When Do Surveys Produce False Positives for Regime Support? A Sensitivity Gradient Approach to Measurement Distortion in Autocratizing Regimes

Jeffrey Stark

Abstract

Scholars of democratic backsliding rely on survey data to track regime legitimacy, yet the same repressive forces that dismantle institutions also distort the instruments used to measure public attitudes. This article develops and validates a "sensitivity gradient" framework—a portable diagnostic for identifying measurement distortion in standard survey batteries without experimental design. The framework exploits a theoretical prediction: because the cost of honest response varies predictably across survey items as a function of institutional coerciveness, preference falsification generates a systematic within-instrument gradient. High-sensitivity items (e.g., trust in police) inflate due to strategic compliance and the selective exit of critics; low-sensitivity items (e.g., abstract democratic values) remain comparatively robust. The divergence itself constitutes a false positive for regime support. Testing this framework using a natural quasi-experiment in Hong Kong (the 2020 National Security Law), the data reveal a "trust paradox" consistent with measurement distortion: trust in police surged ($d \approx +0.49$) while perceived democratic suitability collapsed ($d = -0.54$), and the canonical "critical citizens" correlation reverses—a second diagnostic signal that standard survey relationships have broken down. Cross-national replication using World Values Survey data from Turkey ($r = -0.68$), Latinobarómetro data from Venezuela and Nicaragua, and Afrobarometer data from Burkina Faso establishes both external validity and discriminant validity: the gradient appears when repression is targeted and procedurally legitimated, and is absent or inverted when institutional collapse is genuine. The resulting diagnostic toolkit—five portable signals computable from standard ABS, WVS, Afrobarometer, and Latinobarómetro batteries—offers researchers a practical method for assessing whether survey-based legitimacy claims in autocratizing contexts can be trusted to mean what they appear to say.

1 Introduction

Scholars of democratic erosion face a fundamental identification challenge: the same coercive forces that dismantle democratic institutions also reshape the conditions under which citizens evaluate them. As regimes consolidate control, cross-national surveys often record puzzling spikes in institutional trust and regime satisfaction. Do these shifts reflect genuine "authoritarian resilience" (Nathan 2003) rooted in performance legitimacy and a preference for order? Or do they reflect measurement distortion—a mirage created when critics either hide their true preferences (falsification) or exit the public sphere entirely (selection)? Distinguishing between genuine legitimacy and measurement distortion is critical; misinterpreting the latter as the former risks validating authoritarian narratives of popular support.

This article develops and tests a "sensitivity gradient" framework to adjudicate between these possibilities without recourse to experimental design. While existing methods like list experiments are powerful, they cannot be applied retrospectively to the vast archives of standard survey data used to monitor global democracy. The sensitivity gradient approach fills this gap by exploiting a straightforward theoretical prediction: the cost of honest response under repression is not uniform but varies predictably across survey items.

I argue that high-sensitivity items—those evaluating specific coercive actors like the police or national government—are most susceptible to distortion. Respondents face strong incentives to feign support for these actors ("strategic compliance") or, in the case of critics, to self-select out of the survey pool entirely. Conversely, low-sensitivity items—such as abstract evaluations of "democracy" or "freedom"—often remain safer channels for expressive dissent,

or at least carry lower penalties for non-compliance. If genuine support were rising, we should see uniform improvements across both domains. If measurement distortion is at work, we should observe a "gradient" of divergence: inflated support for coercive institutions coexisting with collapsing evaluations of the political environment.

I test this framework using a natural quasi-experiment created by the disruption of the Asian Barometer Survey (Wave 5) in Hong Kong. Fieldwork was split across the 2019 pro-democracy protests and the post-National Security Law (NSL) period, holding the instrument constant while the cost of dissent shifted exogenously. The analysis yields three findings. First, the sensitivity gradient predicts a "trust paradox": trust in the police surged ($d \approx +0.49$) while democratic suitability scores collapsed ($d = -0.54$), with the magnitude of inflation tracking each institution's coercive role. Second, the canonical "critical citizens" correlation—whereby committed democrats express lower regime satisfaction—reverses in the post-NSL period. Third, this reversal is concentrated among "procedural democrats," a finding that challenges critical theoretical accounts suggesting authoritarian citizens simply possess distinct, non-liberal notions of democracy.

