Online Appendix: When Do Surveys Produce False Positives for Regime Support?

Jeffrey Stark

2026-03-01

A Variable Definitions and Coding

Table A1 presents the full set of variables used in the analysis, with metadata extracted directly from the YAML harmonization specifications used in the data pipeline.

Table A1: Variable definitions, scales, and coding for al
Metadata extracted from YAML harmonization specificat
the full Hong Kong Wave 5 sample.

Variable	ABS W5 Item	Response Scale
Alternative Explanations \& Controls		
Trust in police	q14	1–4 (None at all to A great deal of trust)
Trust in national government	q9	1–4 (None at all to A great deal of trust)
Trust in president/CE	q7	1–4 (None at all to A great deal of trust)
Trust in parliament	q11	1–4 (None at all to A great deal of trust)
Trust in courts	q8	1–4 (None at all to A great deal of trust)
Trust in civil service	q12	1–4 (None at all to A great deal of trust)
Democracy always preferable	q132	1–3 (Democracy is always preferable to Doesn't matter what kind of regime)
Democratic satisfaction	q99	1–4 (Not at all satisfied to Very satisfied)
Democracy suitability	q104	1–10 (Not suitable at all to Completely suitable)

(continued)

Variable	ABS W5 Item	Response Scale
Authoritarian Alternatives		
Extent of current democracy	q100	1–10 (Complete dictatorship to Complete democracy)
Current government dem.\ rating	q101	1–10 (Complete dictatorship to Complete democracy)
Freedom to organize	q116	1–4 (Strongly disagree to Strongly agree)
Democratic Attitudes		
Freedom of speech	q115	1–4 (Strongly disagree to Strongly agree)
Elections free and fair	q38	1–4 (Not free or fair to Completely free and fair)
Rich and poor treated equally	q113	1–4 (Strongly disagree to Strongly agree)
Opposition opportunities	q109	1–4 (Strongly disagree to Strongly agree)
System deserves support	q88	1–4 (Strongly disagree to Strongly agree)
Governance Perceptions		
System pride	q87	1–4 (Strongly disagree to Strongly agree)
System needs change	q90	1–4 (Works fine, no change needed to Should be replaced)
Strongman rule	q137	1–4 (Strongly disapprove to Strongly approve)
Military rule	q139	1–4 (Strongly disapprove to Strongly approve)
Single-party rule	q138	1–4 (Strongly disapprove to Strongly approve)
Institutional Trust		
Political interest	q46	1–4 (Not at all interested to Very interested)
Efficacy: ability to participate	q141	1–4 (Strongly disagree to Strongly agree)
Efficacy: no influence	q143	1–4 (Strongly disagree to Strongly agree)
Demonstration attendance	q79	1–5 (Would never do to Have done >3 times)
China democratic rating	q128	1–10 (Completely undemocratic to Completely democratic)
Satisfaction w/ president/govt	q105	1–4 (Very dissatisfied to Very satisfied)
Political Efficacy \& Participation		
Democratic future (10 yrs)	q103	1–10 (Complete dictatorship to Complete democracy)
Essential element of democracy	q98	1–5 (Electoral (free and fair elections) to Good governance (no waste/corrup

(continued)

Variable	ABS W5 Item	Response Scale
Procedural vs.\ substantive	q98	0–1 (Substantive (redistribution/needs/governance) to Procedural (elections
Willingness to emigrate	q167	1–4 (Not willing at all to Very willing)
System Support		
Democracy is best form	q136	1–4 (Strongly disagree to Strongly agree)
Courts powerless vs.\ leaders	q111	1–4 (Strongly disagree to Strongly agree)
Government responds to people	q121	1–4 (Not at all to Very well)

Note: N statistics refer to the full Hong Kong Wave 5 sample (all fieldwork periods). Scale descriptions reflect the harmonized codi

B Mann-Whitney U Tests

All two-sample t-tests reported in the main text were cross-validated with non-parametric Mann-Whitney U (Wilcoxon rank-sum) tests. This guards against violations of normality assumptions, which are plausible given that many indicators are measured on short ordinal scales (1–4). Table B2 reports both test results side by side.

Table B2: Parametric (t-test) vs. non-parametric (Mann-Whitney U) comparison for al pre/post NSL variables.

Variable	Cohen's d	t p-value		MW p-value		Agree?
democracy_suitability	-0.536	9.62e-07	***	2.77e-05	***	TRUE
dem_extent_current	0.221	0.008743	**	0.003220	**	TRUE
dem_country_present_govt	-0.231	0.000756	***	0.010993	*	TRUE
dem_always_preferable	0.360	4.60e-09	***	3.00e-08	***	TRUE
democracy_satisfaction	0.100	0.112980		0.070960		TRUE

trust_national_government	0.428	2.22e-11	***	3.78e-11	***	TRUE
trust_president	0.377	3.54e-10	***	7.93e-10	***	TRUE
trust_parliament	0.234	0.000138	***	0.000172	***	TRUE
trust_police	0.432	3.82e-12	***	8.19e-12	***	TRUE
trust_military	0.390	2.59e-08	***	3.72e-08	***	TRUE
trust_courts	0.030	0.620259		0.713226		TRUE
gov_free_to_organize	-0.427	6.60e-12	***	5.48e-11	***	TRUE
dem_free_speech	0.408	5.18e-11	***	2.07e-10	***	TRUE
govt_responds_people	-0.436	3.87e-12	***	8.26e-13	***	TRUE
election_free_fair	0.234	0.001732	**	0.004114	**	TRUE
system_deserves_support	-0.182	0.004224	**	0.007508	**	TRUE
system_proud	0.112	0.079234		0.108505		TRUE
system_needs_change	-0.243	0.000108	***	0.000199	***	TRUE
strongman_rule	0.149	0.018812	*	0.008756	**	TRUE
military_rule	0.089	0.149836		0.106124		TRUE
single_party_rule	0.065	0.297797		0.205856		TRUE

Note: Cohen's d computed as (Post-NSL − Protest) / pooled SD; positive values indicate higher post-N

C Demographic Balance Across Periods

A key concern with the within-wave quasi-experiment is that the Protest and Post-NSL sub-samples may differ systematically on demographic covariates, which could account for

some of the observed attitudinal shifts. Table C3 reports means or proportions for age, gender, and education across the two periods, along with t-tests (continuous) or χ^2 tests (categorical) for between-period differences.

Table C3: Demographic balance across fieldwork
standardized mean difference.

Covariate	Protest	Post-NSL	SMD	Test
Age (mean)	50	49.8	0.013	t-test
Gender (% female)	58.6%	55.6%	0.059	Chi-squared
Education level (SE5, ordinal 1-10)	6.09	5.89	0.089	t-test
Age 18-29	16.1%	17.3%	NA	Chi-squared
Age 30-39	14.1%	11.6%	NA	
Age 40-49	17.7%	16.4%	NA	
Age 50-59	18.4%	18.6%	NA	
Age 60+	33.6%	36.1%	NA	

Note: SMD computed as (Protest − Post-NSL) / pooled SD for continuous variables and as the differe

D Stratified Bootstrap and Manski Bounds

D.1 Stratified Percentile Bootstrap

To assess the robustness of the pre/post NSL shifts to sampling variability, I employed a stratified percentile bootstrap with 5,000 iterations, resampling independently within the Protest and Post-NSL periods to preserve the temporal structure of the quasi-experiment.

Table D4 reports the observed Cohen's d alongside 95% bias-corrected and accelerated (BCa) confidence intervals.

Table D4: Stratified bootstrap results (5,000 iterations): BCa 95% confidence intervals fo Cohen's d.

Variable	Observed d	d CI Lo	d CI Hi	CI Type	Sig?
democracy_suitability	-0.536	-0.765	-0.316	BCa	TRUE
dem_extent_current	0.221	0.049	0.387	BCa	TRUE
dem_country_present_govt	-0.231	-0.368	-0.097	BCa	TRUE
dem_always_preferable	0.360	0.235	0.474	BCa	TRUE
democracy_satisfaction	0.100	-0.030	0.217	BCa	FALSE
trust_national_government	0.428	0.300	0.557	BCa	TRUE
trust_president	0.377	0.258	0.499	BCa	TRUE
trust_parliament	0.234	0.114	0.351	BCa	TRUE
trust_police	0.432	0.307	0.553	BCa	TRUE
trust_military	0.390	0.245	0.531	BCa	TRUE
trust_courts	0.030	-0.092	0.148	BCa	FALSE
gov_free_to_organize	-0.427	-0.548	-0.305	BCa	TRUE
dem_free_speech	0.408	0.285	0.528	BCa	TRUE
govt_responds_people	-0.436	-0.558	-0.312	BCa	TRUE
election_free_fair	0.234	0.081	0.380	BCa	TRUE
system_deserves_support	-0.182	-0.307	-0.059	BCa	TRUE
system_proud	0.112	-0.015	0.237	BCa	FALSE

system_needs_change	-0.243	-0.368	-0.120	BCa	TRUE
strongman_rule	0.149	0.025	0.270	BCa	TRUE
military_rule	0.089	-0.027	0.218	BCa	FALSE
single_party_rule	0.065	-0.059	0.188	BCa	FALSE

Note: Cohen's d computed as (Post-NSL − Protest) / pooled SD; positive values indicate higher post-N

