External Validity and Meta-Analysis
Supporting Information

Tara Slough[*] Scott A. Tyson[†]

Contents

S1 Proofs **S-2**

S2 Approach to Existing Meta-Analyses **S-3**
S2.1 Meta-analyses in political science S-3
S2.2 Evaluating existing studies S-3
S2.3 Procedures S-5

S3 Prospective Meta-Analyses **S-7**
S3.1 Dunning et al. (2019) S-7
S3.2 de la O et al. (2021) S-9
S3.3 Slough et al., (2021) S-10
S3.4 Blair et al. (2021) S-12
S3.5 Coppock et al. (2020) S-14

S4 Retrospective Meta-Analyses **S-15**
S4.1 Blair, Christensen, and Rudkin (2021) S-15
S4.2 Incerti (2020) S-16
S4.3 Schwartz and Coppock (2020) S-17

[*] Assistant Professor, New York University. taraslough@nyu.edu
[†] Assistant Professor, University of Rochester. styson2@ur.rochester.edu

S1 Proofs

Proof of Theorem 1. Sufficiency is obvious. For necessity, suppose not. Since studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent, but not measurement harmonized, then for m_1 and m_2:

$$\tau_{m_1}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_2). \tag{1}$$

Applying external validity, at m_2 and (ω', ω''), it must be that for θ_1 and θ_2

$$\tau_{m_2}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_2). \tag{2}$$

Combining (1) and (2),

$$\tau_{m_1}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_1),$$

contradicting divergent validity. □

Proof of Theorem 2. Sufficiency is obvious. For necessity, target-equivalence implies that there are two contrasts, (ω_1', ω_1'') and (ω_2', ω_2''), where

$$\tau_m(\omega_1', \omega_1'' \mid \theta_1) = \tau_m(\omega_2', \omega_2'' \mid \theta_2), \tag{3}$$

and proceeding by contradiction, suppose that $(\omega_1', \omega_1'') \neq (\omega_2', \omega_2'')$. Applying external validity at m and (ω_1', ω_1''), we have that

$$\tau_m(\omega_1', \omega_1'' \mid \theta_1) = \tau_m(\omega_1', \omega_1'' \mid \theta_2). \tag{4}$$

Combining (3) and (4) yields

$$\tau_m(\omega_1', \omega_1'' \mid \theta_2) = \tau_m(\omega_2', \omega_2'' \mid \theta_2),$$

which, since the setting and contrasts were arbitrary, implies that the the treatment effect must be the same at (ω_1', ω_1'') and (ω_2', ω_2'') in any setting. Thus, external validity allows us to suppress the dependence of the treatment effect function on θ. Because $\mathcal{C}$ is a compact subset of $\mathbb{R}^2$, it is a two-dimensional manifold. Define

$$\kappa \equiv \tau_m(\omega_1', \omega_1'' \mid \theta),$$

which by external validity, is the same at almost any $\theta \in \Theta$. We are interested in the level set $\tau_m^{-1}(\kappa) \subset \mathcal{C}$. Since the derivative of $\tau_m(\omega', \omega'' \mid \cdot)$ has full rank for almost every contrast, $(\omega', \omega'') \in \mathcal{C}$, the set of regular points of τ_m is of full measure on $\mathcal{C}$. Thus, if κ is not a regular value, then $\tau_m^{-1}(\kappa)$ does not contain any regular points, and is thus of Lebesgue measure zero. Suppose, instead, that κ is a regular value, and thus, $\tau_m^{-1}(\kappa)$ is a set of regular points. By the Preimage Theorem (e.g., Guillemin and Pollack, 1974, pg. 21), the set $\tau_m^{-1}(\kappa)$ is a submanifold of $\mathcal{C}$, and moreover,

$$\dim \tau_m^{-1}(\kappa) = \dim \mathcal{C} - \dim \mathbb{R} = 2 - 1 = 1.$$

Thus, dim $\tau_m^{-1}(\kappa) <$ dim $\mathcal{C}$, implying that $\tau_m^{-1}(\kappa)$ is a Lebesgue measure zero subset of $\mathcal{C}$, completing the argument.[1] □

[1] The Preimage Theorem applies since all sets in our framework are in $\mathbb{R}$. Otherwise, similar arguments would follow applying the Regular Level Set Theorem, which is equivalent to the Constant Rank Theorem, see Tu (2011, Ch. 9-10).

Proof of Corollary 1. The proof is the same as that of Theorem 2, with the smooth map $\tau_m(\omega_1', \omega'' \mid \theta)$, and replacing $\mathcal{C}$ with Ω, and noting that the Preimage Theorem then implies that

$$\dim \tau_m^{-1}(\kappa; \omega'') = \dim \Omega - \dim \mathbb{R} = 1 - 1 = 0,$$

thus completing the argument. □

S2 Approach to Existing Meta-Analyses

S2.1 Meta-analyses in political science

In Table 1, we identify and classify the most recent meta-analyses in political science, including the four complete Metaketa projects, and published meta-analyses in three leading political science journals (*American Journal of Political Science*, *American Political Science Review*, and *The Journal of Politics*) as well as meta-analyses on political subjects in general science journals.

