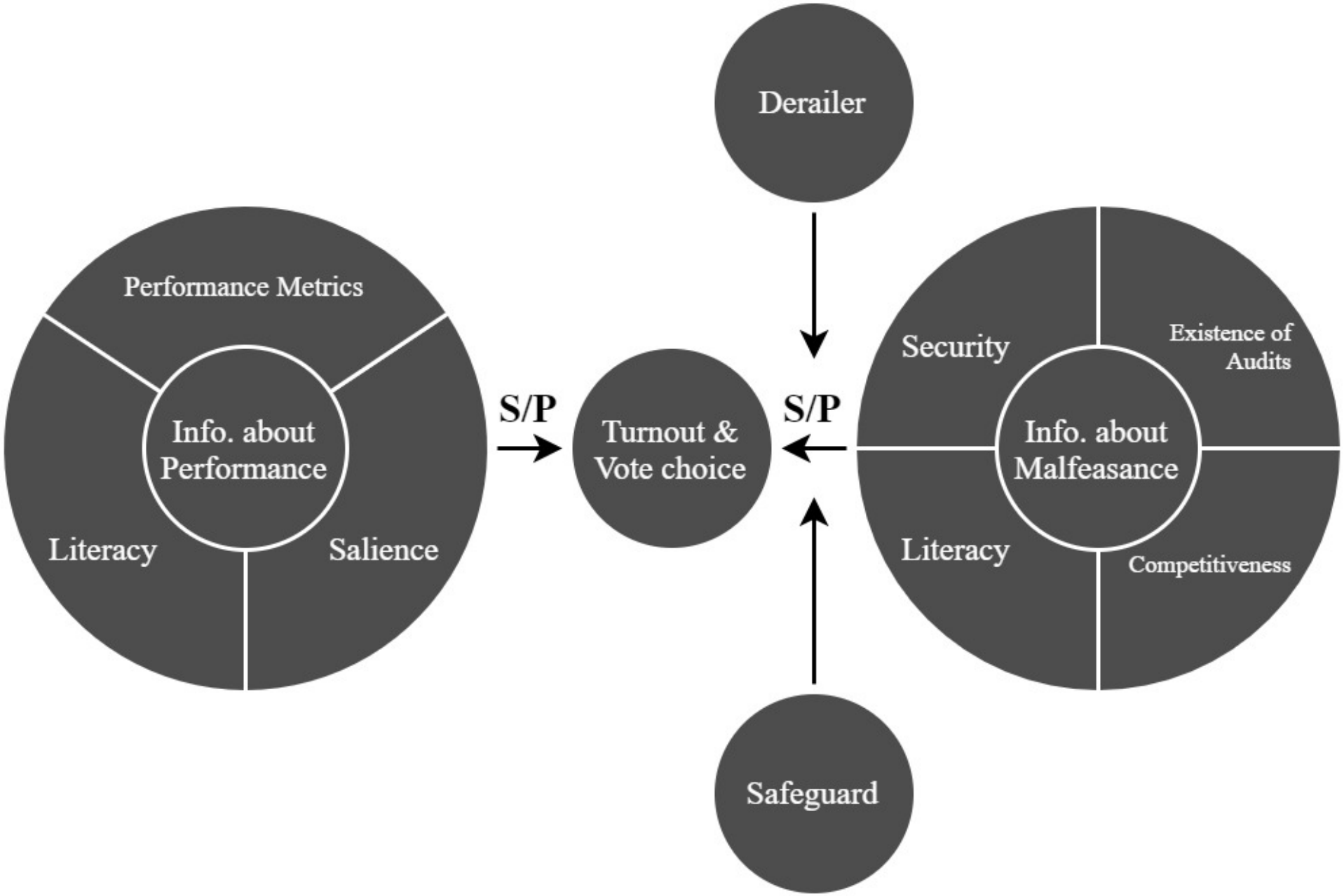

Information
S/P
Literacy
Turnout &
Vote choice


Knowledge
on politics
Information
Literacy
Salience
S/P
Turnout &
Vote choice


Performance Metrics
Info. about Performance
Literacy
Salience
S/P
Turnout & Vote choice
S/P
Security
Existence of Audits
Info. about Malfeasance
Literacy
Competitiveness


Performance Metrics
Info. about Performance
Literacy
Salience
S/P
Performance-based evaluation
Updated Info.
No other info.
No clientalism
S/P
Security
Existence of Audits
Info. about Malfeasance
Literacy
Competitiveness
S/P
Turnout & Vote choice


Derailer
Performance Metrics
Info. about Performance
Literacy
Salience
S/P
Turnout & Vote choice
S/P
Security
Existence of Audits
Info. about Malfeasance
Literacy
Competitiveness


Derailer
Performance Metrics
Info. about Performance
Literacy
Salience
S/P
Turnout & Vote choice
S/P
Security
Existence of Audits
Info. about Malfeasance
Literacy
Competitiveness
Safeguard

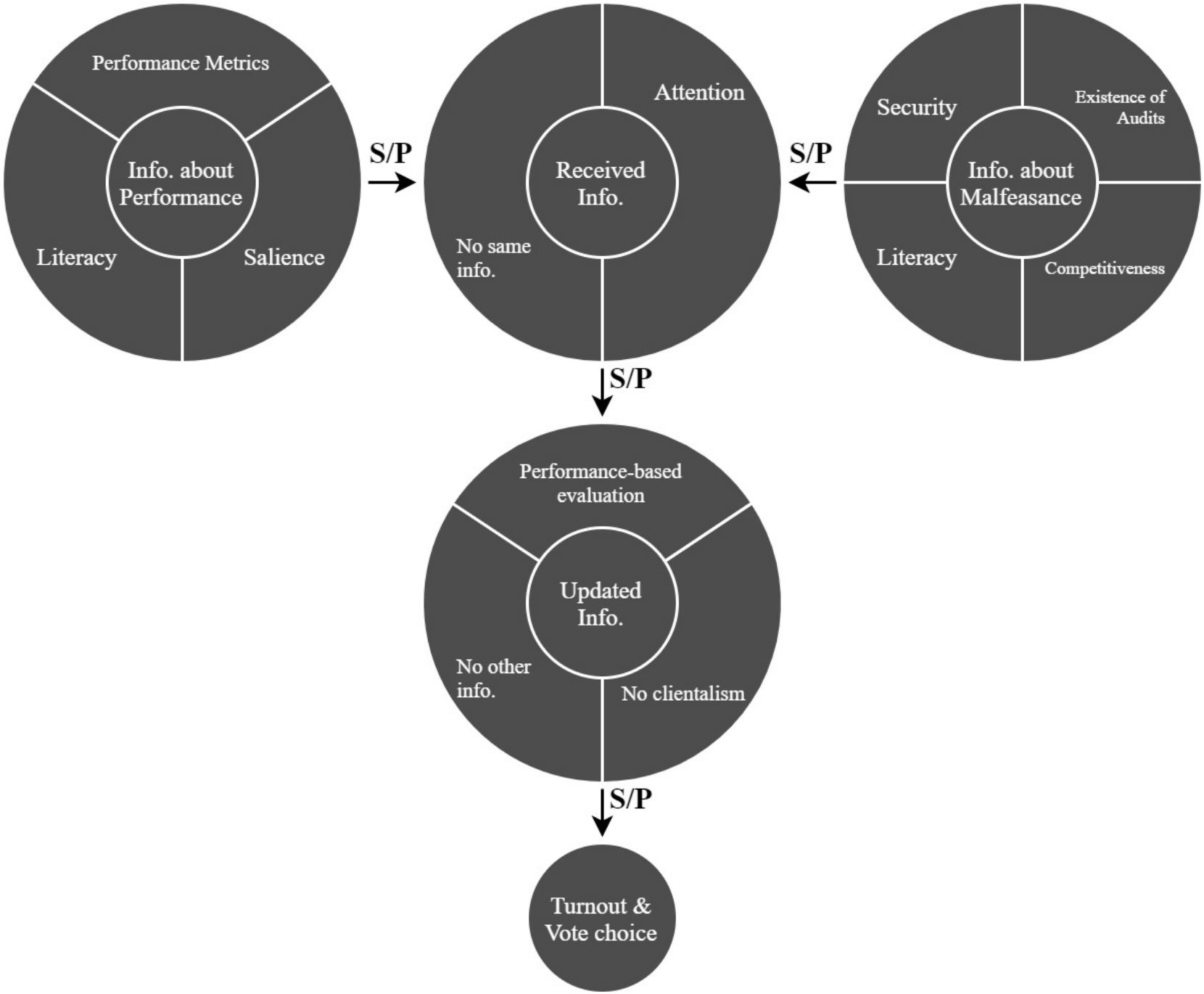

Performance Metrics
Info. about Performance
Literacy
Salience
S/P
Attention
Received Info.
No same info.
S/P
Security
Existence of Audits
Info. about Malfeasance
Literacy
Competitiveness
S/P
Performance-based evaluation
Updated Info.
No other info.
No clientalism
S/P
Turnout & Vote choice