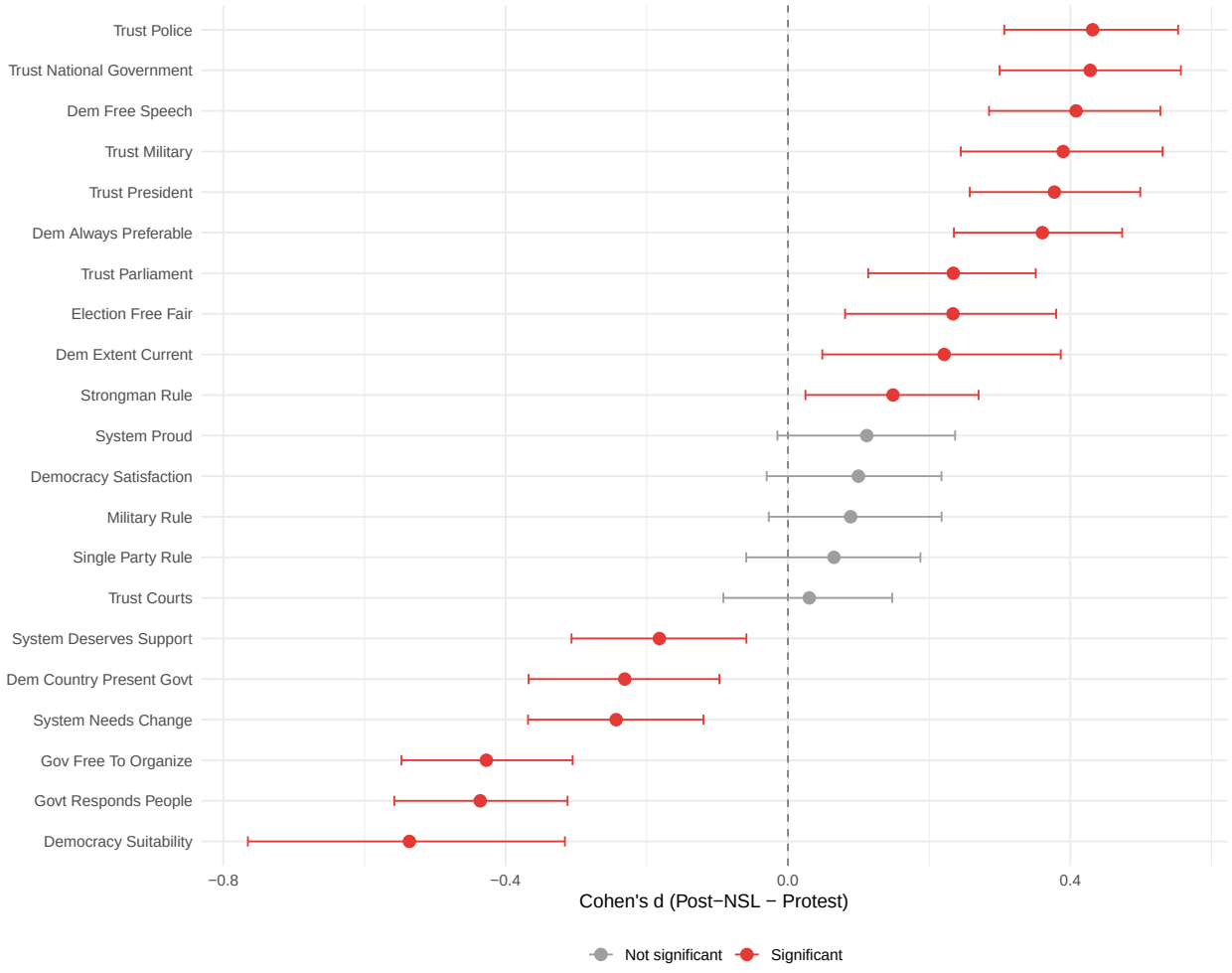

Figure D1: Stratified bootstrap 95% BCa confidence intervals for Cohen's d (Post-NSL − Protest). Positive values indicate higher post-NSL responses. Variables where the CI excludes zero are shown in colour.

D.2 Informal Bounding Analysis

A key concern with the trust increase is that compositional selection — whereby regime critics disproportionately declined to participate in the post-NSL survey — could entirely account for the observed shift. To assess the plausibility of this account, I conduct an informal bounding exercise in the spirit of Manski (2003).

The logic is as follows: if we assume that all non-respondents in the post-NSL period would have reported the minimum possible trust score (1 on the 1–4 scale), what proportion of the target population would need to have been "missing critics" for the trust increase to disappear? I present two scenarios: a worst-case bound (imputing the scale minimum) and a more empirically grounded scenario (imputing the 25th percentile of the protest-period distribution, representing respondents drawn from the bottom quartile of critics). Table D5 and Table D6 present the results for trust in police and trust in the national government, respectively.

Table D5: Bounding analysis for trust in police: worst-c
quartile (impute 25th percentile of protest distribution)

% Missing	N Obs	N Missing	Adj Mean (min)	Still > Protest? (min)	Adj Mean (Q25)	Still > P
0	637	0	2.641	TRUE	2.641	TRUE
5	637	34	2.557	TRUE	2.557	TRUE
10	637	71	2.476	TRUE	2.476	TRUE
15	637	112	2.395	TRUE	2.395	TRUE
20	637	159	2.313	TRUE	2.313	TRUE

25	637	212	2.231	TRUE	2.231	TRUE
30	637	273	2.148	TRUE	2.148	TRUE
35	637	343	2.066	FALSE	2.066	FALSE
40	637	425	1.984	FALSE	1.984	FALSE

Note: Protest-period mean: 2.147. Worst-case imputes scale minimum (1); Q25 imputes the 25th perce

Table D6: Bounding analysis for trust in national govern
and bottom-quartile (impute 25th percentile of protest

% Missing	N Obs	N Missing	Adj Mean (min)	Still > Protest? (min)	Adj Mean (Q25)	Still > P
0	617	0	2.634	TRUE	2.634	TRUE
5	617	32	2.553	TRUE	2.553	TRUE
10	617	69	2.469	TRUE	2.469	TRUE
15	617	109	2.388	TRUE	2.388	TRUE
20	617	154	2.307	TRUE	2.307	TRUE
25	617	206	2.225	TRUE	2.225	TRUE
30	617	264	2.144	FALSE	2.144	FALSE
35	617	332	2.062	FALSE	2.062	FALSE
40	617	411	1.981	FALSE	1.981	FALSE

Note: Protest-period mean: 2.154. Worst-case imputes scale minimum (1); Q25 imputes the 25th perce

The results suggest that the observed trust increase is robust to all but implausibly extreme selection scenarios. Under worst-case imputation (all non-respondents scored at the scale

minimum), the post-NSL trust mean for police remains above its protest-period baseline until approximately 35–40% of the target population is assumed to consist of missing extreme critics. Under the more empirically grounded bottom-quartile scenario, the trust increase persists even at higher assumed non-response rates. Selection alone would therefore need to reach implausibly extreme levels—removing over one-third of the survey-eligible population from the respondent pool, all of whom would have to hold the most negative possible view—to fully account for the observed shift.

D.3 Ruling Out Differential Baseline Distrust

A compositional selection account offers a plausible alternative to the preference falsification interpretation: if protest-aligned respondents disproportionately exited the survey frame following NSL implementation, and if those respondents happened to distrust coercive institutions more than they distrusted political or democratic institutions at baseline, observed gradient-consistent shifts could reflect compositional change rather than strategic misreporting. We test this directly.

The selection account generates a specific prediction: items with greater baseline variance in trust—reflecting deeper pre-existing disagreement—should shift more sharply after NSL implementation, as the critics most likely to exit were clustered at the low end of these items. We operationalize baseline variance as the pre-period standard deviation for each survey item and test whether it independently predicts the magnitude of post-NSL shifts (Cohen's d) after accounting for sensitivity rank.

We conduct the analysis at the item level ($n = 6$: five individual trust items plus a democratic

attitudes index formed from three z-scored items re-expressed in trust-scale units).[1] Cross-item regressions regress Cohen's d on sensitivity rank (M1), baseline SD (M2), and both jointly (M3). Results are summarized in Table D7 and Table D8.

Item	Sensitivity rank	Protest-period SD
Trust in Police	1	1.158
Trust in Nat'l Government	1	1.151
Trust in President/CE	2	0.992
Trust in Parliament	3	0.989
Trust in Military	3	1.172
Trust in Courts	4	1.018
Democratic Attitudes Index	5	0.689

Note: Sensitivity rank: 1 = most coercive/sensitive (police, national government), 5 = least sensitive (c

The results are unambiguous. Sensitivity rank alone explains 79% of cross-item variance in post-NSL shifts (M1: $\beta = -0.094$, $p = .007$). Adding baseline SD to the model leaves sensitivity rank essentially unchanged ($\beta = -0.102$, $p = .053$) while contributing no independent explanatory power (M2 baseline SD coefficient: $\beta = -0.088$, $p = .80$). The partial correlation for sensitivity rank, controlling for baseline SD, is $r = -0.805$ ($p = .029$);

[1]The democratic attitudes index averages z-scored versions of democracy suitability, democracy always preferable, and democratic satisfaction. Collapsing these three rank-5 items to a single index is theoretically motivated—all three are "low-sensitivity abstract evaluations" in the framework's typology—and avoids pseudo-replication from treating operationalizations of the same construct as independent observations.

the partial correlation for baseline SD, controlling for rank, is $r = -0.062$ ($p = .14$). The Spearman rank correlation between sensitivity and shift is $\rho = -0.946$ ($p = .001$).