In the panels of Table 1, we distinguish between *prospective* and *retrospective* meta-analyses. The treatment-harmonized RCTs constitute prospective meta-analyses since the constituent studies (or sites) were designed with an eye to formal synthesis. In retrospective meta-analysis, researchers collect and synthesize estimates from a variety of existing studies. We identify one study, Kalla and Broockman (2018), which uses both approaches to synthesize existing experiments on persuasion while incorporating a number of new experiments. We classify the constituent study design as experimental or observational, a distinction described by Rosenbaum (2002). All of the meta-analyses that we identify use fixed- or random-effects estimators, which we relate to our framework below.[2]

The meta-analyses in Table 1 analyze the findings of 755 constituent studies. Table S1 provides an accounting of the number of studies reported in each meta-analysis. Note that in some cases, studies generate multiple treatment effect estimates or multiple studies are reported per paper. We endeavor to define the number of studies in a symmetric manner across the meta-analyses we have identified.

S2.2 Evaluating existing studies

Theorem 3 shows that external validity and harmonization are necessary and sufficient for target-equivalence in meta-studies. This implies that **limited or insufficient harmonization of any two constituent studies is a sufficient condition for lack of target-equivalence**. As we note, harmonization—of both contrasts and measurement strategies—should be is assessed and judged in terms of their *construct validity* with the underlying construct they are meant to represent in each study. Specifically, the analyst needs to evaluate the extent to which an empirical object—a measure of an instrument or outcome—corresponds to an underlying substantive concept. Harmonization means that measurement strategies and contrasts are identical *in the model*, meaning they represent the same construct, but does not mean that they are the same in a literal or material sense.

[2] Some of the studies also employ meta-regression estimators that build upon the random- and fixed-effects estimators that we discuss.

Study	Type	N studies	Elaborated here
Dunning et al. (2019)	Prospective	6	✓
de la O et al. (2021)	Prospective	6	✓
Slough et al. (2021)	Prospective	6	✓
Blair et al. (2021)	Prospective	6	✓
Coppock, Hill, and Vavreck (2020)	Prospective	59	✓
Blair, Christensen, and Rudkin (2021)	Retrospective	37	✓
Blair, Coppock, and Moor (2020)	Retrospective	105*	
Eshima and Smith (2022)	Retrospective	16	
Godefroidt (2021)	Retrospective	326†	
Incerti (2020)	Retrospective	24‡	✓
Kertzer (2020)	Retrospective	48§	
Schwarz and Coppock (2022)	Retrospective	67	✓
Kalla and Broockman (2018)	Mixed	49	
Total		755	

Table S1: Enumeration of studies in the meta-analyses described in Table 1.

* We calculate the number of studies as the number of estimates reported in the eight principal meta-analyses reported in Blair, Coppock, and Moor (2020) Figure 4. They additionally meta-analyze some list experiments on topics for which there were less than three accumulated studies.

† Godefroidt (2021) analyzes 1,733 unique estimates from 326 studies reported across 241 manuscripts. We count the number of studies.

‡ Incerti (2020) analyzes 8 field and 18 survey experiments in separate meta-analyses. Our elaboration considers two of the survey experiments that he reports.

§ Kertzer (2020) analyzes 48 experiments in 26 studies.

As in Figure S1, our harmonization refers to the idea that study-level attributes (i.e., instruments or measurement strategies) measure a common construct. In this figure, it is clear that harmonization would hold if $\omega' = \omega'_1 = \omega'_2$ etc.

Ultimately arguments about harmonization rely on a positive argument by the analyst that follows from substantive and contextual factors, and the details of these arguments will vary from case to case. Because we are not experts in every study (or area of inquiry) represented in Table 1, we do not presuppose that our ability to develop (or critique) the authors' arguments. Instead, we focus on identifying potential threats to harmonization that would ideally be addressed, thus clarifying the interpretation of what those meta-analyses conclude. Our analysis, therefore **does not establish harmonization or a lack thereof**. It aims to highlight the considerations that it raises for existing studies. Authors of future analyses can use this framework to develop the arguments necessary to support harmonization or motivate the adoption of alternative meta-analytic models when harmonization fails (see p. 28).