TABLE 2.1 *The Metaketa Initiative: Extant challenges and pillars*

	Extant challenges		Pillars of the Metaketa Initiative
1.	Confounding in observational research	1.	Randomized controlled trials
2.	Limited external validity of single RCTs	2.	Multiple studies in diverse contexts
3.	Heterogeneous, scattered findings	3.	Meta-analysis with overall finding
4.	Diversity of interventions	4.	"Common arm" intervention
5.	Noncomparable measures, impeding aggregation	5.	Harmonized measurement of inputs, outcomes, and controls
6.	Researcher incentives for innovation over replication	6.	"Alternative arm" intervention
7.	Private data	7.	Open data and replication code
8.	Errors in data or code	8.	Third-party data analysis
9.	Fishing (data mining, specification searching, multiple hypotheses)	9.	Pre-analysis plans with limited number of specified hypotheses
10.	Publication bias	10.	Publication of all registered analyses

TABLE 3.1 *Metaketa experiments: Common and alternative intervention arms*

Study	Focus of Common Informational Treatment	Focus of Alternative Treatment Arm(s)
Benin (Chapter 4)	Legislative performance (relative to department and national averages; high-dosage)	Civics lesson on importance of legislative performance; public provision of info.; low or high dosage
Brazil (Chapter 9)	Accounting irregularities (acceptance or rejection of municipal accounts by auditors)	Municipal education outcomes (ranking of municipalities)
Burkina Faso (Chapter 8)	Quality of municipal services provided by previous incumbent party (relative to other municipalities)	Invitation to participate in municipal government meetings
India (Chapter 10)	Criminal backgrounds of politicians (info. provided by survey enumerators)	Criminal backgrounds of politicians (info. provided by local intermediaries)
Mexico (Chapter 5)	Unauthorized/misallocated spending (relative to opposition municipalities in the same state)	Unauthorized/ misallocated spending, publicly provided via loudspeakers; or not benchmarked
Uganda 1 (Chapter 6)	Voter-candidate policy alignment and candidate characteristics (via videos, in general elections)	Public provision of common-arm information (via videos); in primary and general elections
Uganda 2 (Chapter 7)	Budget irregularities (provided over SMS; high and low density)	Quality of service provision (provided over SMS; high and low density)

The common informational treatments provided information on politician performance privately to individual voters in the month prior to an election; the information was disseminated by flyer, text message, or video, depending on the study. These interventions echo both previous experimental treatments in the research literature and those promoted by donor organizations that advocate for transparency.

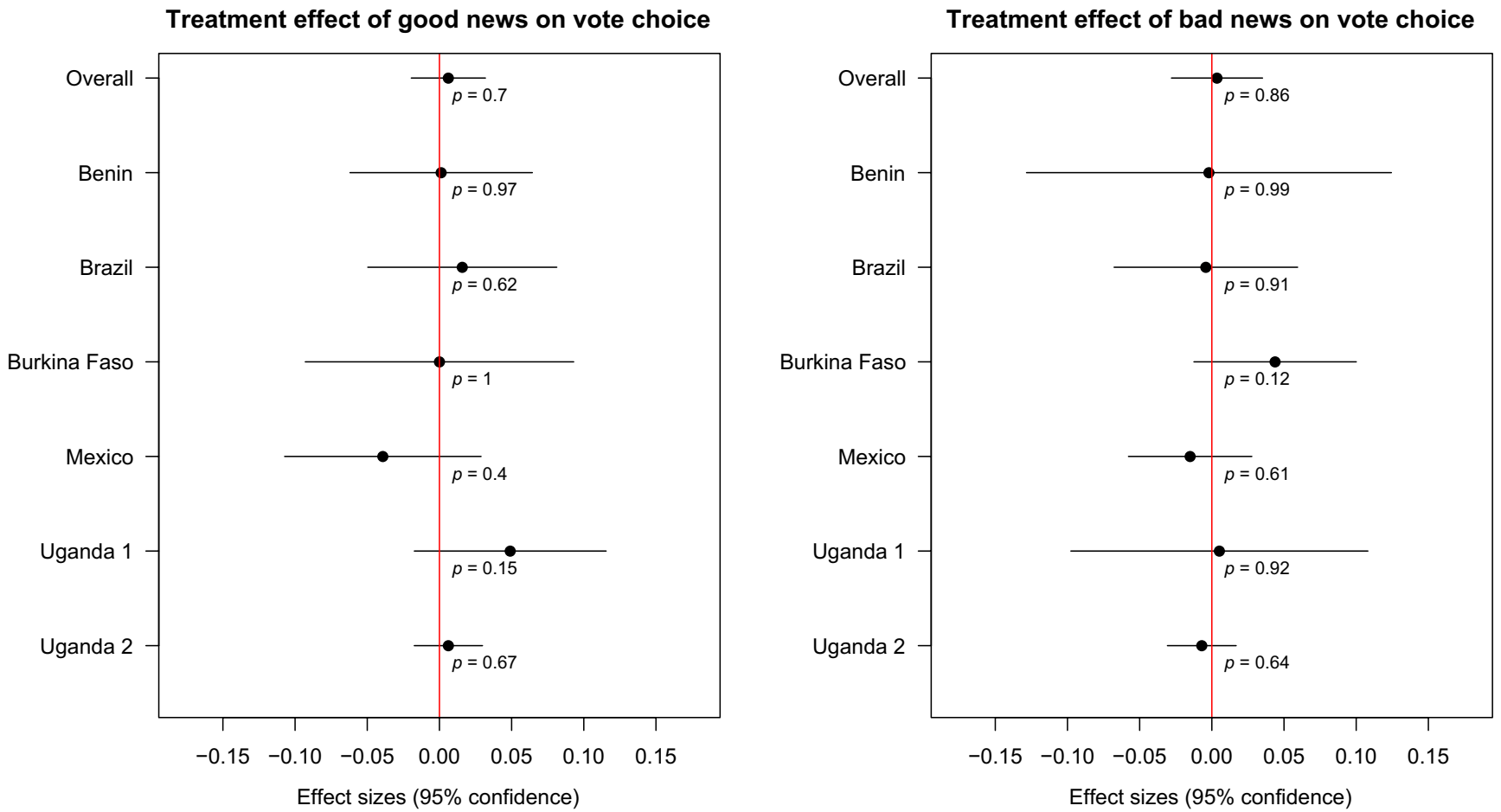


Fig. 2. Meta-analysis: Country-specific effects on vote choice. Estimated change in the proportion of voters who support an incumbent after receiving good news (**left**) or bad news (**right**) about the politician, compared to receiving no information. Unadjusted estimates. For estimating the average of the study-specific effects (top row), each study is weighted by the inverse of its size. Horizontal lines show 95% CIs for the estimated change. Entries under each estimate show *p*-values calculated by randomization inference. In all cases, the differences are close to zero and statistically insignificant.