The selection account's key prediction—that baseline variance should positively predict post-NSL shifts—is directly contradicted. The clearest discriminating observation involves police and courts. These two items have nearly identical pre-period standard deviations (police SD = 1.16; courts SD = 1.02), indicating similar degrees of baseline disagreement. Yet their post-NSL shifts diverge dramatically: police confidence increased by $d = 0.46$, while court confidence changed by only $d = 0.05$. If baseline heterogeneity were driving the gradient, items with similar variance should shift similarly. Instead, the difference precisely tracks coercive sensitivity: police represent the front line of NSL enforcement; courts retained meaningful judicial independence. Baseline distrust cannot explain this divergence; institutional sensitivity can.

Table D8: Cross-item regression: sensitivity rank and baseline disagreement
Cohen's d. M1: sensitivity rank alone. M2: baseline SD alone. M3: both jo
M4: M3 with standardized predictors.

Model	Term	$b	$SE	$t	$p
M1: Sensitivity rank only	Intercept	0.565	0.066	8.582	0.000354
M1: Sensitivity rank only	Sensitivity rank	-0.094	0.022	-4.366	0.007248
M2: Baseline SD only	Intercept	-0.315	0.332	-0.948	0.386622
M2: Baseline SD only	Baseline SD	0.609	0.321	1.900	0.115822
M3: Both jointly (key test)	Intercept	0.676	0.426	1.586	0.187911
M3: Both jointly (key test)	Sensitivity rank	-0.102	0.038	-2.715	0.053247

M3: Both jointly (key test)	Baseline SD	-0.088	0.333	-0.264	0.804805
M4: Standardized	Intercept	0.000	0.209	0.000	1.000000
M4: Standardized	Sensitivity rank (std.)	-0.962	0.354	-2.715	0.053247
M4: Standardized	Baseline SD (std.)	-0.094	0.354	-0.264	0.804805

Note: $n = 6$ observations (5 trust items + democratic attitudes index). Model fit: M1 adj. $R^2 = 0.751$

A supplementary analysis disaggregating the democratic attitudes index into its three component items ($n = 9$) confirms the main finding ($\rho = -0.846$, $p = .004$) and yields an additional piece of evidence. *Democracy Suitability* has the highest baseline SD in the battery (2.98)—indicating deep pre-existing disagreement—yet posts the sharpest *negative* shift ($d = -0.522$). Under the selection account, high baseline variance should predict upward shift (as critics exit); instead, this item moves in the wrong direction entirely. The selection account cannot accommodate this pattern. The divergence of *Democracy Suitability* from other democratic items is consistent with item asymmetry: evaluating whether democracy "suits" Hong Kong implicitly references regime performance, making falsification psychologically costly in a way that endorsing democracy as an abstract ideal ("democracy is always preferable") is not (see Section 2.2 of the main manuscript for the theoretical argument).

Taken together, these tests support the conclusion that the sensitivity gradient reflects strategic response processes rather than differential attrition of baseline critics.

E Supplementary Figures

E.1 Age-Stratified Pre/Post NSL Shifts

Figure E2 presents the pre/post NSL mean comparison within each age cohort for trust in police, trust in government, democracy suitability, and freedom to organize. If the trust increase were driven entirely by compositional selection — specifically, by younger, more critical respondents declining to participate in the post-NSL period — we would expect the shift to appear primarily or exclusively in younger age groups. The figure shows that while the magnitude of change varies across cohorts, the direction of the trust increase is broadly consistent across age groups, which is more consistent with preference falsification operating across the population than with a purely compositional account.

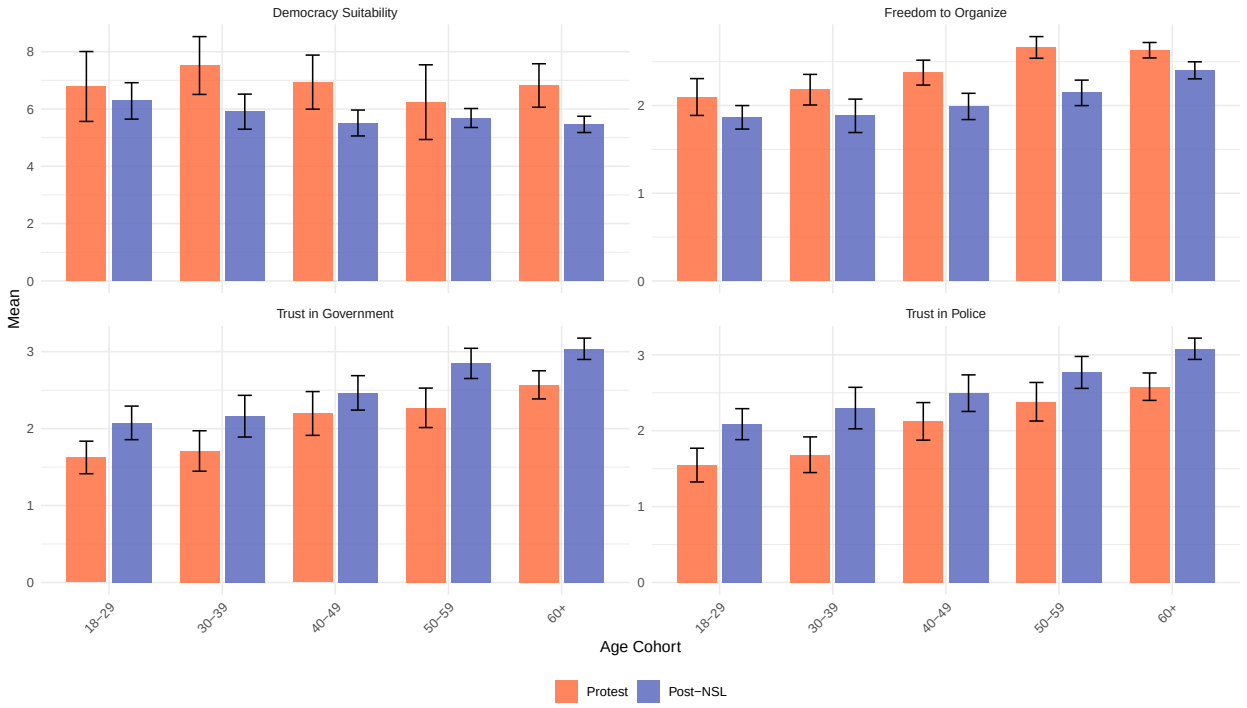

Figure E2: Pre/Post NSL mean shifts within age cohorts for key indicators. Error bars show 95% confidence intervals.

F Falsification Tests and Multiple Comparisons

F.1 Placebo-Adjacent Items

To assess whether the observed attitudinal shifts are specifically attributable to politically sensitive mechanisms rather than generalized confounding (e.g., pandemic effects, time trends), I pre-specified a set of items with lower expected political sensitivity. Table F9 reports the pre/post NSL comparison for these placebo-adjacent items.

Table F9: Placebo-adjacent items: pre-specified items with lower expected political sensitivity Small or non-significant effects are consistent with NSL-specific mechanisms rather than generalized confounding.

Variable	Protest Mean	Post-NSL Mean	Delta	Cohen's d	p-value	Sig
rich_poor_treated_equally	2.294	2.285	-0.008	-0.011	0.85304	ns
political_interest	2.216	2.091	-0.124	-0.169	0.00564	**
trust_courts	2.675	2.728	0.053	0.053	0.38560	ns

Note: Cohen's d computed as (Post-NSL − Protest) / pooled SD; positive values indicate higher post-

F.2 Benjamini-Hochberg FDR Correction

The main analysis tests multiple variables simultaneously. To guard against inflated Type I error rates, Table F10 reports raw and Benjamini-Hochberg adjusted p-values for all pre/post NSL comparisons. The four pre-specified primary outcomes are flagged.

Table F10: Raw and FDR-corrected (Benjamini-Hochberg) p-values for all pre/post NSL comparisons.