With respect to external validity, we do not assess these studies directly. In principle, within our

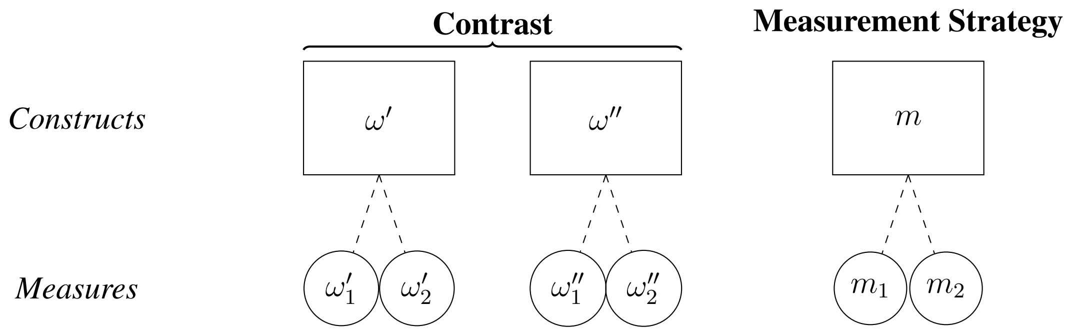

Figure S1: Relationship between constructs and their measures in two studies, indexed by the subscripts 1 and 2. The dashed lines linking constructs correspond to the discussion of construct validity in Adcock and Collier (2001).

framework, authors would specify a mechanism and describe the set of settings, Θ, where a mechanism could present for some non-empty subset of units or observations, (i.e. $|\mathcal{D}| > 0$). Some studies in Table 1 are more precise in the specification of the mechanism than others. For example Dunning, Grossman, Humphreys, Hyde, McIntosh, and Nellis (2019) and Incerti (2020) clearly posit *voter updating from an informational signal* as a mechanism. They examine the effects of this mechanism on voter beliefs and vote choice. Dunning (2012) identifies subsets of experimental subjects for which this mechanism should have a positive or negative effect, and defines $\mathcal{D}$ in this way.[3] In order to evaluate external validity of this mechanism in either meta-analysis, authors should make a positive case for the scope conditions of the mechanism. By positing these scope conditions, or describing Θ, researchers could better justify that the cases they study fall within these conditions.

Since authors generally have more expertise or understanding of the mechanisms they propose, we suggest that authors make a positive argument that specifies (i) the mechanism(s) of interest and (ii) the scope conditions for each mechanism to justify claims to external validity. Where such analysis suggests that external validity is more local than the set of settings, authors can follow the guidance on p. 29.

S2.3 Procedures

Given the large number of studies in Table S1, we pursue a limited elaboration of constituent studies to assess the prevalence of these potential issues. From each of the elaborated meta-studies, we describe two randomly-selected constituent studies. Lack of harmonization in any two constitutent studies is sufficient to establish a lack of target-equivalence.

We characterize constituent studies using the meta-analysis article or supplemental information. Where insufficient information is provided in these documents, we identify papers reporting the results of constituent studies. As such, we prioritize the characterization of a design in a meta-study article over those in articles documenting the constitutent studies whenever the meta-study

[3] Incerti (2020) looks at corruption revelation which should, in theory, lead to downward updating by voters.

article is sufficiently detailed.

We selected the studies for elaboration based on two criteria. First, we elaborate all of the prospective meta-analyses because harmonization is more explicitly discussed in these meta-analyses than in the retrospective meta-analyses. As such, these constitute harder cases for identifying a lack of harmonization. We further, select three of the retrospective meta-analyses among the set of meta-analyses for which we can identify the constituent studies from the meta-analysis manuscript. We therefore discuss two randomly selected studies from the following meta-analyses: Blair, Christensen, and Rudkin (2021), Incerti (2020), and Schwarz and Coppock (2022).

S3 Prospective Meta-Analyses

S3.1 Dunning et al. (2019)

Dunning, Grossman, Humphreys, Hyde, Mcintosh, Nellis, Adida, Arias, Bicalho, Boas, Buntaine, Chauchard, Chowdhury, Gottlieb, Hidalgo, Holmlund, Jablonski, Kramon, Larreguy, Lierl, Marshall, McClendon, Melo, Nielson, Pickering, Platas, QuerubÍn, Raffler, and Sircar (2019) analyze estimates of causal effects of pre-electoral dissemination of politician performance information on voting behavior. The meta-analysis includes six experiments.[4] This is a prospective meta-analysis: the six studies were designed with the aim of ultimately combining the studies in a meta-analysis. As in many studies, there are multiple outcomes or measurement strategies. We consider one outcome: turnout, though this exercise could be extended to elaborate additional measurement strategies.

Potential harmonization concerns:

ω'': Prospective and retrospective informational treatments do not necessarily convey the same information to voters.

ω': The control state, which aims to capture voters' priors, has a different relationship to the treatment – the information provided. In Uganda, voters' priors were better correlated with the randomly assigned informational signal than in Burkina Faso. This means that voters in Burkina Faso may have had a greater scope for learning than voters in Uganda.

m: Self-reported intention to vote before an election is not necessarily the measure as validated voter turnout measured after the election.

[4] A seventh study was terminated in the field before the collection of endline data so it is omitted from the meta-analysis.