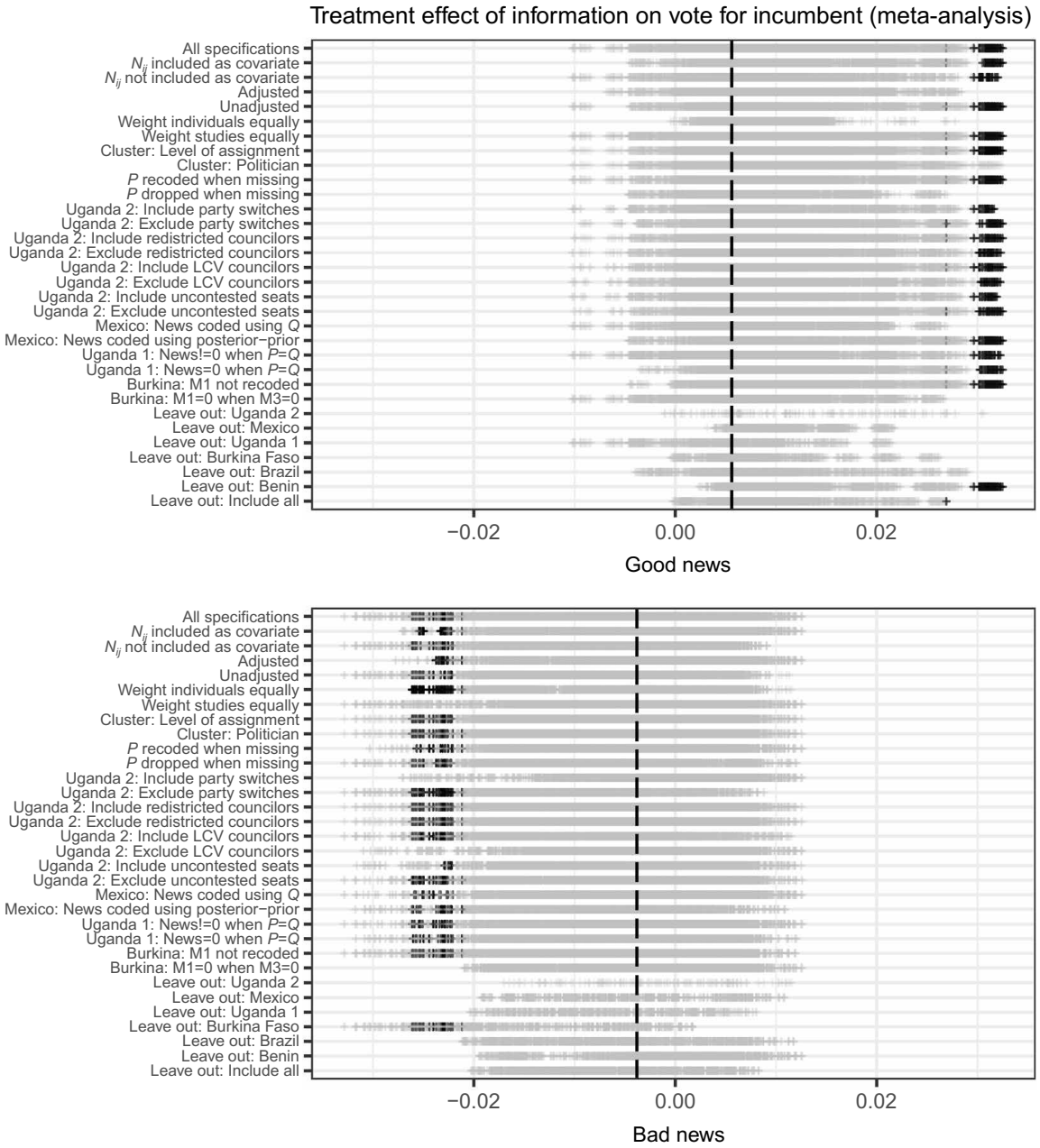
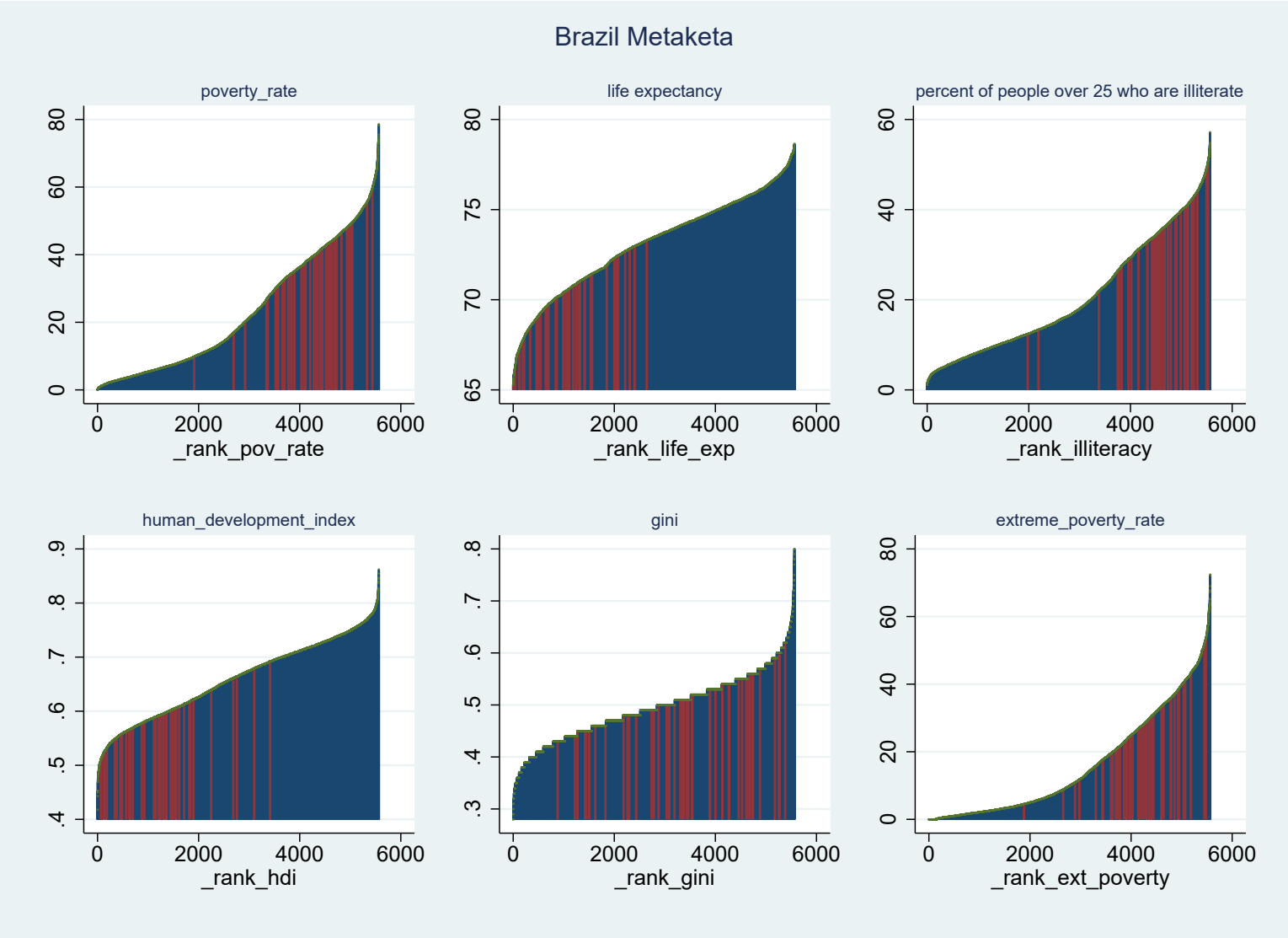
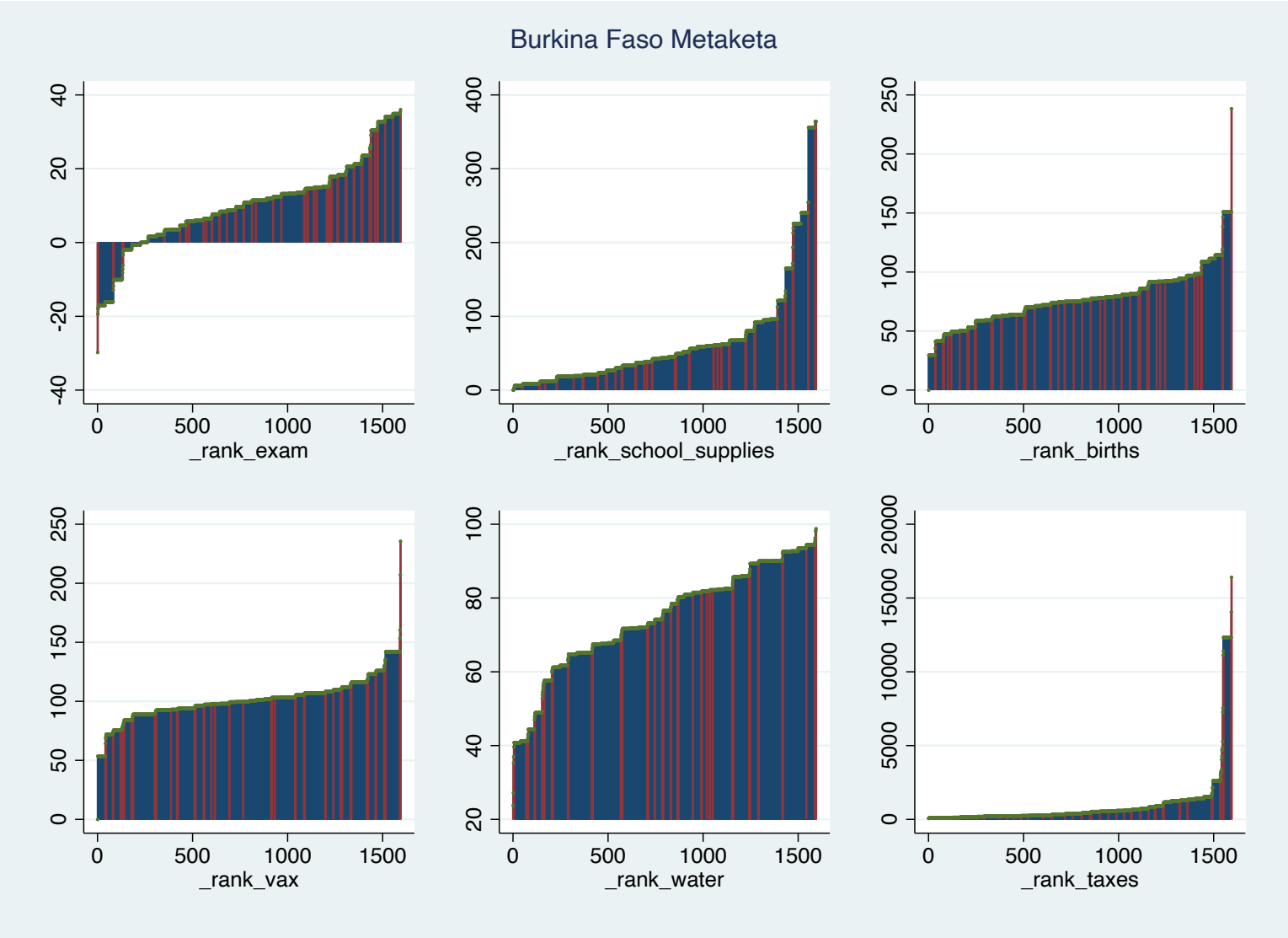