Variable	Primary?	Cohen's d	p (raw)		p (BH-adj)	
govt_responds_people	FALSE	-0.480	2.58e-14	***	7.47e-13	***
trust_police	TRUE	0.460	2.67e-13	***	3.87e-12	***
trust_national_government	TRUE	0.451	3.19e-12	***	3.08e-11	***
trust_president	FALSE	0.404	6.86e-11	***	4.98e-10	***
dem_always_preferable	FALSE	0.364	1.23e-08	***	7.09e-08	***
dem_free_speech	FALSE	0.360	1.47e-08	***	7.09e-08	***
gov_free_to_organize	TRUE	-0.357	2.21e-08	***	9.15e-08	***
trust_military	FALSE	0.379	4.41e-08	***	1.60e-07	***
democracy_suitability	TRUE	-0.522	7.21e-08	***	2.32e-07	***
trust_parliament	FALSE	0.273	1.09e-05	***	3.17e-05	***
system_needs_change	FALSE	-0.263	5.09e-05	***	0.000134	***
election_free_fair	FALSE	0.284	8.64e-05	***	0.000209	***
system_deserves_support	FALSE	-0.248	9.67e-05	***	0.000216	***
pol_discuss	FALSE	-0.214	0.000467	***	0.000927	***
efficacy_no_influence	FALSE	-0.215	0.000479	***	0.000927	***
efficacy_ability_participate	FALSE	-0.206	0.001143	**	0.002073	**
dem_country_present_govt	FALSE	-0.218	0.001551	**	0.002646	**
nat_willing_emigrate	FALSE	-0.184	0.004440	**	0.007154	**
political_interest	FALSE	-0.169	0.005640	**	0.008608	**
strongman_rule	FALSE	0.157	0.013091	*	0.018982	*

dem_extent_current	FALSE	0.208	0.014007	*	0.018982	*
system_proud	FALSE	0.155	0.014400	*	0.018982	*
democracy_satisfaction	FALSE	0.143	0.025456	*	0.032097	*
gov_courts_powerless	FALSE	0.139	0.032766	*	0.039592	*
military_rule	FALSE	0.096	0.120134		0.139355	
single_party_rule	FALSE	0.091	0.145502		0.162291	
trust_courts	FALSE	0.053	0.385596		0.414159	
gov_opposition_opportunities	FALSE	0.025	0.712385		0.737827	
rich_poor_treated_equally	FALSE	-0.011	0.853037		0.853037	

Note: Cohen's d computed as (Post-NSL − Protest) / pooled SD; positive values indicate higher post-N

G Critical Citizens Correlation: CI and Period Decomposition

The main text reports that Hong Kong is the only country among the sixteen ABS Wave 5 polities where the correlation between democratic preference ("democracy is always preferable") and democratic satisfaction is positive. This section reports the precise estimate with 95% confidence intervals and decomposes the correlation by fieldwork period.

Table G11: Critical citizens correlation (democratic preference × democratic satisfaction): overall and by fieldwork period, with 95% confidence intervals.

Period	N	r	95% CI Lo	95% CI Hi	p-value

Overall (Wave 5)	980	0.279	0.221	0.336	< 2e-16
Protest period	410	0.184	0.089	0.276	0.000179
Post-NSL period	534	0.328	0.250	0.401	7.7e-15

The positive correlation is concentrated in the post-NSL period, consistent with a mechanism that operates more strongly under authoritarian constraint.

G.1 Cross-National Comparison

Table G12 reports the Pearson correlation between democratic preference ("democracy is always preferable") and democratic satisfaction for all fifteen ABS Wave 5 polities with sufficient data ($N \geq 50$).

Table G12: Critical citizens correlation (democratic preference × democratic satisfaction) across ABS Wave 5 polities, sorted by correlation magnitude.

Polity	N	r	95% CI Lo	95% CI Hi	p-value
Hong Kong	**980**	**0.279**	**0.221**	**0.336**	**< 2e-16**
Singapore	951	0.013	-0.051	0.076	0.69702
Malaysia	1197	-0.049	-0.106	0.007	0.08797
Myanmar	1291	-0.052	-0.107	0.002	0.05934
Indonesia	1218	-0.060	-0.116	-0.004	0.03671
Mongolia	1144	-0.060	-0.118	-0.002	0.04129
Vietnam	1026	-0.074	-0.135	-0.013	0.01725
China	3468	-0.081	-0.114	-0.048	1.72e-06

Philippines	1156	-0.095	-0.152	-0.038	0.00118
India	3988	-0.102	-0.133	-0.071	1.00e-10
Thailand	959	-0.134	-0.196	-0.071	3.16e-05
South Korea	1190	-0.157	-0.212	-0.101	5.17e-08
Japan	920	-0.177	-0.239	-0.114	6.57e-08
Taiwan	1185	-0.184	-0.239	-0.129	1.65e-10
Australia	1572	-0.187	-0.234	-0.139	7.41e-14

Hong Kong is the only polity with a positive correlation ($r = 0.279$). Singapore shows a near-zero positive value ($r = 0.013$, $p = .70$); all other polities are negative, with established democracies (Australia, Taiwan, Japan, South Korea) showing the strongest negative correlations as expected by the critical citizens framework.

G.2 Decomposition by Democratic Conception Type

Table G13 decomposes the critical citizens correlation by respondents' conception of democracy, using the ABS forced-choice item asking which element is most essential to democracy. Respondents are classified as holding procedural conceptions (free expression or free elections) or substantive conceptions (basic necessities or clean governance).

Table G13: Critical citizens correlation decomposed by democratic conception type and fieldwork period.

Period	Conception	N	r	95% CI Lo	95% CI Hi	p-value
Protest	Substantive	211	0.059	-0.077	0.192	0.398

Protest	Procedural	193	0.095	-0.047	0.233	0.190
Post-NSL	Substantive	231	0.056	-0.074	0.184	0.398
Post-NSL	Procedural	289	0.355	0.250	0.452	5.4e-10

The positive correlation is concentrated entirely among respondents with procedural conceptions of democracy ($r = 0.355$, $p < .001$). Respondents with substantive conceptions show near-zero correlations in both periods. A Fisher z-test confirms the difference between conception groups in the post-NSL period is statistically significant ($z = 3.55$, $p < .001$). See Sections 4.6 and 5.3 of the main text for discussion.

G.3 Fine-Grained Decomposition by Essential Element

Table G14 provides a finer decomposition using the four specific response options.

Table G14: Critical citizens correlation by specific essential element of democracy and fieldwork period.

Period	Essential Element	N	r	95% CI Lo	95% CI Hi	p-value
Protest	Free expression	123	0.092	-0.087	0.265	0.312
Protest	Free elections	70	0.089	-0.149	0.318	0.462
Protest	Basic necessities	157	0.060	-0.098	0.214	0.459
Protest	Clean governance	54	0.020	-0.249	0.287	0.884
Post-NSL	Free expression	150	0.321	0.169	0.458	6.25e-05
Post-NSL	Free elections	139	0.421	0.273	0.549	2.5e-07
Post-NSL	Basic necessities	190	0.007	-0.135	0.149	0.921

Post-NSL	Clean governance	41	0.284	-0.026	0.544	0.0719

H Differential Item Non-Response

Table H15 reports item-level response rates for key survey variables, comparing the protest period and post-NSL period sub-samples.

Table H15: Item-level response rates by fieldwork period.

Item	Response % (Protest)	Response % (Post-NSL)	Δ (pp)
Political interest	99.2	97.8	-1.4
Willing to emigrate	89.0	88.0	-1.0
Rich/poor treated equally	97.7	97.2	-0.5
Extent of current democracy	54.8	45.3	-9.5
Democracy suitability	38.3	41.6	3.3
Democracy always preferable	93.9	88.8	-5.1
Democratic satisfaction	91.3	87.3	-4.0
Free to speak without fear	94.3	90.4	-3.9
Freedom to organize	93.0	89.9	-3.1
Elections free and fair	70.6	69.5	-1.1
System deserves support	93.2	87.0	-6.2
Trust: police	94.9	94.2	-0.7
Trust: courts	97.9	97.6	-0.3

Trust: parliament	96.2	96.0	-0.2
Trust: president/CE	98.3	98.4	0.1
Trust: civil service	96.2	96.9	0.7
Trust: national govt	89.0	91.3	2.3

Trust items maintain near-identical response rates across periods (mean change: +0.3 pp). Normative democracy items decline by a mean of 4.6 pp, governance items by 2.7 pp, while control items remain stable. This differential pattern is consistent with politically motivated non-response on items requiring explicit democratic commitments.

I Robust Standard Errors

To address potential heteroskedasticity and assess sensitivity to standard error computation, Table I16 reports covariate-adjusted estimates with both conventional OLS standard errors and heteroskedasticity-consistent (HC2) robust standard errors for the four primary outcomes.

Table I16: Covariate-adjusted Post-NSL effect estimates with con
standard errors.

Outcome	b	SE (conv)	SE (HC2)	p (conv)	p (HC2)
Trust in Police	0.491	0.068	0.068	1.49e-12	1.08e-12
Trust in National Government	0.469	0.069	0.069	1.8e-11	1.76e-11
Democracy Suitability	-1.181	0.218	0.247	1.06e-07	2.51e-06
Freedom to Organize	-0.312	0.044	0.044	4.25e-12	1.83e-12

Note: All models control for age, gender, edu_clean. HC2 standard errors are heteroskedasticity-consis

The HC2 robust standard errors are substantively indistinguishable from conventional OLS standard errors across all four primary outcomes, confirming that heteroskedasticity does not meaningfully affect inference.

J Cross-National Sensitivity Gradient: Turkey and Russia

The main analysis documents a sensitivity gradient in Hong Kong's ABS Wave 5 data: trust in coercive institutions surged post-NSL ($d \approx +0.49$ for police) while democratic evaluations collapsed ($d = -0.54$), with the magnitude of inflation tracking each institution's coercive role ($r = -0.85$ between pre-specified sensitivity rankings and observed effect sizes). Differential item non-response reinforced this pattern, with normative democracy items showing a mean response rate decline of 4.6 percentage points while trust items remained stable across fieldwork periods. A natural question is whether these patterns are specific to Hong Kong's particular political context and the ABS instrument, or whether the sensitivity gradient captures a more general phenomenon of survey measurement distortion under autocratization.