Attribute		Burkina Faso study	Uganda study #1
Setting	θ	39 rural municipalities in Burkina Faso during the 2016 Burkinabe municipal elections.	265 villages located in 11 Ugandan parliamentary constituencies during the 2015-2016 Ugandan parliamentary elections.
Contrast	ω''	With an enumerator, subjects viewed informational flashcards on public goods provision benchmarked to national and regional averages. (This is a *retrospective* informational treatment.)	Screening of "meet the candidates" videos in villages. (This is a *prospective* informational treatment.)
	ω'	Pure control (status quo condition). No information was provided by researchers or their partners. Theoretically, this corresponds to voters' "prior" beliefs. The correlation between this prior and the informational signal provided (as measured by the researchers) is 0.14 (95% CI: [0.08, 0.19]).	Pure control (status quo condition). No information was provided by researchers or their partners. Theoretically, this corresponds to voters' "prior" beliefs. The correlation between this prior and the informational signal provided (as measured by the researchers) is 0.32 (95% CI: [0.25, 0.38]).
Measurement strategy	m	A binary indicator capturing intention to vote as measured in a pre-electoral survey.	A binary indicator for turnout coded from a phone survey within 48 hours of the election among respondents who correctly answered a factual question about the biometric screening procedure.
Estimand	τ	Two conditional ATEs. The authors condition on the relationship between a voter's prior and the information provided: in other words, they define two subgroups in which the informational treatment represented "good news" or "bad news." They estimate the conditional ATE for each of these two subgroups.	Two conditional ATEs. The authors condition on the relationship between a voter's prior and the information provided: in other words, they define two subgroups in which the informational treatment represented "good news" or "bad news." They estimate the conditional ATE for each of these two subgroups.

Table S2: Two randomly selected studies from Dunning, Grossman, Humphreys, Hyde, Mcintosh, Nellis, Adida, Arias, Bicalho, Boas, Buntaine, Chauchard, Chowdhury, Gottlieb, Hidalgo, Holmlund, Jablonski, Kramon, Larreguy, Lierl, Marshall, McClendon, Melo, Nielson, Pickering, Platas, QuerubÍn, Raffler, and Sircar (2019). Note that there were two studies in Uganda. Uganda #1 corresponds to the study by Platas and Raffler.

S3.2 de la O et al. (2021)

de la O et al. (2021) analyze estimates of causal effects of assignment to information about formalization and/or assistance in the formalization process on formalization behavior of households or firms. The meta-analysis includes six experiments. The six studies were designed with the aim of ultimately combining the studies in a meta-analysis (per a public pre-analysis plan). We focus on one outcome, or measurement strategy: formalization, though this exercise could be extended to consider the other measurement strategies in the meta-analysis.

Attribute		Democratic Republic of Congo	Nigeria
Setting	θ	824 households eligible to receive a property title in Kananga, DRC.	641 vendors in markets across 37 Local Community Development Areas and 20 Local Government Areas in Lagos, Nigeria.
Contrast	ω''	The Kananga provincial government provides information on the costs and benefits of property titling to households and offers assistance to complete paperwork and discounted rates for obtaining a legal property title.	A think tank provides information on the costs and benefits of formalization and offers assistance in filling out registration paperwork.
	ω'	Pure control (status quo condition). Ostensibly, households can pursue property titles through the local government, though households in the sample have chosen not to do so.	Pure control (status quo condition). Ostensibly, market vendors can pursue market vendor registration, a form of formalization, though vendors in the sample have largely chosen not to do so.
Measurement strategy	m	Binary indicator for formalization by endline. Formalization refers to having a property title.	Binary indicator for formalization by endline. Formalization refers to registration as a market vendor.
Estimand	τ	ITT of assignment to information/assistance treatment on formalization behavior. However, compliance was perfect in the DRC study, so the ITT is equivalent to the ATE.	ITT of assignment to information/assistance on formalization behavior. However, compliance was perfect in the Nigeria study, so the ITT is equivalent to the ATE.

Table S3: Two randomly selected studies from de la O et al. (2021).

Potential harmonization concerns:

ω'': The articulated costs and benefits of formalization may be different for households than for vendors. Moreover, households in DRC were provided fee reductions for formalization while vendors in Nigeria were not.

ω': The constraints to titling property versus registering a small business may be very different. Very little information is provided on the "control" state in either context.

S3.3 Slough et al., (2021)

Slough et al. (2021) analyze estimates of causal effects of community monitoring of natural resources on resource conservation. The meta-analysis includes six experiments. The six studies were designed with the aim of ultimately combining the studies in a meta-analysis (per a public pre-analysis plan). We focus on one outcome, or measurement strategy: natural resource status, though this exercise could be extended to consider the other measurement strategies in the meta-analysis.