Fig. 4. Robustness of findings across specifications: Vote for incumbent. Estimates across all specifications of the overall treatment effect of the common informational intervention on vote for incumbent. The vertical axis lists all considered specification choices. The top row shows the collection of estimates across all specifications. Each subsequent row holds fixed a given specification choice and shows the distribution of treatment effect estimates, varying all other choices. Darkened vertical lines show estimates for which $p < 0.05$. The dashed vertical line indicates the estimated average treatment effect reported in table S5.

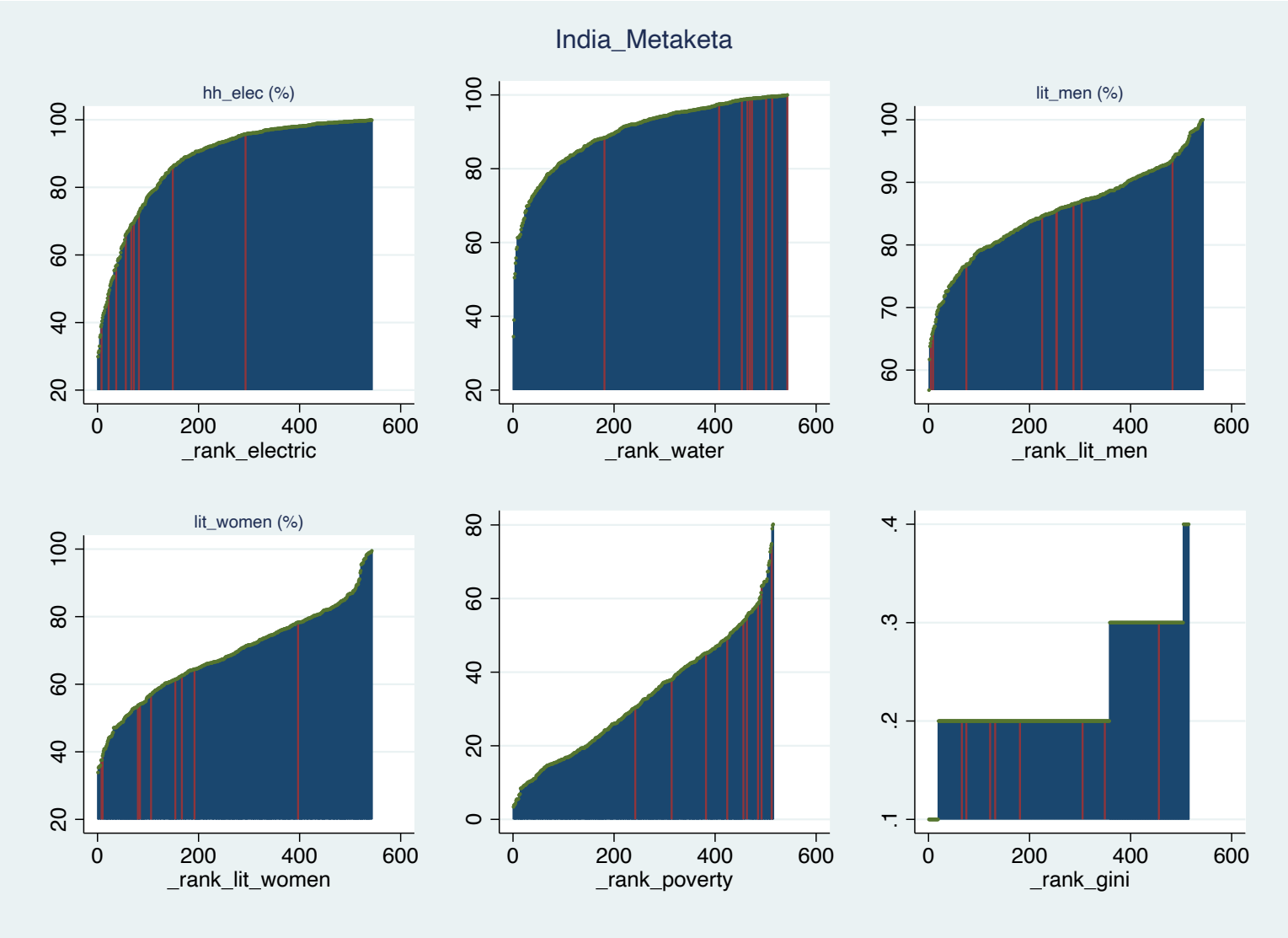
Brazil Sub-national (Misc)