To assess this, we apply the same logic to two independent cases using an entirely different survey program: the World Values Survey (WVS). Turkey (Wave 6: 2011, Wave 7: 2018) experienced dramatic autocratization following the 2016 failed coup attempt, with mass purges of civil servants, academics, and journalists, declaration of a state of emergency, and

constitutional changes consolidating executive power. Russia (Wave 6: 2011, Wave 7: 2017) underwent intensified authoritarian consolidation following the 2014 Crimea annexation, including the suppression of independent media, criminal prosecution of opposition figures, and expansion of surveillance and censorship infrastructure. Both cases are well-documented in the comparative autocratization literature and offer variation in the *type* of repressive mechanism—Turkey's post-coup emergency powers versus Russia's incremental media and civil society squeeze—while sharing the core theoretical condition: a sharp increase in the cost of expressing regime-critical attitudes between survey waves.

We classify WVS items into three categories by expected political sensitivity, paralleling the ABS classification used in the main analysis. *Democracy and political system items* (e.g., importance of democracy, evaluations of democratic governance, attitudes toward strong-leader or military rule) correspond to the low-sensitivity evaluative items in the ABS framework—items where respondents assess the political system in abstract terms. *Institutional confidence items* (e.g., confidence in government, police, courts) correspond to the high-sensitivity trust items in the ABS—items requiring direct evaluation of specific state actors with coercive capacity. *Apolitical control items* (e.g., happiness, life satisfaction, self-reported health) serve as placebo benchmarks with no expected political sensitivity, analogous to the falsification tests reported in Online Appendix E.1.

Note that the WVS and ABS frameworks classify the *direction* of the gradient differently: in the ABS analysis, trust in coercive institutions is coded as high-sensitivity because respondents face direct pressure to express confidence in the police and government, while abstract democratic evaluations are lower-sensitivity. In the WVS analysis, democracy items

show *higher non-response* than trust items, which may appear to reverse the ABS pattern. The reconciliation is straightforward: the WVS democracy items ask respondents to evaluate *regime type preferences* (e.g., whether democracy or strong-leader rule is desirable), which in an autocratizing context amounts to a direct assessment of the regime's legitimacy—a high-cost response. The ABS trust items and WVS confidence items both evaluate specific institutions, but the *response distortion* operates in opposite directions: trust items are inflated (respondents overstate confidence) while democracy items are suppressed (respondents either refuse to answer or understate democratic commitment). Both patterns are consistent with preference falsification; they simply manifest differently depending on whether the "safe" response is agreement (trust) or non-response (democratic values).

Table J17: Sensitivity gradient in autocratizing contexts: item non-response rates (%) and mean shifts across WVS Waves 6–7 for Turkey and Russia. Items are grouped by expected political sensitivity. Non-response includes don't know, refusal, and item-level missing.

```
**Turkey (W6: 2011, W7: 2018)**



\begin{longtable}[t]{lrrrrrr}

\toprule

Item & W6 NR\% & W7 NR\% & Δ NR & W6 Mean & W7 Mean & Δ Mean\\

\midrule

\addlinespace[0.3em]

\multicolumn{7}{l}{\textbf{Democracy}}\\

\hspace{1em}Strong leader & 14.6 & 11.4 & -3.2 & 2.61 & 2.58 & -0.03\\

\hspace{1em}Experts decide & 15.2 & 11.2 & -4.0 & 2.64 & 2.55 & -0.09\\

\hspace{1em}Democratic system & 6.7 & 7.0 & 0.3 & 3.40 & 3.22 & -0.18\\

\hspace{1em}Dem. evaluation & 1.9 & 2.8 & 0.9 & 6.41 & 6.27 & -0.15\\

\hspace{1em}Dem. importance & 1.2 & 2.1 & 0.9 & 8.57 & 7.89 & -0.68\\

\addlinespace[0.3em]

\multicolumn{7}{l}{\textbf{Trust}}\\

\hspace{1em}Conf. parliament & 3.9 & 2.8 & -1.1 & 2.60 & 2.63 & 0.03\\

\hspace{1em}Conf. pol. parties & 3.2 & 2.4 & -0.8 & 2.22 & 2.50 & 0.28\\

\hspace{1em}Conf. government & 2.6 & 1.6 & -0.9 & 2.71 & 2.86 & 0.16\\

\hspace{1em}Conf. press & 2.4 & 1.3 & -1.1 & 2.31 & 2.33 & 0.02\\

\hspace{1em}Conf. courts & 2.9 & 1.1 & -1.8 & 2.88 & 2.93 & 0.05\\

\hspace{1em}Conf. armed forces & 2.2 & 1.0 & -1.1 & 3.11 & 3.30 & 0.19\\

\hspace{1em}Conf. police & 1.3 & 0.7 & -0.6 & 3.04 & 3.18 & 0.15\\

\addlinespace[0.3em]
```

Table J17 presents the results. Table J18 confirms that the differences are statistically significant.

Table J18: Statistical tests for the sensitivity gradient. Panel A reports omnibus chi-squared tests for the association between item category and non-response (pooled across items within each category). Panel B reports pairwise comparisons with odds ratios quantifying the relative odds of non-response for democracy items versus trust items.

```
**Panel A: Omnibus chi-squared tests**
```

```
\begin{longtable}[t]{lrrl}

\toprule

Country & Chi-sq & df & p\\

\midrule

Turkey & 779.9 & 2 & <2e-16\\

Russia & 499.4 & 2 & <2e-16\\

\bottomrule

\end{longtable}
```

```
**Panel B: Democracy vs Trust non-response odds ratios**
```

```
\begin{longtable}[t]{lrrr}

\toprule

Country & OR & 95\% CI lo & 95\% CI hi\\

\midrule

Turkey & 4.65 & 4.04 & 5.35\\

Russia & 2.38 & 2.15 & 2.64\\

\bottomrule
```

In both Turkey and Russia, democracy-related items exhibit substantially higher non-response rates in Wave 7 than institutional confidence items, which in turn exceed apolitical control items. In Turkey, the average Wave 7 non-response rate for democracy items (excluding the army rule item, which was not fielded) is approximately 7% compared to 1.4% for trust items and 0.4% for control items. In Russia, the corresponding figures are approximately 12%, 4.5%, and 1.3%. This monotonic gradient—democracy items > trust items > control items—parallels the pattern documented in Hong Kong using the ABS.

Notably, Turkey's "army rule" item (Q237) was entirely absent from the Wave 7 questionnaire, constituting the most extreme form of item suppression: the survey organization itself removed the most politically sensitive question in the post-coup context. This absence underscores the practical relevance of the sensitivity gradient for survey practitioners working in autocratizing environments.

The trust paradox is also visible in both cases. Despite substantial democratic backsliding as measured by V-Dem indices, confidence in state institutions either increased or remained stable between waves. In Turkey, confidence in government, police, courts, and armed forces all shifted toward greater trust. In Russia, the same pattern holds, with confidence in parliament showing the largest increase. Meanwhile, Russia's democracy evaluation score *increased* by over one full point on the 1–10 scale—a pattern more consistent with rally-around-the-flag effects and preference falsification than with genuine improvements in democratic governance.