Attribute		Costa Rica study	Uganda study
Setting	θ	161 rural villages in semi-arid regions of Costa Rica facing low groundwater levels.	110 rural forest-edge villages in Uganda.
Contrast	ω''	Community workshops initiated a community monitoring program for groundwater. Monitors were selected, trained, and incentivized to monitor the resource. They then disseminated their findings to citizens and to local elected water organization boards.	Community workshops initiated a community forest monitoring program. Monitors were selected, trained, and incentivized to monitor the resource. They then disseminated their findings to citizens and in village meetings.
	ω'	Pure control (status quo condition). There is no existing community monitoring of groundwater levels/quality. Community members do not regularly receive information about groundwater levels/quality nor are there community meetings/fora focused on water use issues.	Pure control (status quo condition). Forests are monitored by National Forest Authority officials, who are not members of the community. In 28% of communities, this monitoring occurred at least weekly at baseline. Forest issues are discussed in community meetings. At baseline, 45% of community members reported discussing forest issues in these fora in the last month.
Measurement strategy	m	A z-score index comprised of (a) well electricity usage and (b) chemical measures of water quality.	A z-score index comprised of (a) remote-sensed measures of tree-cover loss and (b) on-the ground assessment of forest quality in a sample of forest transects.
Estimand	τ	Intent to treat (ITT) effect of assignment to community monitoring. Note that monitoring occurred in approximately 80% of treatment communities in each quarter of the intervention (the principal measure of compliance with treatment assignment).	Intent to treat (ITT) effect of assignment to community monitoring. Note that monitoring occurred in 90-100% of treatment communities in each quarter of the intervention (the principal measure of compliance with treatment assignment).

Table S4: Two randomly selected studies from Slough et al. (2021).

Potential harmonization concerns:

ω'': The monitoring process and information generated by monitoring are different because the resource systems (ground water and forests) are very different. Additionally, the process of disseminating findings to a "management authority" for the resource varies because the institutional context is different in Costa Rica and Uganda.

ω': Monitoring was absent in the status quo condition in Costa Rica but relatively frequent in Uganda. This means the treatment introduced monitoring in Costa Rica but only augmented the amount of existing monitoring in Uganda.

m: Measures of water usage/quality and deforestation/forest quality may not constitute an equivalence class. If this is the case, the z-score normalization will not address lack of measurement harmonization.

S3.4 Blair et al. (2021)

Blair et al. (2021) analyze estimates of causal effects of community policing on crime victimization of citizens. The meta-analysis includes six experiments. This is a prospective meta-analysis: the six studies were designed with the aim of ultimately combining the studies in a meta-analysis. As in many studies, there are multiple outcomes or measurement strategies. We consider one outcome: crime victimization, as measured through endline surveys. This exercise could be extended to incorporate the many additional measurement strategies.

Attribute		Colombia study	Pakistan study
Setting	θ	347 police beats (*cuadrantes*) encompassing the majority of Medellín, Colombia from mid-2018 to mid-2019.	108 police beats in Sheikhupura and Nankana Sahib districts of Punjab, Pakistan in 2018 and 2019.
Contrast	ω''	Beat officers introduce bi-monthly town hall meetings. There are no watch forums. Beat officers conduct daily foot patrols. Citizens can provide feedback via a hotline or mobile application. The police engage in problem-oriented policing.	Beat officers introduce monthly town hall meetings and watch forums. They conduct occasional foot patrols. They encourage use of the pre-existing hotline for citizen feedback. They introduce problem-oriented policing.
	ω'	Pure control (status quo). Beat police officers do not conduct town hall meetings or watch forums. They do conduct daily foot patrols. Citizens can provide feedback via a hotline or mobile application. The police engage in problem-oriented policing.	Pure control (status quo). Beat police officers do not conduct town hall meetings or watch forums. They conduct occasional foot patrols. Citizens can provide feedback via a hotline. The police do not engage in problem-oriented policing.
Measurement strategy	m	z-score index of personal exposure to violent crime and non-violent crime as well as community exposure to violent crime and non-violent crime. Colombia is omitted from the meta-analysis for this outcome because different crimes were measured than in other sites, in part due to different legal classifications of crime.	z-score index of personal exposure to violent crime and non-violent crime as well as community exposure to violent crime and non-violent crime.
Estimand	τ	For relevant outcomes, ITT effects of assignment to community monitoring. There are multiple measures of the "first stage" ATE on compliance (exposure to community policing). The ATE on community awareness in Colombia was 0.838 standard deviations (95% CI: [0.66, 1.02]).	ITT effects of assignment to community monitoring. There are multiple measures of the "first stage" ATE on compliance (exposure to community policing). The ATE on community awareness in Pakistan was 0.406 standard deviations (95% CI: [0.02, 0.80]).

Table S5: Two randomly selected studies from Blair et al. (2021).

Potential harmonization concerns:

- ω'': The introduction of community policing consists of different elements. In Colombia, the treatment introduced bi-monthly town hall meetings. In Pakistan, the treatment introduced monthly town-hall meetings, increased the frequency of foot patrols, and introduced problem-oriented policing for the first time.