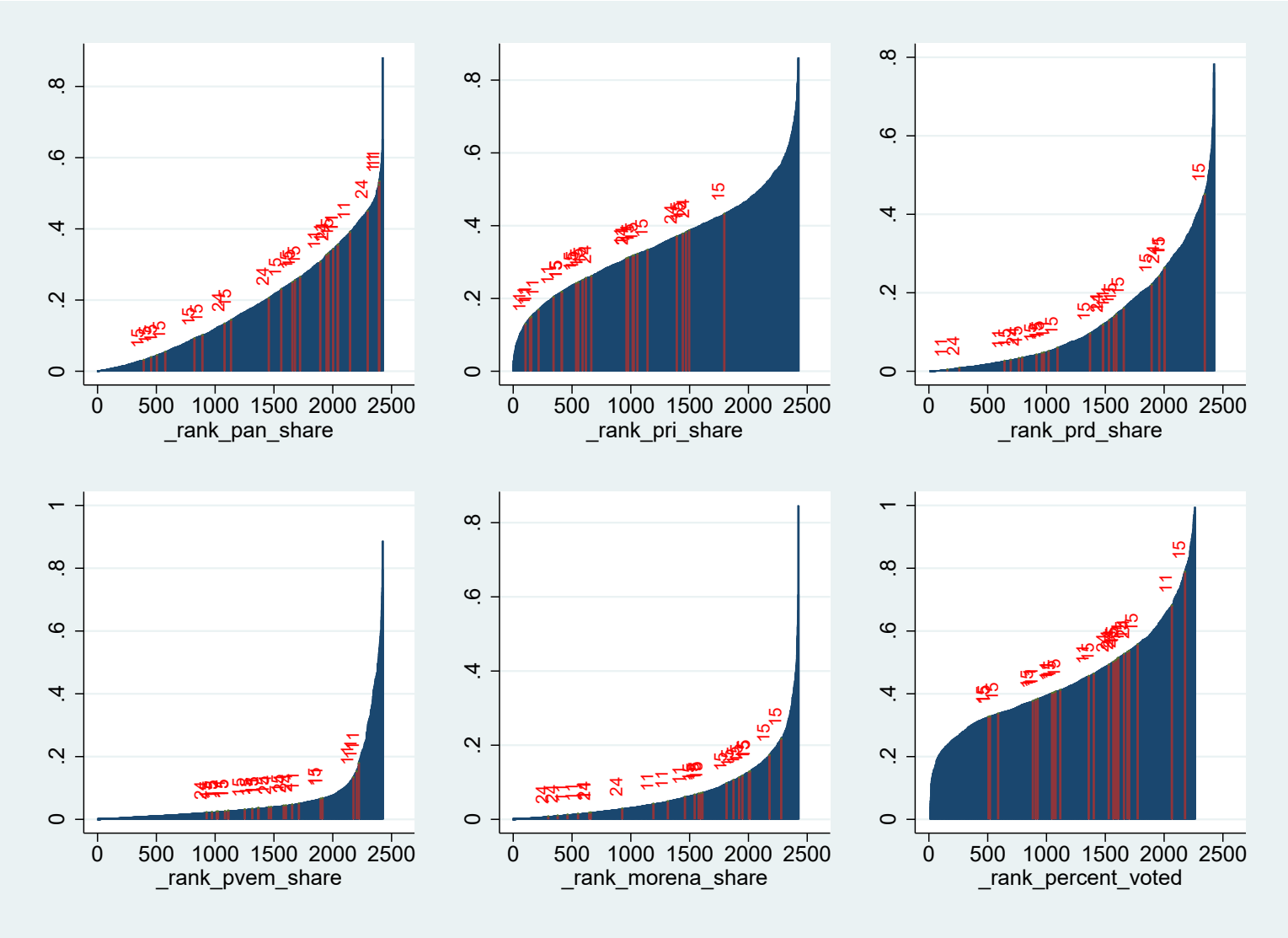
Burkina Faso Sub-national (Misc)



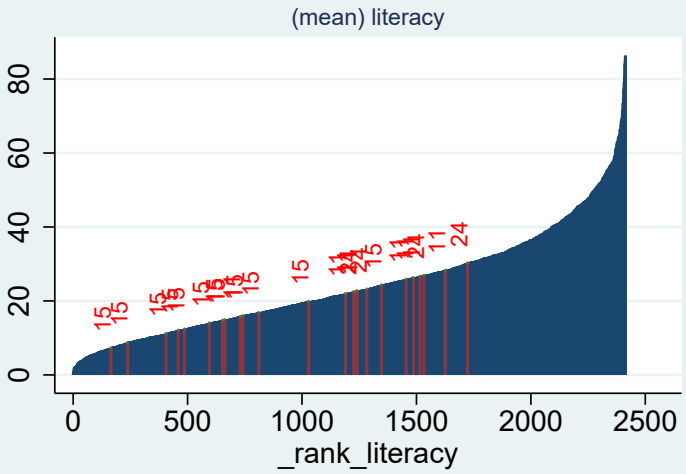
India Sub-national (Misc)

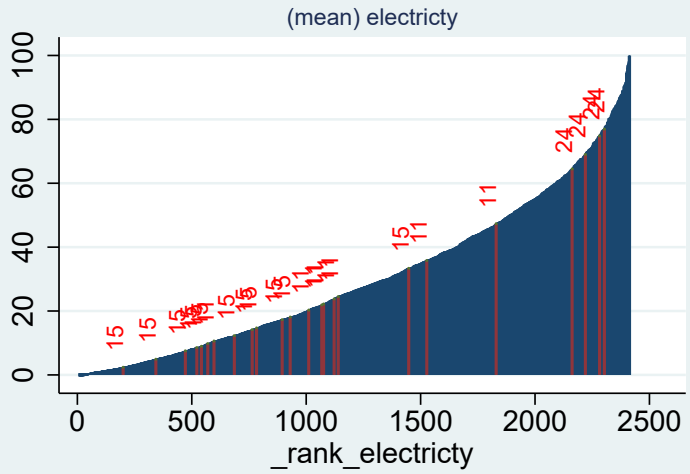


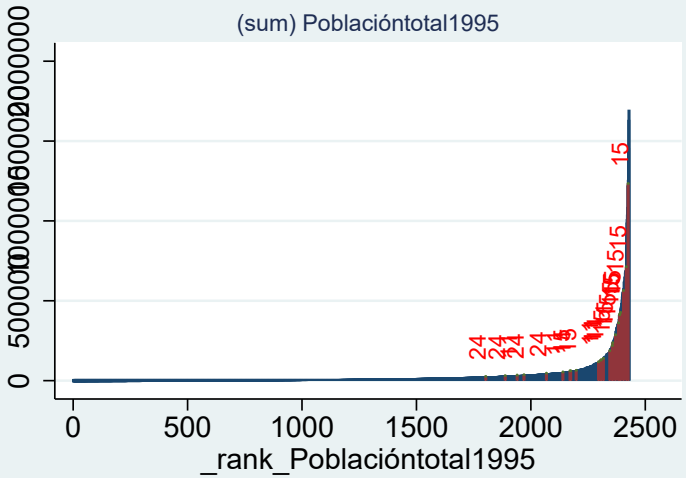
Mexico Sub-national (Electoral)

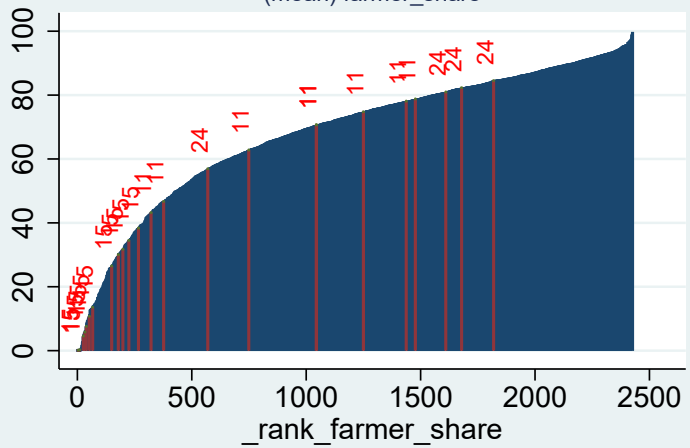


Mexico Sub-national (Misc)