I adjudicate among competing mechanisms through age-stratified analysis, information environment tests, and rally effect diagnostics (Section 4.5), finding that no single alternative explanation accounts for all observed patterns. I then apply the same framework to Turkey using World Values Survey data spanning the 2016 coup attempt, finding consistent evidence of the gradient across a different autocratization context and survey instrument (Section 4.6). The framework is best suited to contexts where autocratization is rapid and survey fieldwork spans the transition, though cross-wave comparisons (as in Turkey) also reveal gradient

patterns when waves bracket clear autocratization shocks. The approach requires only standard trust and democratic evaluation items available in all major cross-national survey programs.

These findings contribute to three literatures. First, and most directly, they offer a practical methodological advance for the study of democratic backsliding (Waldner and Lust 2018): a diagnostic toolkit that can be applied retrospectively to data already collected in ABS, WVS, Afrobarometer, and Latinobarómetro archives, without experimental design, and that has now been validated in five distinct political contexts.[1] Second, the findings engage debates on authoritarian legitimacy (Nathan 2003; Tang 2016; Dukalskis and Gerschewski 2017), providing empirical grounds to question, in at least some autocratizing contexts, whether measured "performance legitimacy" reflects genuine public approval or is partly an artifact of survey non-response and preference falsification. Third, they bridge critical theory and survey methodology, operationalizing the concern that quantitative metrics can inadvertently reproduce the logic of domination by masking the silence of the marginalized—and offering a concrete means of detection.

---

[1]Preliminary evidence from published Levada Center tracking polls in Russia shows patterns consistent with the framework (trust increases coinciding with democratic support decreases around the February 2022 invasion of Ukraine), but individual-level microdata access would be required for the item-level diagnostics that constitute the key tests. Online Appendix J reports available WVS-based evidence for Russia, where Waves 6 (2011) and 7 (2017) bracket the post-Crimea annexation period rather than the 2022 invasion.

2 The Validity Problem in Autocratizing Regimes

2.1 Survey Validity Under Political Pressure

The concern that survey measures may not mean the same thing in authoritarian and democratic contexts is not new. Kuran (1995)'s foundational work on preference falsification demonstrated that citizens under repressive regimes systematically misrepresent their private beliefs in public, generating a gap between stated and genuine preferences that can persist at equilibrium. Wedeen (2000) extended this insight by arguing that authoritarian regimes do not require genuine belief—only the public performance of compliance, such that "acting as if" one supports the regime becomes the regime's operative goal. Simpser (2013) demonstrates a parallel logic in the electoral domain: governments manipulate elections not merely to win but to signal dominance and convey the futility of resistance, a communicative function that applies equally to survey responses when citizens treat their answers as public acts rather than private disclosures. Empirical assessments of this phenomenon have proliferated in recent years. Tannenberg (2022) shows cross-nationally that respondents who believe the government commissioned a survey report systematically higher trust in autocracies but not in democracies, providing direct evidence that self-censorship inflates trust measures under authoritarian conditions. List experiments and endorsement experiments in China suggest that direct survey measures substantially overstate trust in sensitive institutions (Nicholson and Huang 2023; Blair et al. 2014; Bullock et al. 2011). Jiang and Yang (2016) exploit a political purge in China to show that responses to politically sensitive and insensitive survey items diverge following repressive events—an empirical strategy closely analogous to the present study. Kobayashi and Chan (2022) provide direct

panel-level evidence from Hong Kong, demonstrating that pro-democracy respondents were more likely to drop out of political polls following the NSL, and that those who remained falsified potentially sensitive past behavior. More recently, Shamaileh (2025) demonstrates that nonresponse rates alone are insufficient proxies for preference falsification, since respondents under repression shift toward the regime-preferred response rather than simply declining to answer—a pattern consistent with Simpser (2013)'s concept of strategic compliance, whereby citizens signal obedience through regime-favorable responses rather than merely withdrawing.