ω': Colombian police were already engaged in more aspects of the bundled community policing treatment than their Pakistani counterparts. Specifically, Colombian police were conducting more foot patrols and engaging in problem-oriented policing, unlike the police in Pakistan.

m: Crimes are defined differently in different settings with different laws. Definitions of victimization therefore depend on the underlying legal status of crimes. By fixing the text of questions about crime victimization in different legal settings, the surveys capture different subsets of victimization in different contexts.

S3.5 Coppock et al. (2020)

Coppock, Hill, and Vavreck (2020) analyze estimates of causal effects of political advertisements on support for the targeted candidates. The meta-analysis includes 59 experiments. The 59 studies were (seemingly) designed with the aim of ultimately combining the studies in a meta-analysis. We focus on one outcome, or measurement strategy: favorability toward the targeted candidate, though this exercise could be extended to consider anticipated vote choice.

Note that the design of many of these experiments was a 2×2 factorial design with two different political advertisements. We consider a simplified comparison between the a single advertisement treatment and the relevant comparison condition.

Attribute		Week of May 23, 2016, Ad #1	Week of August 22, 2016, Ad #1
Setting	θ	Representative sample of 1000 Americans in the YouGov survey research panel, surveyed during the week of May 23, 2016.	Representative sample of 1000 Americans in the YouGov survey research panel, surveyed during the week of August 22, 2016.
Contrast	ω''	Respondents were assigned to watch anti-Trump ad labeled "Quotes." (See `https://time.com/4258101/anti-trump-ad-women-quotes/` for a video.)	Respondents were assigned to watch anti-Clinton ad labeled "2 Americas: Immigration." (See `https://www.washingtonpost.com/video/politics/donald-trump-two-americas-immigration-campaig2016/08/19/b2091a70-6607-11e6-b4d8-33e931b5a26d_video.html`)
	ω'	Respondents were assigned to watch a placebo advertisement for car insurance.	Respondents were assigned to watch a placebo advertisement for car insurance.
Measurement strategy	m	Favorability rating of Trump on a five-point scale.	Favorability rating of Clinton on a five-point scale.
Estimand	τ	ATE of assignment to the political ad (versus placebo) on favorability.	ATE of assignment to the political ad (versus control) on favorability.

Table S6: Two randomly selected studies from Coppock, Hill, and Vavreck (2020).

Because the 2016 US presidential race was between two candidates, Trump and Clinton, although the ratings in the two studies are literally different, the measurement strategy in each study represents the same construct, and thus, there are unlikely to be harmonization concerns in this study.

S4 Retrospective Meta-Analyses

S4.1 Blair, Christensen, and Rudkin (2021)

Blair, Christensen, and Rudkin (2021) conduct a retrospective meta-analysis of 37 studies on commodity shocks and conflict. These studies are observational, though most seek to estimate an average treatment effect on the treated (ATT) of price shocks on conflict. This is a retrospective meta-analysis drawing on studies published between 2010 and 2020. Note that the measures of conflict and the contrasts are quite different across studies. The authors normalize estimates for target-equivalence, a step which we have omitted in the characterization of the studies. Note however, that such a normalization cannot address issues of limited comparability, as defined by our paper.

Attribute		Idrobo et al. (2014)	Parker and Vadheim (2017)
Setting	θ	Colombia. The panel considers violence or conflict events over time and space in Colombia.	Democratic Republic of Congo. The panel considers conflict events over time and space in the DRC.
Contrast	ω''	Price of gold. This is ultimately a continuous instrument, though ω'' could be thought of as a high price of gold.[5]	Three provinces in eastern DRC affected by the 2010 Dodd-Frank Act, which discouraged manufacturers from sourcing tin, tungsten, and tantalum from those regions.
	ω'	Price of gold. This is ultimately a continuous instrument, though ω' could be thought of as a low price of gold.	Provinces in eastern DRC that were not targeted by the 2010 Dodd-Frank Act.
Measurement strategy	m	Homicide rates, massacres, or forced displacement rate at the municipality-quarter level.	Uses ACLED data to code indicators for "looting" and for "battles" at the territory-month level.
Estimand	τ	An ATT of gold prices identified by a differences-in-differences design.	An ATT of the Dodd Frank act identified by a differences-in-differences design.

Table S7: Two randomly selected studies from Blair, Christensen, and Rudkin (2021).

Potential harmonization concerns:

ω'': Increased prices of gold (induced by the recession) are different from price shocks induced by the Dodd-Frank Act regulations.

ω': Little information is provided to characterize the setting at baseline (with low prices of gold or without Dodd-Frank Act regulations), though the minerals and the structure of the mining industry is presumably somewhat different in Colombia and the DRC.

m: The measures of violence are very different. Forced displacement and looting, for example, measure very different substantive phenomena.