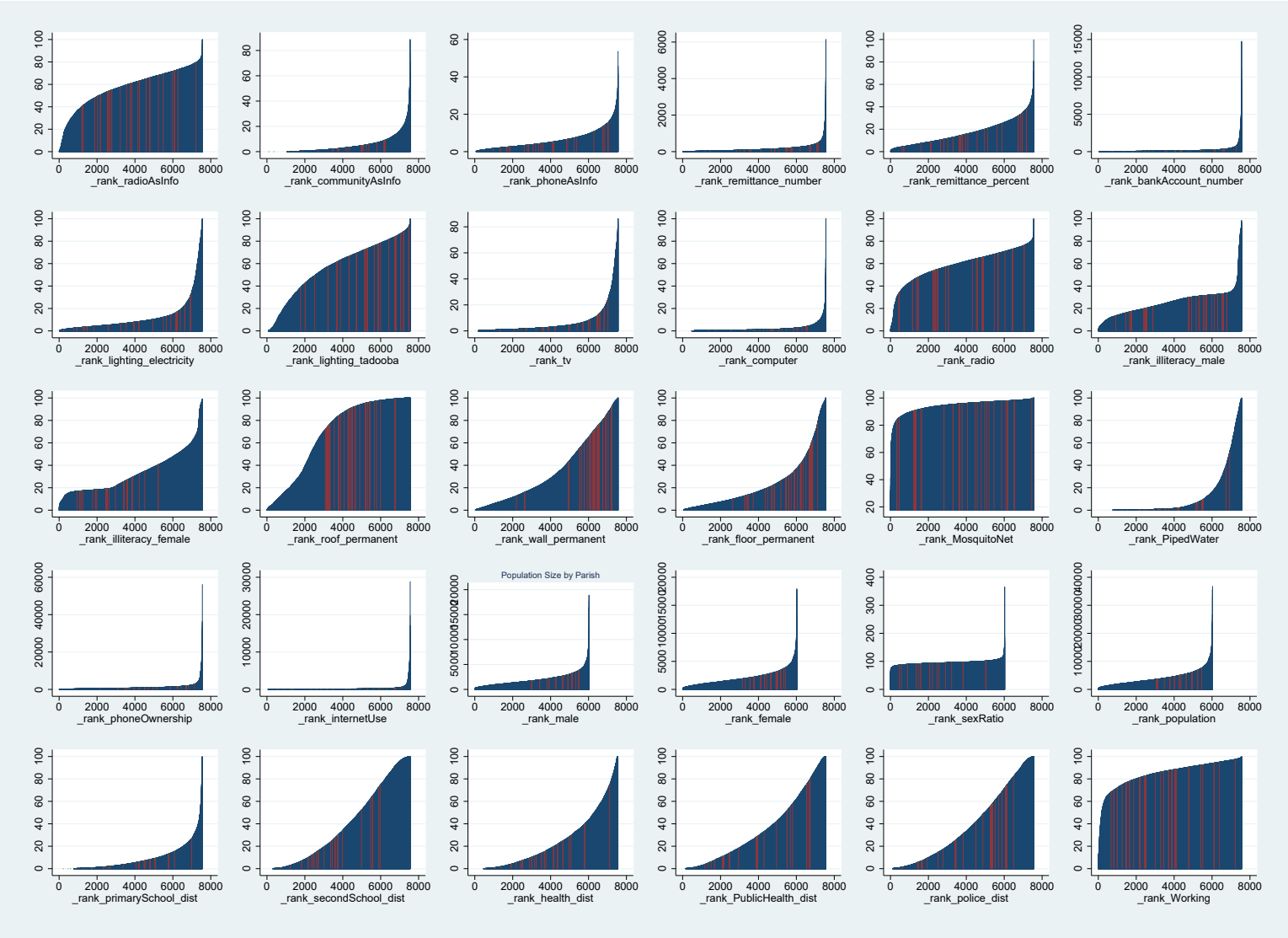




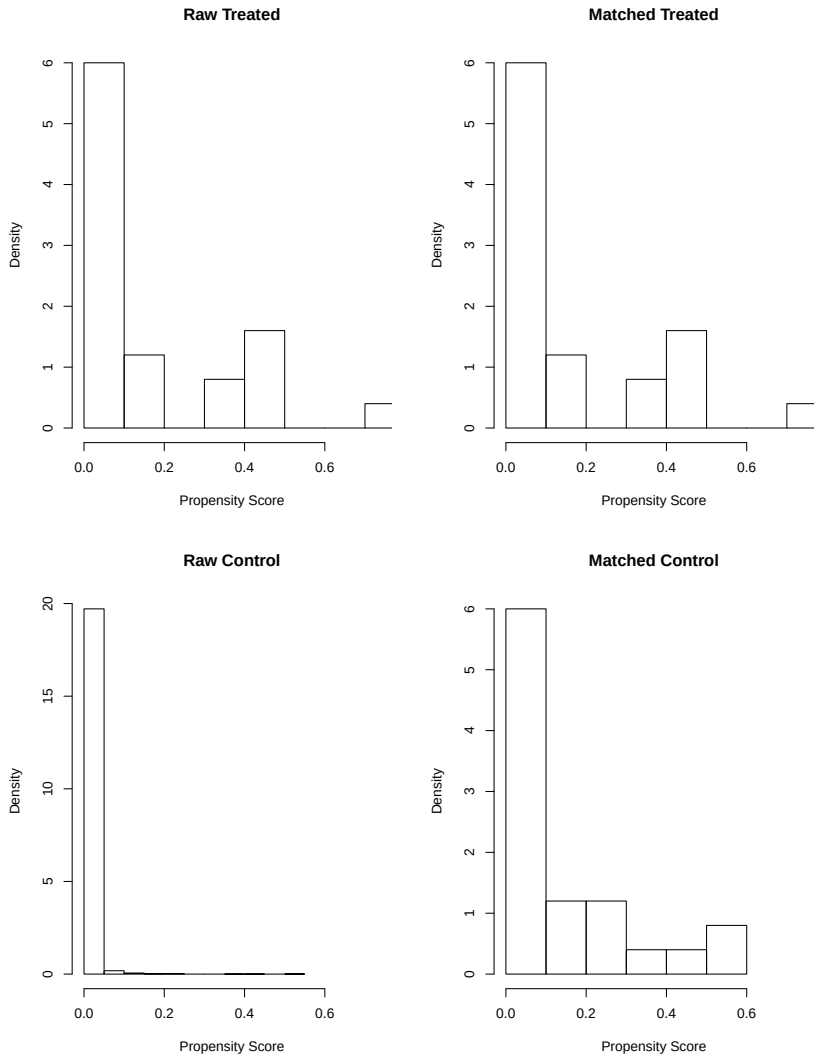




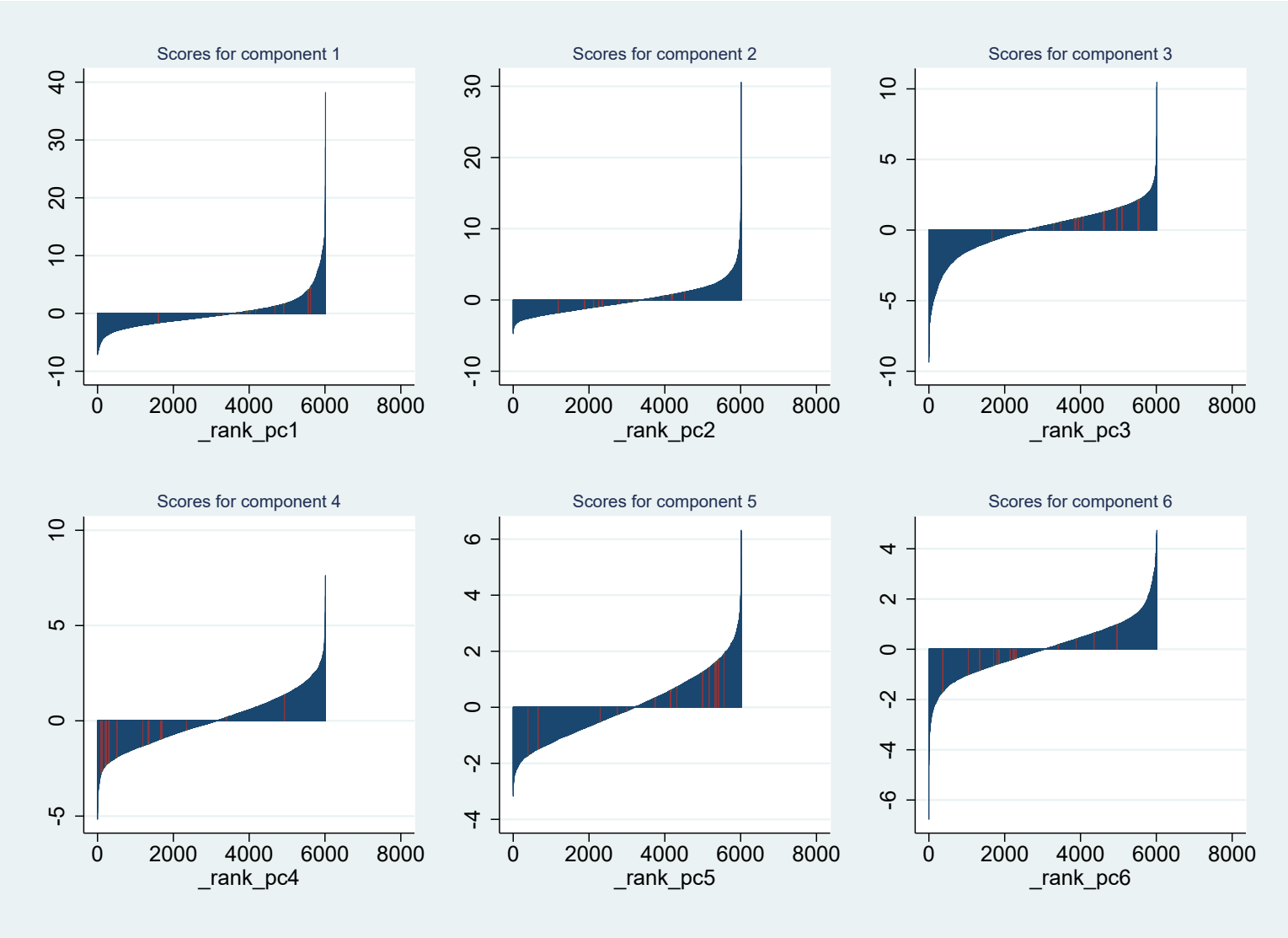