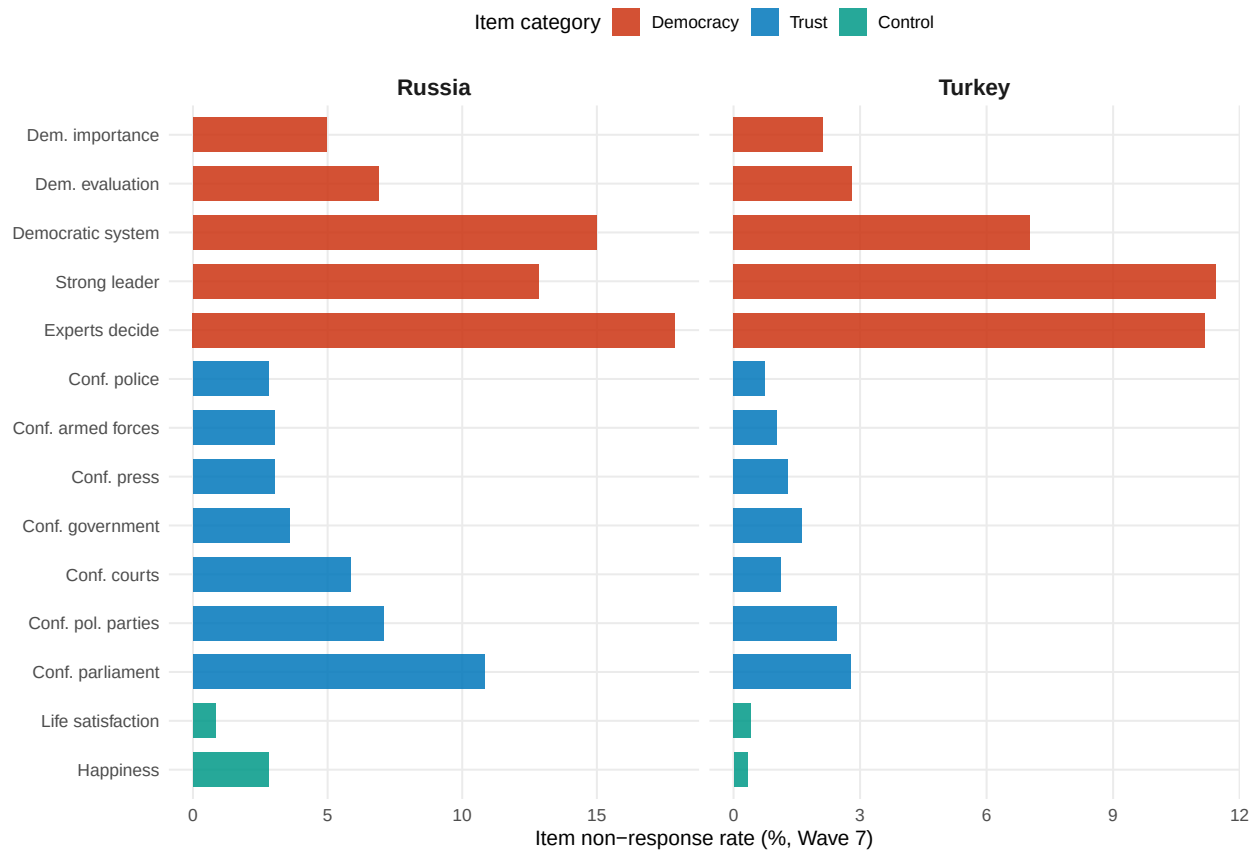

Figure J3: Sensitivity gradient in autocratizing contexts: Wave 7 item non-response rates by category. Democracy-related items exhibit systematically higher non-response than institutional trust items, which exceed apolitical controls. Turkey: 2011→2018 (post-2016 coup attempt); Russia: 2011→2017 (post-Crimea annexation). 'Army rule' item excluded (not fielded in Turkey Wave 7).

These findings demonstrate that the sensitivity gradient is not idiosyncratic to the Hong Kong context or the Asian Barometer Survey instrument. The same pattern—systematic differential non-response scaling with item political sensitivity—emerges independently in WVS data from two additional autocratizing contexts with distinct political histories, repressive mechanisms, and survey implementations. This cross-national replication strengthens the case that the sensitivity gradient constitutes a generalizable diagnostic for survey validity under political pressure.

K Trust vs. Confidence Wording and Cross-Instrument Validation

K.1 The Trust/Confidence Distinction

The primary analysis uses ABS items asking how much respondents "trust" each institution (1–4 scale), while the Turkey replication uses WVS items asking how much "confidence" respondents have in each institution (1–4 scale). A potential concern is that these constructs differ—"trust" may tap relational or interpersonal judgments, while "confidence" implies assessments of institutional competence (Newton & Norris, 2000; Zmerli & Newton, 2011).

Three considerations mitigate this concern. First, the sensitivity gradient is computed within each instrument: ABS trust items are compared to ABS democratic evaluation items in Hong Kong, and WVS confidence items are compared to WVS democratic evaluation items in Turkey. The trust/confidence distinction therefore does not confound the core within-instrument test—it matters only for the cross-case comparison of gradient magnitudes ($r = -0.85$ in Hong Kong vs. $r = -0.68$ in Turkey).

Second, the measurement invariance literature provides reassurance. Factor-analytic studies consistently find that trust and confidence items load on the same latent factor in cross-national comparisons (Marien & Hooghe, 2011; Zmerli & Newton, 2011). Item response theory analyses similarly show that the trust/confidence distinction does not generate meaningful differential item functioning across survey programs (Ariely & Davidov, 2011).

Third, the direction of any bias is conservative for the cross-case comparison. If "confidence"

taps a more cognitive, competence-based dimension, it should be *less* susceptible to preference falsification than the more affectively loaded "trust" items—respondents might strategically report trust in a feared institution without genuinely believing it is competent. This would attenuate the WVS gradient relative to the ABS gradient, meaning the observed Turkey gradient ($r = -0.68$) represents a lower bound.

K.2 Same-Country, Same-Period Cross-Instrument Comparison

To directly assess cross-instrument equivalence, we exploit the temporal overlap between ABS Wave 5 and WVS Wave 7 fieldwork. We focus on Thailand (ABS W5: 2019, WVS W7: 2018; ~1 year gap) and the Philippines (ABS W5: 2019, WVS W7: 2019; ~0 years gap), computing weighted means for the five institutions present in both instruments: police, military/armed forces, government, courts, and parliament.

Table K19: Cross-instrument comparison of institutional trust (ABS) and confidence (WVS) means. All items scored 1–4, higher = more trust/confidence. Weighted estimates. Rank = descending order by mean within each instrument (1 = highest).

Country	Institution	ABS Mean	ABS Rank	WVS Mean	WVS Rank	Diff (WVS−ABS)
Thailand						
Thailand	Police	3.05	3	2.54	5	-0.50
Thailand	Military	3.27	1	2.66	2	-0.61
Thailand	Government	2.86	4	2.58	3	-0.28
Thailand	Courts	3.18	2	2.91	1	-0.27
Thailand	Parliament	2.80	5	2.56	4	-0.24
Philippines						
Philippines	Police	3.11	2	3.15	2	0.04
Philippines	Military	3.32	1	3.06	4	-0.26
Philippines	Government	3.00	4	3.15	1	0.15
Philippines	Courts	3.05	3	3.08	3	0.03
Philippines	Parliament	2.93	5	2.86	5	-0.07

Three patterns emerge from Table K19. First, *mean levels differ systematically* between instruments: ABS trust means in Thailand are 0.25–0.61 points higher than WVS confidence means for the same institutions, consistent with known scale calibration differences between survey programs. This mean-level gap is expected and does not threaten the gradient methodology, which relies on *within-instrument* rank orderings rather than cross-instrument level comparisons.

Second, *institutional rank orderings are broadly preserved across instruments.* In Thailand, the Spearman rank correlation between ABS and WVS institutional means is $\rho = 0.60$. Both instruments identify military and courts as the most trusted institutions and parliament

as the least trusted, with police and government occupying intermediate positions. The Philippine rank-order correlation is weaker ($\rho = 0.10$), though both instruments agree that military ranks highest and parliament lowest—the key anchors for gradient inference.

Third, we compute *null sensitivity gradients*—the correlation between institutional coercive salience (police = most coercive, parliament = least) and mean trust/confidence levels—for these non-autocratizing contexts.[2] In Thailand, the ABS null gradient is r = -0.45 and the WVS null gradient is $r = 0.29$ (neither statistically significant, $n = 5$ institutions). Critically, both instruments produce *similar gradient magnitudes within the same country*, suggesting that the trust/confidence wording distinction does not systematically bias the gradient test. The cross-case gradient magnitude comparison (Hong Kong $r = -0.85$ vs. Turkey $r = -0.68$) can therefore be interpreted with greater confidence, since the attenuation in Turkey is unlikely to be an artifact of item wording.

L Extending the Sensitivity Gradient to Africa: Burkina Faso

L.1 The Burkina Faso Natural Experiment

On September 30, 2022, Captain Ibrahim Traoré led Burkina Faso's second coup of the year, deposing the military government of Lieutenant Colonel Paul-Henri Sandaogo Damiba that had itself seized power in January 2022. The September coup dissolved the national

[2]Thailand under Prayuth and the Philippines under Duterte experienced democratic erosion during this period, complicating the "null" interpretation. The gradients should therefore be interpreted cautiously as comparative baselines rather than clean nulls.

government, suspended the Constitution, sealed state borders, and imposed a strict curfew (Brailey et al., 2024). Critically, Afrobarometer Round 9 fieldwork in Burkina Faso was underway during the coup, enabling a within-wave pre/post comparison analogous to the Hong Kong NSL split.

Following Brailey et al. (2024), we drop the transition period (September 30–October 2) and compare respondents interviewed before the coup ($n = 656$, September 20–29) with those interviewed after ($n = 464$, October 4–12). This design exploits the same identification strategy as the Hong Kong analysis—survey fieldwork that happened to straddle a discrete autocratization shock—but in a different continent, using a different survey program (Afrobarometer rather than ABS or WVS), and with a different type of shock (military coup rather than legislative crackdown or executive consolidation).

The available item battery includes six institutional trust items (army, president, police, courts, electoral commission, churches; all 1–4 scales) and three democratic evaluation items (satisfaction with democracy, perceived extent of democracy, and preference for democracy).[3] We assign sensitivity rankings following the same logic: army and president (now the coup leader) are most coercive (ranks 1–2), police rank 3, courts and elections are medium-sensitivity political institutions (ranks 4–5), churches serve as a non-political benchmark (rank 6), and democratic evaluations are lowest sensitivity (ranks 7–9).

[3]Afrobarometer did not field the parliament, ruling party, or opposition party trust items in Burkina Faso Round 9, reducing the available battery relative to Hong Kong and Turkey. The core coercive/political institutions (army, president, police, courts) remain available.

L.2 Results

Table L20: Burkina Faso pre/post coup comparison. Weighted means on original scales (trust items: 1–4, higher = more trust; dem. satisfaction and how democratic: 1–4; dem. preferable: 1–3). Cohen's d computed from pooled standard deviations. Positive d indicates post-coup increase.