S4.2 Incerti (2020)

Incerti (2020) meta-analyses both survey and field experiments on corruption information and vote choice. Given the overlap between the field experiments and the studies in Dunning, Grossman, Humphreys, Hyde, Mcintosh, Nellis, Adida, Arias, Bicalho, Boas, Buntaine, Chauchard, Chowdhury, Gottlieb, Hidalgo, Holmlund, Jablonski, Kramon, Larreguy, Lierl, Marshall, McClendon, Melo, Nielson, Pickering, Platas, QuerubÍn, Raffler, and Sircar (2019), we focus on the survey experimental meta-analysis. These studies generally provide survey respondents with some vignette about corruption of an incumbent or candidate in order to measure effects on vote intentions toward the candidate. This is a retrospective meta-analysis drawing on studies published between 2014 and 2020.

Attribute		Avenburg (2019)	Banerjee et al. (2014)
Setting	θ	4,894 Brazilian respondents recruited using Facebook, though most analyses are conducted on the 1,506 respondents that passed informational screener questions. The dates of the experiment are not clear.	5,105 male respondents from rural Sitapur, Uttar Pradesh India. The experiment was fielded in 2010.
Contrast	ω''	Vignette provides information that the candidate has accounts rejected by the Audit Court **and** information on the Audit Courts procedures and mechanisms leading to that decision.	Vignette states: "It is common knowledge that the candidate has accepted a bribe of Rs 10/20 lakh from a contractor." Vignette also varies election type, caste, and party.
	ω'	Vignette provides information that the candidate has accounts rejected by the Audit Court without further information.	Vignette states: "The candidate has a reputation for honesty." Vignette also varies election type, caste, and party.
Measurement strategy	m	Binary indicator for self-reported vote choice for candidate in vignette.	Binary indicator for self-reported vote choice for candidate.
Estimand	τ	ATE of corrupt candidate with procedural information vs. corrupt candidate without procedural information.	ATE of corrupt vs. honest candidate.

Table S8: Two randomly selected studies from Incerti (2020). The characterization of Avenburg (2019) focuses on the "procedural vignette" treatment condition among three treatment arms that provide additional information in addition to the basic corruption vignette. The characterization of Banerjee et al. (2014) focuses on the "strong" not the "weak" corruption vignette. It is unclear precisely what contrasts are analyzed in Incerti (2020).

Potential harmonization concerns:

ω'': The description of the corrupt actions of a hypothetical politician are different across the vignettes in the two experiments. It is hard to qualitatively or quantitatively assess the comparative severity of these actions.

ω': In Avenburg (2019), the "control" condition is a corrupt politician (with less detailed accusations). In Banerjee et al. (2014), the "control" condition is an honest politician. As such, the comparison is between two corrupt politicians in Avenburg (2019) but between a corrupt and an honest politician in Banerjee et al. (2014).

S4.3 Schwartz and Coppock (2020)

Schwarz and Coppock (2022) consider 67 preference elicitation survey experiments on candidate gender. All experiments vary the gender of hypothetical or real candidates in order to estimate the effects of gender on respondent support for a candidate. This is a retrospective meta-analysis drawing on studies published or written between 1984 and 2020 on six continents.

Attribute		Fox and Smith (1998): UCSB study	Wüest and Pontusson (2017)
Setting	θ	Sample of 173 University of California Santa Barbara students in the late 1990s. Respondents viewed hypothetical candidates for an unspecified election.	Sample of 4,500 Swiss citizens of voting age in 2017. Respondents viewed hypothetical candidates for the Swiss National Council.
Contrast	ω''	Hypothetical candidate is female. This is conveyed by candidate names. The ideology (liberal, moderate, or conservative) was also varied independently of gender.	Hypothetical candidate is female. In the conjoint setting, the other manipulated attributes were: wealth (salary), education, occupation, experience, and residence in respondent's canton.
	ω'	Hypothetical candidate is male. This is conveyed by candidate names. The ideology (liberal, moderate, or conservative) was also varied independently of gender.	Hypothetical candidate is male. In the conjoint setting, the other manipulated attributes were: wealth (salary), education, occupation, experience, and residence in respondent's canton.
Measurement strategy	m	0-100 feeling thermometer converted to binary choice among a pair of candidates.	Forced binary vote choice among pair of two candidates.
Estimand	τ	ATE of female vs. male candidate on vote choice.	ATE (AMCE) of female vs. male candidate on vote choice.

Table S9: Two randomly selected studies from Schwarz and Coppock (2022).

Potential harmonization concerns:

- ω'': Conveying that a candidate is female though her name versus a direct statement of gender may constitute a different treatment.
- ω': Conveying that a candidate is male though his name versus a direct statement of gender may constitute a different treatment.
- m: While Schwarz and Coppock (2022) compare binary vote choice, the conversion from a feeling thermometer to a binary choice may be different than direct elicitation of vote choice.

Supplementary Appendix: References

Adcock, Robert, and David Collier. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* 95 (3): 529–546.