Uganda 1 Sub-national (Misc)



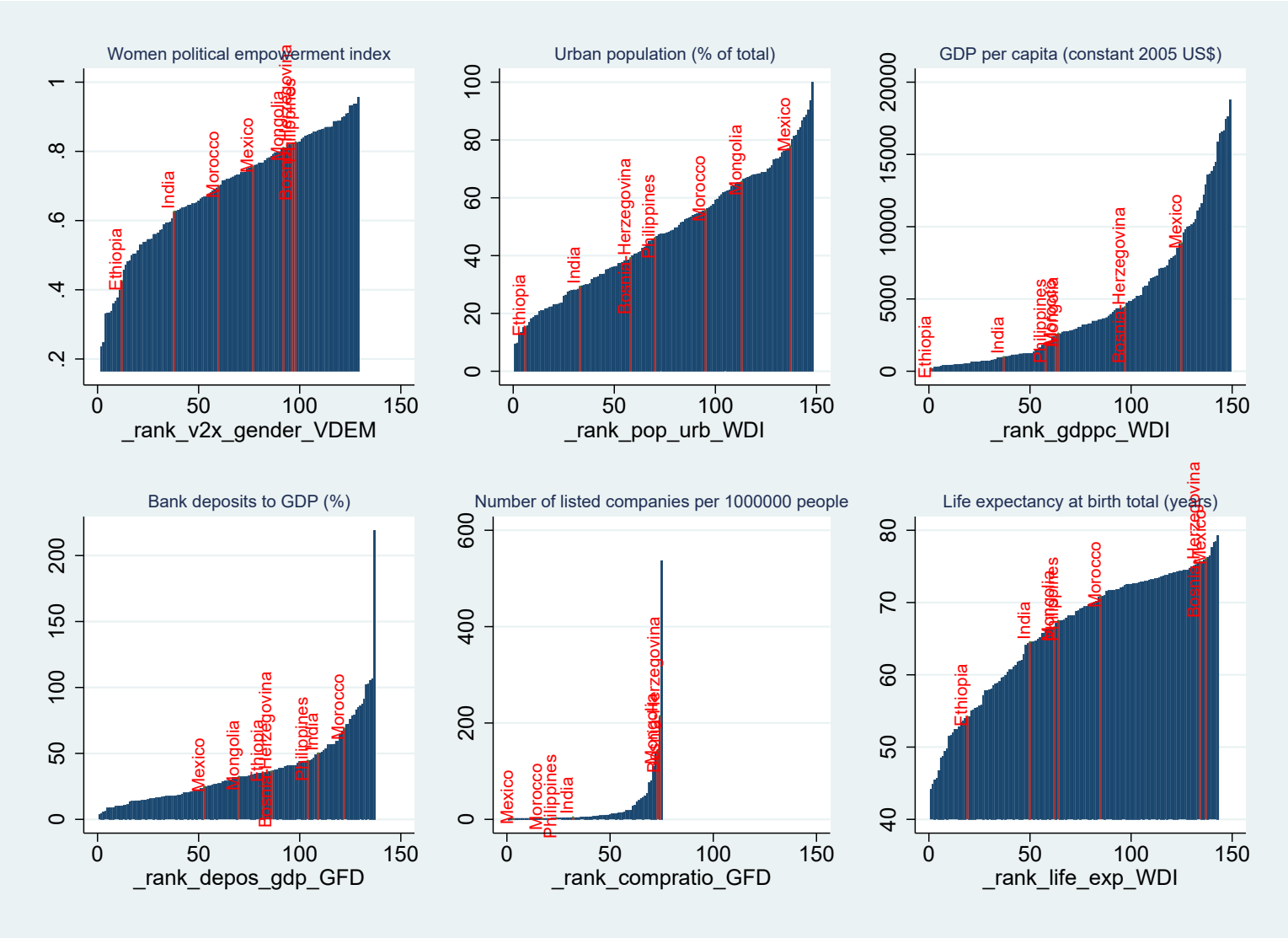
Uganda 1 Sub-national (Matching)