These studies establish that survey bias under authoritarianism is real and consequential. What remains underdeveloped is a framework for identifying bias *within* a given survey instrument without experimental manipulation. The sensitivity gradient approach proposed here fills this gap by exploiting a straightforward theoretical prediction: because the cost of honest response varies across items as a function of their political sensitivity, preference falsification should operate *differentially* within the same survey. Trust in the police—the institution most directly associated with coercive enforcement—should be the most inflated; trust in the judiciary, which retains some independence, should be less affected; abstract democratic evaluations, which do not directly implicate specific institutions, should remain comparatively honest. This differential generates testable predictions about the *pattern* of observed shifts, allowing analysts working with standard survey batteries to distinguish falsification from genuine attitude change without recourse to designed experiments.

The micro-level logic connecting repression to differential survey response runs through three sequential steps. First, autocratization shifts the perceived cost of honest response

asymmetrically across item types. When a regime deploys coercive institutions visibly against political opponents, respondents update their beliefs about the risk of expressing dissatisfaction with those specific institutions: the police who arrested protesters, the government that enacted emergency legislation. Abstract evaluations of "democracy as a system" carry lower perceived risk because they are harder for authorities to interpret as directed criticism and because regimes typically claim democratic legitimacy themselves, making abstract democratic endorsements compatible with compliance. Second, this cost asymmetry produces two behavioral responses. Among respondents who remain in the survey pool, high-stakes items elicit strategic misrepresentation: respondents report regime-favorable values not because their attitudes have changed but because honest expression now carries costs they are unwilling to bear (Kuran 1995; Simpser 2013). Among the subset of respondents most likely to express dissent—those with the strongest prior dissatisfaction—the elevated cost of participation induces selective exit: refusal, emigration, or withdrawal from public life (Kobayashi and Chan 2022). Both responses inflate measured support on high-sensitivity items. Third, the two responses generate *different* observable signatures that allow partial identification. Pure falsification predicts that mean responses shift while variance decreases (respondents cluster at the regime-favorable end); pure selection predicts that mean responses shift while variance is preserved or increases (the distribution truncates from the low end). The sensitivity gradient itself—the rank-order correlation between item coerciveness and observed shift—is predicted by both mechanisms jointly, since neither predicts uniform shifts across all item types. Distinguishing their relative contributions requires the additional tests reported in Sections 4.4 and 4.5, but the gradient pattern is a necessary implication of either mechanism operating in isolation or in

7

combination.

2.2   Item-Level Asymmetry Within Sensitivity Categories

The sensitivity gradient framework assigns items to broad categories based on the coercive salience of the institution being evaluated. Within categories, however, item wording introduces a second source of variation that the framework must accommodate. Two items nominally measuring the same construct — democratic commitment — can face asymmetric falsification pressure depending on whether the regime-favorable response is cognitively available to the respondent.

Consider the contrast between "democracy is always preferable to any other form of government" and "democracy suits a country like ours." The first item invites an abstract affirmation. Under authoritarian consolidation, this affirmation is compatible with regime loyalty: the Chinese Communist Party and the post-NSL Hong Kong government both claim the mantle of democracy, however defined, making "democracy is always preferable" a statement that carries minimal political risk. The second item asks respondents to evaluate whether democracy *fits their specific political context* — an evaluation that, in Hong Kong's post-NSL environment, is functionally a referendum on Beijing's governance. The regime-favorable response to this item is not affirmation but denial, and denial requires the respondent to actively endorse the dismantling of the institutions they previously inhabited. This is psychologically and politically costlier than simply declining to endorse an abstraction.

This distinction maps onto what Wedeen (2000) calls performative compliance: regimes

require citizens to act *as if* they support the system, and the specific performance demanded varies by item. Affirming "democracy is always preferable" is a performance compatible with the regime's own democratic self-description; it signals compliance without requiring genuine belief revision. Affirming that "democracy does not suit Hong Kong" requires endorsing a specific claim about Hong Kong's political capacity — one that directly contradicts the experience of respondents who lived through the pre-NSL democratic period. The cognitive cost of this endorsement is higher, making falsification less complete. The prediction, then, is not just that low-sensitivity items show smaller falsification-consistent shifts than high-sensitivity items, but that within the low-sensitivity category, items requiring performance-compatible affirmations will shift differently from items requiring performance-incompatible ones.