Avenburg, Alejandro. 2019. "Public Costs versus Private Gain: Assessing the Effect of Different Types of Information about Corruption Incidents on Electoral Accountability." *Journal of Politics in Latin America* 11 (1): 71–108.

Banerjee, Abhijit, Donald P. Green, Jeffery McManus, and Rohini Pande. 2014. "Are Poor Voters Indifferent to Whether Elected Leaders Are Criminal or Corrupt? A Vignette Experiment in Rural India." *Political Communication* 31 (3): 391–407.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry about Sensitivity Bias: A Social Referent Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114 (4): 1297–1315.

Blair, Graeme, Darin Christensen, and Aaron Rudkin. 2021. "Do Commodity Price Shocks Cause Armed Conflict? A Meta-Analysis of Natural Experiments." *American Political Science Review* 115 (2): 1–8.

Blair, Graeme, Jeremy M. Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman, Dotan Haim, Zulfiqar Hameed, Rebecca Hanson, Ali Hasanain, Dorothy Kronick, Benjamin S. Morse, Robert Muggah, Fatiq Nadeem, Lily L. Tsai, Matthew Nanes, Tara Slough, Nico Ravanilla, Jacob N. Shapiro, Barbara Silva, Pedro C. L. Souza, and Anna M. Wilke. 2021. "Community policing does not build citizen trust in police or reduce crime in the Global South." *Science* 374 (6571): eabd3446.

Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6 (eabc4046): 1–6.

de la O, Ana, Donald P. Green, Peter John, Rafael Goldszmidt, Anna-Katharina Lenz, Martin Valdivia, Cesar Zucco, Darin Christensen, Francisco Garfiras, Pablo Balán, Augustin Bergeron, Gabriel Tourek, Jonathan Weigel, Jessica Gottlieb, Adrienne LeBas, Janica Magat, Nonso Obikili, Jake Bowers, Nuole Chen, Christopher Grady, Matthew Winters, Nikhar Gaikwad, Gareth Nellis, Anjali Thomas, and Susan Hyde. 2021. "Fiscal Contracts? A Six-country Randoized Experiment on Transaction Costs, Public Services, and Taxation in Developing Countries." Working paper.
URL: `https://nikhargaikwad.com/resources/De-La-O-et-al_2021.pdf`

Dunning, Thad. 2012. *Natural experiments in the social sciences: a design-based approach.* Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig Mcintosh, Gareth Nellis, Claire L. Adida, Eric Arias, Clara Bicalho, Taylor C. Boas, Mark T. Buntaine, Simon

Chauchard, Anirvan Chowdhury, Jessica Gottlieb, Daniel F. Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Daniel L. Nielson, Paula M. Pickering, Melina R. Platas, Pablo QuerubÍn, Pia Raffler, and Neelanjan Sircar. 2019. "Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials." *Science Advances* 5 (7).

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* New York: Cambridge University Press.

Eshima, Shusei, and Daniel M. Smith. 2022. "Just a Number? Voter Evaluations of Age in Candidate Choice Experiments." *Journal of Politics* Forthcoming.

Fox, Richard L., and Eric R. A. N. Smith. 1998. "The Role of Candidate Sex in Voter Decision-Making." *Political Psychology* 19 (2): 405–419.

Godefroidt, Amélie. 2021. "How Terrorism Does (and Does Not) Affect Citizens' Political Attitudes: A Meta-Analysis." *American Journal of Political Science* forthcoming.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/ajps.12692*

Guillemin, Victor, and Alan Pollack. 1974. *Differential topology*. AMS Chelsea Publishing.

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114 (3): 761–774.

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–166.

Kertzer, Joshua D. 2020. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* forthcoming.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/ajps.12583*

Parker, Dominic P., and Bryan Vadheim. 2017. "Resource Cursed or Policy Cursed? US Regulation of Conflict Minerals and Violence in the Congo." *Journal of the Association of Environmental and Resource Economists* 4 (1): 1–49.

Rosenbaum, Paul R. 2002. *Observational Studies*. Second edition ed. New York: Springer.

Schwarz, Susanne, and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84 (2).

Slough, Tara, Daniel Rubenson, Ro'ee Levy, Francisco Alpizar Rodriguez, María Bernedo del Carpio, Mark T. Buntaine, Darin Christensen, Alicia Cooperman, Sabrina Eisenbarth, Paul J.

Ferraro, Louis Graham, Alexandra C. Hardman, Jacob Kopas, Sasha McLarty, Anouk S. Rigterink, Cyrus Samii, Brigitte Seim, Johannes Urpelainen, and Bing Zhang. 2021. "Adoption of Community Monitoring Improves Common Pool Resource Management Across Contexts." *Proceedings of the National Academy of Sciences* 10.1073: 1–10.

Tu, Loring W. 2011. *An Introduction to Manifolds*. Springer.

Wüest, Reto, and Jonas Pontusson. 2017. "Do Citizens Prefer Affluent Representatives? Evidence from a Survey Experiment In Switzerland.".
URL: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3077598`