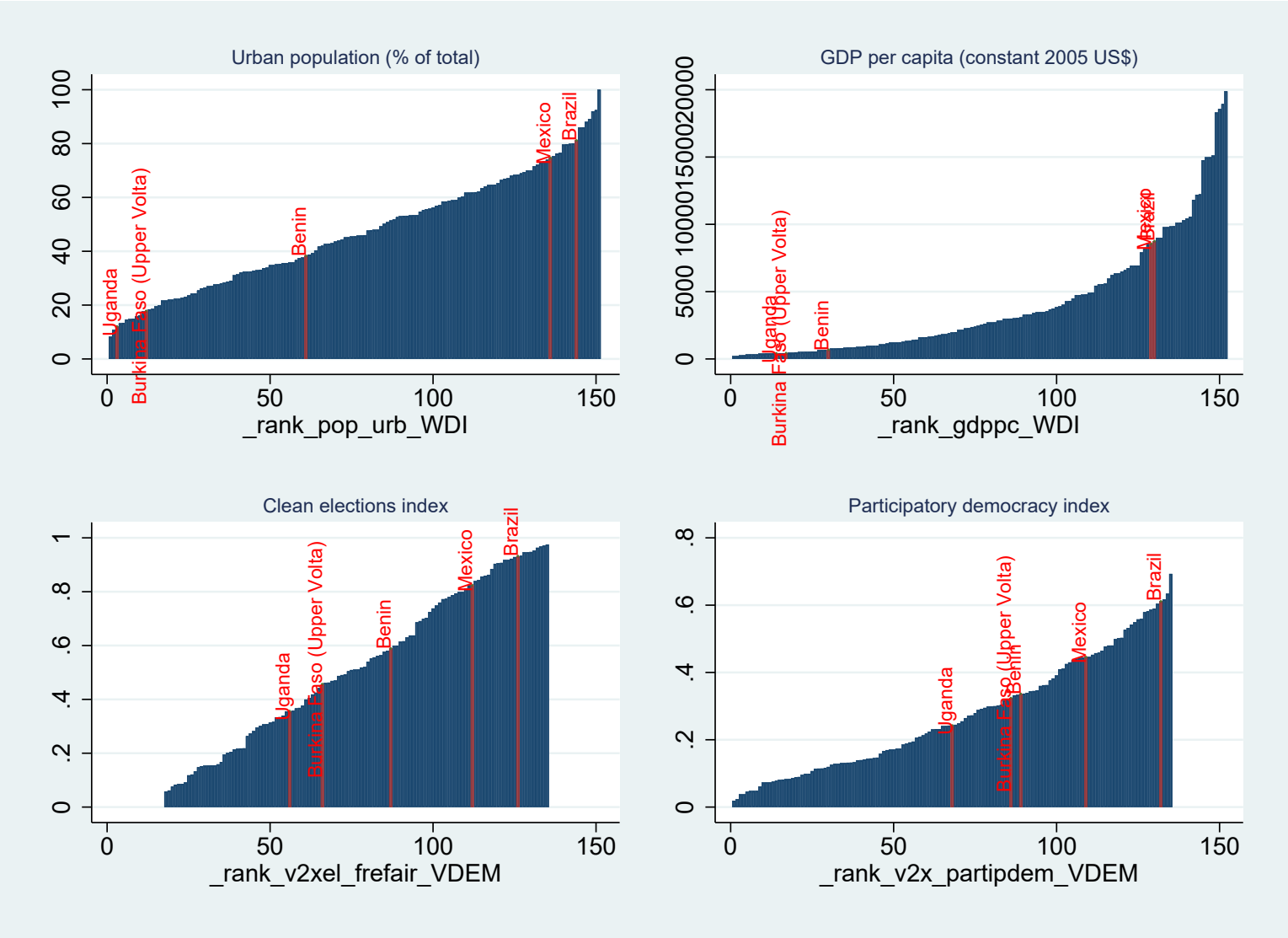
Uganda 1 Sub-national (PCA)



Meager Country Level



Metaketa I Country Level (Old)



Policing Metaketa Country Descriptives Table

Table 2. Descriptive statistics for study sites. Sources are provided in the supplementary materials, section A.7.

	Brazil	Colombia	Liberia	Pakistan	Philippines	Uganda
Political freedoms	Partly free	Partly free	Partly free	Partly free	Partly free	Not free
Regime type	Electoral democracy	Electoral democracy	Electoral democracy	Electoral autocracy	Electoral autocracy	Electoral autocracy
Corruption score	45/100	39	32	31	46	26
Criminal justice score	34/100	34	31	35	31	31
Income category	Upper middle	Upper middle	Low	Lower middle	Lower middle	Low
Inequality (Gini coefficient)	54	50	35	33	44	42
Study site	Santa Catarina	Medellín	Monrovia	Punjab Province	Sorsogon Province	–
Type	State	Large city	Large city	Two districts	Province	Country
Rate of crime victimization (%)						
Simple assault	1	5	6	5	3	6
Burglary	4	15	17	16	2	19
Armed robbery	0	6	3	10	0	2
Murder	1	9	7	–	1	9
Trust in police (%)	79	47	46	23	86	62
Citizen cooperation (%)	1	5	-	2	1	5
Police capacity indicators						
Vehicle	√					
Motorbike	√	√			√	√
Gun	√	√		√	√	
Radio	√	√		√	√	√
Computer	√	√			√	
Printer	√	√			√	
Camera	√	√			√	
Officers per capita	1:473	1:333	1:950	1:560	1:991	1:910
Budget per officer	$56,000	$18,000	$3642	$3400	$18,000	-
Citizens per station	-	143,000	21,428	500,000	44,444	-
Officer rotation rate	-	15 months	-	1 month	6 months	17 months

Resource Management Metaketa Country Descriptives Table

Table 1. Features of the research contexts and experimental designs

	Brazil	China	Costa Rica	Liberia	Peru	Uganda
Contextual features of CPR						
Resource	Groundwater	Surface water	Groundwater	Forest	Forest	Forest
Community	Rural villages	Urban microneighborhoods	Rural villages	Villages	Indigenous communities	Villages
Primary threat to resource	Drought, overuse	Individual, industrial pollution	Drought, overuse	Overcutting by residents	Extraction by outsiders	Overcutting by residents
Components of harmonized interventions						
Community workshops	✓	—	✓	✓	✓	✓
Monitor selection, training, incentives	✓	✓	✓	✓	✓	✓
Monitoring of the resource	✓	✓	✓	✓	✓	✓
Dissemination to citizens	✓	✓	✓	✓	✓	✓
Dissemination to management bodies	—	(Alternate arm)	✓	✓*	✓*	✓*
Experimental design						
Alternate treatment arm	Conservation plan making	Dissemination to government	—	Negotiation training	—	SMS reminders
Experimental design	Three-arm†	2×2 factorial	Two-arm	2×2 factorial	Two-arm	Three-arm†
No. of monitoring communities (N_M)	80	80	81	60	39	60
No. of nonmonitoring communities (N_{-M})	40	80	80	60	37	50
Common outcome measurement						
Duration of implementation, mo	12	15	12	12	13	12
Primary compliance measure	SMS reports received	Dissemination posters	Reports submitted	Monitoring walks completed	Reports submitted	Reports submitted
Primary resource outcome	Well electricity usage	Pollutant concentration in water	Well electricity usage, water quality	Deforestation	Deforestation	Deforestation, forest quality
Endline citizen survey	✓	✓	✓	✓	✓	✓

N_M denotes the number of communities assigned to any treatment condition with community monitoring, and N_{-M} denotes the number assigned to any treatment condition without community monitoring.

*In the forest studies, the community constitutes at least one of the possibly overlapping management bodies.

†In both three-arm designs, communities assigned to the alternative treatment arm received both monitoring and the alternative treatment.

External Validity Errors (And Examples)

	Null *(Actually True)*	Null *(Actually False)*
Null *(Rejected)*	**Type 1 EV Error**: -False positive (No EV, but EV inferred)	*Correct EV Inference* -True positive
Null *(Not Rejected)*	*Correct EV Inference* -True negative	**Type 2 EV Error**: -False negative (Is EV, but not inferred)

External Validity Errors (And Examples)

	Null *(Actually True)*	Null *(Actually False)*
Null *(Rejected)*	**Type 1 EV Error**: -False positive (No EV, but EV inferred)	*Correct EV Inference* -True positive
Null *(Not Rejected)*	*Correct EV Inference* -True negative	**Type 2 EV Error**: -False negative (Is EV, but not inferred)

- Type 1 EV Errors: XXXX examples
- Type 2 EV Errors: XXXX examples

Four Primary Approaches to External Validity

		Study/Sample Type	
		Single	*Multiple*
External Validity Inference	*Generalizability* (G)	(Random) Sampling	Same Target Synthesis
	Transportability (T)	Principled Prediction	Synthetic Prediction

Four Primary Approaches to External Validity

		Study/Sample Type	
		Single	*Multiple*
External Validity Inference	*Generalizability* (G)	(Random) Sampling	Same Target Synthesis
	Transportability (T)	Principled Prediction	Synthetic Prediction

But...

1. (Random) sampling may neglect causal structure
2. Same Target Synthesis (e.g., meta-analysis) relies on selected/available studies
3. Principled Prediction and Synthetic Prediction almost always occurs *post hoc*