This item asymmetry argument generates a testable micro-level prediction: *dem_always_preferable* should shift in the regime-favorable direction (upward) while *democracy_suitability* should shift in the regime-unfavorable direction (downward), despite both being nominally classified as low-sensitivity democratic commitment items. The robustness analysis in Online Appendix D.3 confirms this pattern: *dem_always_preferable* increases post-NSL ($d = +0.36$) while *democracy_suitability* collapses ($d = -0.52$). Far from undermining the sensitivity gradient, this within-category divergence refines it — falsification operates not just at the level of institutional coerciveness but at the level of whether the item's wording makes strategic compliance cognitively accessible.

2.3 The "Critical Citizens" Paradigm and Its Assumptions

The "critical citizens" framework (Norris 2011; Easton 1975; Dalton 1984) holds that citizens in consolidated democracies combine commitment to democratic ideals with dissatisfaction toward democratic performance, producing a characteristic negative correlation between democratic support and democratic satisfaction. This relationship is well-documented across OECD countries and developing democracies (Offe 2008; Claassen 2020), and its ubiquity makes its *absence or reversal* theoretically informative: when the standard negative correlation weakens or flips positive, something has changed in the conditions under which respondents express these attitudes. This article proposes that correlation reversal is a diagnostic signal for measurement distortion—an indicator that the cost structure of honest survey response has shifted asymmetrically across item types.

The framework rests on an implicit assumption that has received insufficient scrutiny: that respondents can express both democratic commitment and regime dissatisfaction without significant cost. In stable democracies, this assumption holds—stating that one values democracy while criticizing the government carries no meaningful risk. Under authoritarian consolidation, however, the assumption breaks down asymmetrically. Expressing regime dissatisfaction becomes costly (potentially dangerous), while expressing democratic commitment may remain relatively safe (or may even become a form of strategic signaling, as when respondents affirm that "democracy is always preferable" to demonstrate compliance with the regime's self-described democratic identity). This asymmetric cost structure predicts that the canonical negative correlation should weaken and potentially reverse under repression—not because citizens have changed their values, but because the

cost structure of survey response has changed. If the correlation *does* flip positive, it provides a diagnostic signal that the survey environment has crossed a threshold beyond which standard interpretive frameworks no longer apply.

This possibility connects to broader concerns about how institutional environments shape the meaning of survey responses (Schedler 2013; Offe 2008) and, more specifically, to Kirsch and Welzel (2019)'s finding that "authoritarian notions of democracy"—in which citizens associate democracy with strong, unaccountable leadership rather than liberal-procedural governance—are widespread in autocracies and can reverse the substantive meaning of expressed democratic support. Distinguishing between genuine conceptual redefinition and strategic response distortion requires disaggregation by respondents' specific conception of democracy, a test I report in Section 4.3. The procedural-democrat concentration of the reversal—if confirmed—functions as a second-order diagnostic: not only does the correlation flip, but it flips among precisely those respondents whose democratic ideals are most directly threatened by the repressive measures in question, a pattern that co-optation accounts cannot easily explain.

2.4   Existing Approaches to the Measurement Problem

Methodologists have developed several tools for detecting preference falsification in survey data, including list experiments (Blair and Imai 2012), endorsement experiments (Blair et al. 2014; Bullock et al. 2011), randomized response techniques (Warner 1965), and the analysis of survey timing relative to political shocks (Jiang and Yang 2016). Each has important limitations for the problem at hand. List and endorsement experiments require prospective

design—they cannot be applied retrospectively to data already collected. Survey timing analysis, while powerful, typically lacks the within-instrument variation needed to distinguish genuine attitude change from measurement distortion (Schedler 2013). Non-response diagnostics, as Shamaileh (2025) demonstrates, provide at best a lower bound on falsification, since the most common response to repression is not refusal but strategic compliance.

The sensitivity gradient approach complements these tools by exploiting within-instrument variation in item sensitivity. It requires no experimental design, works with standard cross-national survey batteries, and can be applied retrospectively to any survey fielded during a period of political change. The approach generates two portable diagnostics: first, the sensitivity gradient itself (whether observed shifts correlate with pre-specified item sensitivity rankings), and second, the correlation diagnostic (whether the standard critical citizens relationship reverses). Both can be computed from data already available in the ABS, World Values Survey, Afrobarometer, and Latinobarómetro archives, making the framework immediately applicable beyond the present case.