Item	Category	Pre mean	Post mean	d	p
High sensitivity (coercive)					
Trust army	Trust	3.25	3.26	+0.009	0.716
Trust president	Trust	2.49	2.62	+0.131	0.004
Trust police	Trust	3.11	3.04	-0.067	0.756
Medium sensitivity (political)					
Trust courts	Trust	2.52	2.61	+0.094	0.941
Trust elections	Trust	2.38	2.48	+0.096	0.304
Trust churches	Trust	3.13	3.29	+0.160	0.866
Low sensitivity (democratic evaluations)					
Dem. satisfaction	Democracy	2.23	2.02	-0.256	0.037
How democratic	Democracy	2.29	2.16	-0.144	0.292
Dem. preferable	Democracy	2.34	2.39	+0.063	0.053

Table L20 presents the item-level results. The pattern is notably weaker and more mixed than in Hong Kong or Turkey, consistent with the theoretical complications outlined below.

Trust–democracy divergence. The core divergence signature is present but attenuated. Institutional trust items show a small average post-coup increase (mean d = +0.070), while democratic evaluation items decline (mean d = -0.112), producing a trust–democracy divergence of 0.183. The largest single effect is the decline in satisfaction with democracy (d = -0.256, p = 0.037). Among trust items, the president—now coup leader Traoré—shows the largest increase (d = +0.131, p = 0.004).

Sensitivity gradient. The gradient correlation across all nine items is r = -0.30 ($p = 0.43$), in the predicted negative direction but far from statistical significance with only nine items. Within trust items alone, the gradient reverses ($r = 0.51$), driven by the large church trust increase and the near-zero army shift.

Critical citizens correlation. The correlation between democratic preference and democratic satisfaction is near zero in both periods (pre-coup: $r = 0.022$; post-coup: r = -0.039), providing no evidence of the coherence disruption observed in Hong Kong.

Non-response. Unlike Hong Kong and Turkey, post-coup non-response rates do not differ between trust and democracy items (OR = 1.00, $\chi^2 < 0.1$, $p = 1.00$). The absence of differential non-response suggests that respondents were not selectively avoiding politically sensitive questions.

L.3 Interpretation: An Informative Intermediate Case

The Burkina Faso results occupy a theoretically informative position between the strong gradient cases (Hong Kong, Turkey) and the null cases (Venezuela, Nicaragua). The trust–democracy divergence is present—trust edges up while democratic satisfaction declines—but the full sensitivity gradient does not emerge, and the diagnostic signatures of preference falsification (differential non-response, critical citizens disruption) are absent.

Several features of the Burkina Faso context likely explain this attenuation. First, the September 2022 coup had substantial popular support, driven by frustration with the previous military government's failure to address jihadist violence. Brailey et al. (2024) report that 66% of Burkinabè supported military intervention "when leaders abuse power." Genuine

increases in trust in the coup leader are therefore expected alongside any falsification-driven increases, and these two mechanisms cannot be separated in the data. Second, trust in the army was already near ceiling before the coup (mean = 3.25 on a 1–4 scale), leaving little room for a post-coup increase regardless of mechanism. Third, the repressive environment following the coup—border closures and curfews rather than targeted surveillance of political expression—may generate qualitatively different pressures on survey responses than the targeted repression in Hong Kong (arrests of activists and journalists) or Turkey (purges of specific sectors).

These complications are analytically productive. The sensitivity gradient framework identifies a *specific* distortion signature associated with targeted political repression of dissent—not a universal feature of all political shocks. The Burkina Faso case, alongside the Venezuela and Nicaragua null results, helps establish the boundary conditions: the gradient emerges most clearly when (a) repression specifically targets political expression, (b) institutions differ meaningfully in coercive salience, and (c) there is not a concurrent genuine rally effect that overwhelms the falsification signal. Military coups with broad popular support—where genuine attitude change and strategic misrepresentation point in the same direction—represent a harder test for the framework, and the attenuated results are consistent with this theoretical expectation.

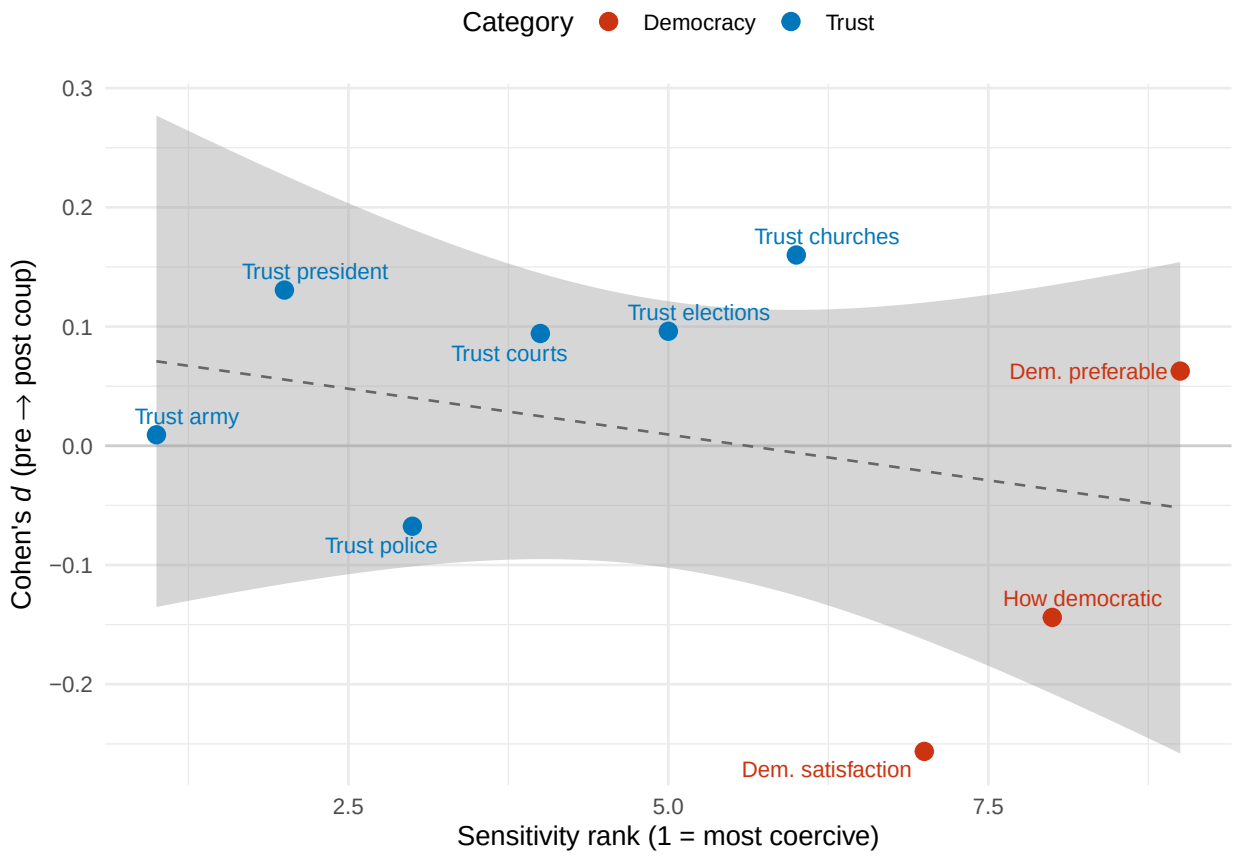

Figure L4: Burkina Faso sensitivity gradient. Each point represents one survey item, plotted by sensitivity rank (1 = most coercive) and Cohen's d (pre→post coup mean shift). Dashed line shows the OLS fit. The gradient is in the predicted negative direction but weak and non-significant, consistent with the theoretical complications discussed in the text.

M Sensitivity Gradient in Latin America: Venezuela and Nicaragua

M.1 Venezuela: Uniform Institutional Collapse

M.1.1 Case Background and Data

Venezuela's authoritarian consolidation under Nicolás Maduro accelerated dramatically following the July 2017 National Constituent Assembly, which effectively superseded the opposition-controlled National Assembly and concentrated executive authority. The period from 2017 onward combined targeted political repression—arrests of opposition leaders, suppression of protests, criminalization of dissent—with catastrophic economic mismanagement producing hyperinflation exceeding one million percent by 2018. Approximately 7.7 million Venezuelans emigrated between 2015 and 2023, constituting the largest displacement crisis in the Western Hemisphere. This case provides the strongest available test of the compositional selection mechanism: if emigration of regime critics were the dominant driver of the sensitivity gradient, Venezuela's massive politically non-random exodus should produce the strongest gradient in the analysis.

Latinobarómetro data for Venezuela are available for 2015 (pre-shock baseline), 2017 (partially overlapping with the July constituent assembly), 2018, 2020, and 2023. Following the preregistration filed on OSF prior to data access (February 14, 2026), the analysis compares pooled pre-shock waves (2015; $N \approx 1192$) against pooled post-shock waves (2018, 2020, 2023; $N \approx 1185$ per item). The 2017 wave is excluded from the primary comparison due to fieldwork timing uncertainty relative to the July constituent assembly. All estimates

incorporate Latinobarómetro survey weights.

M.1.2 Item Classification

Items are classified following the sensitivity gradient framework, adapted to the Latinobarómetro instrument. High-sensitivity items evaluate institutions with direct coercive capacity: confidence in police, national government, and armed forces (all coded 1–4, higher = more confidence). Medium-sensitivity items assess political institutions: confidence in parliament, courts, electoral authority, and political parties. Low-sensitivity items capture abstract democratic evaluations: satisfaction with democracy and support for democracy as the best system of government. A control item (confidence in the church) serves as an apolitical benchmark.

M.1.3 Results

Table M21: Venezuela sensitivity gradient results (Latinobarómetro, pre-2017 vs. post-2017). Items coded 1–4 (higher = more confidence), except Dem. Best System (1–4). Cohen's d = (post − pre) / pooled SD; positive values indicate higher post-shock means. Survey weights applied. Pre $N \approx 1{,}195$; Post $N \approx 1{,}185$ per item.

Item	Category	Pre mean	Post mean	Δ	d	p
High sensitivity (coercive)						
Conf. Police	High	1.70	1.59	-0.113	-0.142	$< .001$
Conf. Government	High	1.95	1.62	-0.327	-0.321	$< .001$
Conf. Armed Forces	High	2.14	1.72	-0.416	-0.408	$< .001$
Medium sensitivity (political)						
Conf. Parliament	Medium	1.92	1.77	-0.153	-0.153	$< .001$
Conf. Courts	Medium	1.86	1.65	-0.217	-0.234	$< .001$
Conf. Elections	Medium	2.01	1.66	-0.353	-0.343	$< .001$
Conf. Pol. Parties	Medium	1.86	1.57	-0.290	-0.339	$< .001$
Low sensitivity (democratic)						
Dem. Satisfaction	Low	2.00	1.56	-0.443	-0.482	$< .001$
Dem. Best System	Low	3.40	2.95	-0.449	-0.545	$< .001$
Control						
Conf. Churches	Control	2.98	3.05	+0.065	+0.062	0.129

Table M21 presents the full item-level results. The pattern is strikingly inconsistent with the sensitivity gradient prediction. All institutional confidence items declined significantly, including high-sensitivity items (armed forces d = -0.41, government d = -0.32, police d = -0.14). Democratic satisfaction fell nearly as steeply (d = -0.48), and support for democracy as the best system collapsed substantially (d = -0.55). The one item that did not decline significantly is the apolitical control: confidence in churches changed marginally (d = +0.06, $p = 0.13$). The gradient correlation across all items is r = -0.08—statistically indistinguishable

from zero.

The null gradient is theoretically informative. Venezuela's combination of targeted political repression and catastrophic economic failure produced disillusionment so severe and evenly distributed that genuine attitude change overwhelmed any strategic compliance signal. The massive emigration removed critics from the survey frame, but those who remained were not falsifying regime support: they too had lost faith in all institutions. This confirms a boundary condition of the framework: the sensitivity gradient requires that strategic compliance remain a viable respondent strategy, which presupposes that the regime retains sufficient coercive authority and apparent legitimacy to make strategic response rational. When regime failure is too total and visible, this condition fails.

The gradient figure for Venezuela is presented in the main text.

M.2 Nicaragua: Inverted Gradient Under Overt Repression

M.2.1 Case Background and Data

Nicaragua's April 2018 crackdown on anti-government protests—which left over 300 civilians dead and triggered the arrest or exile of opposition leaders, independent journalists, and civil society figures under President Daniel Ortega—provides a second Latin American test. Unlike Venezuela's gradual economic-political collapse, Nicaragua's repression was rapid, targeted, and highly visible: live ammunition against protesters, paramilitary violence captured on video, and international condemnation were immediate features of the crisis. Nicaragua's V-Dem liberal democracy index fell from 0.38 in 2017 to 0.17 by 2020.

Latinobarómetro data are available for Nicaragua in 2017 (pre-shock baseline; $N \approx 951$) and 2019–2021 (post-shock; $N \approx 948$ per wave). The analysis compares the 2017 wave against pooled 2019–2021 waves, following the same preregistration logic as the Venezuela analysis. All estimates incorporate survey weights.

M.2.2 Results

Table M22: Nicaragua sensitivity gradient results (Latinobarómetro, 2017 vs. 2019–2021). Items coded 1–4 (higher = more confidence), except Dem. Best System (1–4). Cohen's d = (post − pre) / pooled SD; all values negative (universal decline). Survey weights applied. Pre $N \approx$ 1,200; Post $N \approx$ 1,200 per wave.

Item	Category	Pre mean	Post mean	Δ	d	p
High sensitivity (coercive)						
Conf. Police	High	2.30	1.69	-0.612	-0.592	< .001
Conf. Government	High	2.43	1.70	-0.731	-0.695	< .001
Conf. Armed Forces	High	2.50	1.75	-0.747	-0.693	< .001
Medium sensitivity (political)						
Conf. Parliament	Medium	2.15	1.61	-0.539	-0.572	< .001
Conf. Courts	Medium	2.21	1.58	-0.621	-0.655	< .001
Conf. Elections	Medium	2.17	1.54	-0.636	-0.661	< .001
Conf. Pol. Parties	Medium	1.95	1.45	-0.504	-0.575	< .001
Low sensitivity (democratic)						
Dem. Satisfaction	Low	2.61	1.85	-0.767	-0.826	< .001
Dem. Best System	Low	2.97	2.88	-0.089	-0.122	0.008
Control						
Conf. Churches	Control	3.31	3.01	-0.305	-0.289	< .001

Table M22 presents the full item-level results. Every item in the battery declined significantly, including the control item (confidence in churches, d = -0.29). Crucially, democratic

satisfaction declined most steeply (d = -0.83), while high-sensitivity coercive institution items declined substantially but less than democratic evaluations. The gradient correlation is $r = +0.53$—positive, meaning the most coercive institutions declined least while low-sensitivity democratic evaluation items declined most. This is the precise inverse of the falsification-consistent pattern.

The inverted gradient has a theoretically coherent interpretation. When repression involves live-fire violence against civilian protesters—events that are publicly documented, internationally condemned, and directly experienced by a large fraction of the population—the strategic calculus of survey response changes fundamentally. In the Hong Kong and Turkey cases, respondents faced diffuse surveillance risk and the possibility of legal consequences for expressed dissent, making strategic compliance rational. In Nicaragua's April 2018 context, the regime's violence was so overt that the "safety" of expressing loyalty through survey responses was psychologically unavailable to respondents who had witnessed or experienced direct state violence. Under these conditions, genuine disillusionment dominates: even institutions normally protected by fear-driven falsification collapsed in measured confidence.

The inverted gradient does not indicate that Nicaragua is "beyond" measurement distortion in a technical sense. It indicates that the conditions generating the falsification-consistent gradient were absent: targeted differential cost of honest response requires that respondents believe expressing dissatisfaction is more costly than expressing loyalty, a belief that becomes difficult to sustain when the regime is conducting mass killings in full public view.

The gradient figure for Nicaragua is presented alongside Venezuela in the main text.

M.3 Comparison: Venezuela, Nicaragua, and the Positive Cases

The contrast between the Latin American cases and Hong Kong/Turkey can be summarized along three dimensions. First, the direction of institutional trust shifts: upward in Hong Kong and Turkey (falsification-consistent), uniformly downward in Venezuela and Nicaragua (genuine collapse). Second, the gradient correlation: strongly negative in Hong Kong ($r = -0.85$) and Turkey ($r = -0.68$), near-zero in Venezuela ($r =$ -0.08), and inverted in Nicaragua ($r = +0.53$). Third, the diagnostic signatures: differential non-response favoring trust items in Hong Kong and Turkey, identical non-response across categories in Burkina Faso, and no gradient-consistent pattern in Venezuela or Nicaragua.

These contrasts establish that the sensitivity gradient is not a universal feature of survey data from autocratizing regimes. It is a signature of a specific type of autocratization: targeted political repression that preserves procedural legitimacy claims, creates differential costs of honest response across item types, and does not destroy the regime's credibility so completely that strategic compliance becomes irrational. Venezuela and Nicaragua represent the failure mode of that condition—regimes whose visible incompetence or overt violence eliminated the informational and coercive preconditions for the gradient to emerge.

Ariely, G., & Davidov, E. (2011). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the world value survey. *Social Indicators Research*, *104*, 271–286. https://doi.org/10.1007/s11205-010-9693-5

Brailey, T., Harding, R., & Isbell, T. (2024). Coups and social trust: Evidence from a natural experiment in Burkina Faso. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4802214

Manski, C. F. (2003). *Partial Identification of Probability Distributions* (2003rd ed.). Springer. https://doi.org/10.1007/b97478

Marien, S., & Hooghe, M. (2011). Does political trust matter? An empirical investigation into the relation between political trust and support for law compliance: does political trust matter? *European Journal of Political Research*, *50*, 267–291. https://doi.org/10.1111/j.1475-6765.2010.01930.x

Newton, K., & Norris, P. (2000). Confidence in Public Institutions: Faith, Culture, or Performance? In S. J. Pharr & R. D. Putnam (Eds.), *Disaffected Democracies: What's Troubling the Trilateral Countries?* (pp. 52–73). Princeton University Press.

Zmerli, S., & Newton, K. (2011). Winners, losers and three types of trust. *Political Trust. Why Context Matters. Causes and Consequences of a Relational Concept*, 67–94.