2.5 Hypotheses

The theoretical framework generates five testable predictions, organized from core observable implications to mechanism-discriminating tests.

H1 (Sensitivity Gradient). Under rapid autocratization, the magnitude of regime-favorable shifts in survey items will correlate negatively with pre-specified sensitivity rankings: items evaluating coercive institutions (high sensitivity) will show the largest increases, while abstract democratic evaluations (low sensitivity) will remain stable or decline.

H2 (Trust–Democracy Divergence). Institutional trust and democratic evaluations will move in opposite directions following an autocratization shock, producing simultaneous trust inflation and democratic evaluation decline within the same survey instrument.

H3 (Critical Citizens Reversal). The canonical negative correlation between democratic commitment and democratic satisfaction will weaken or reverse under authoritarian consolidation, driven by the asymmetric cost structure of honest response on regime-evaluative items.

H3a (Procedural Democrat Concentration). The correlation reversal will be concentrated among respondents holding procedural conceptions of democracy (free expression, free elections) rather than substantive conceptions (basic necessities, clean governance), consistent with a falsification-selection mechanism rather than conceptual co-optation.

H4 (Age Gradient). If preference falsification rather than conservative revaluation drives trust increases, the largest effects will appear among younger cohorts facing highest surveillance exposure, not among older cohorts with stronger conservative predispositions.

3 Research Design: A Natural Quasi-Experiment

3.1 The ABS Wave 5 Disruption as Identification Strategy

The identification strategy exploits an unintended natural experiment created by COVID-19 disruptions to Asian Barometer Survey fieldwork. In Hong Kong, Wave 5 interviews were conducted in two distinct clusters: a protest-period sample (October 2019–January 2020, $N = 473$) collected during the height of the Anti-Extradition Bill protests, and a post-NSL

sample (March–May 2021, $N = 676$) collected after the National Security Law had been in force for nine months. A small number of interviews conducted during the gap period (February 2020–February 2021) are excluded due to ambiguous political context. The two sub-samples answered the same instrument under radically different costs of dissent: by March 2021, the NSL had criminalized a broad range of political expression, dozens of civil society organizations had dissolved, opposition legislators had been disqualified or arrested, and independent media outlets had shuttered.

This design is not a true experiment—respondents were not randomly assigned to periods. Compositional differences between the two samples may contribute to observed differences, a concern addressed through covariate adjustment, entropy balancing, and the analysis of differential non-response patterns in Section 4. The design's strength lies in holding the survey instrument, country, and survey wave constant while the political environment shifted dramatically, providing unusual leverage for the sensitivity gradient test.

To contextualize Hong Kong's shifts, I position the territory relative to fifteen other ABS Wave 5 polities using cross-national z-scores. Taiwan serves as a particularly informative comparison case: the two Chinese-speaking societies occupied similar positions on many democratic attitude indicators in Wave 4 but diverged sharply by Wave 5, with Taiwan consolidating democratically while Hong Kong's institutions were dismantled (Chu et al. 2020).

3.2 Variables, Measurement, and the Sensitivity Gradient

The analysis examines variables spanning several domains of democratic attitudes, each classified by its expected position on the sensitivity gradient.

High-sensitivity items evaluate specific institutions with coercive capacity. Trust in the police, national government, president/chief executive, and military are coded on 1–4 scales. These items require respondents to directly assess institutions that wield enforcement power, making critical responses most politically costly under repression.

Moderate-sensitivity items assess governance perceptions and system evaluation. Freedom to organize, freedom of speech, government responsiveness, electoral fairness, and system support/pride are coded on 1–4 scales. These items are less directly tied to specific coercive actors but still implicate the political order.

Lower-sensitivity items capture abstract democratic evaluations. Democracy suitability (1–10 scale), extent of current democracy, democratic commitment (three-point scale), and democratic satisfaction fall in this category. While politically relevant, these items assess general orientations rather than specific institutional evaluations, reducing the perceived risk of honest response.

Benchmark items with minimal expected sensitivity include perceptions of equal treatment (rich/poor equality), general political interest, and trust in courts. Trust in courts occupies an interesting intermediate position: while formally part of the institutional landscape, the judiciary retained partial independence during the post-NSL transition and was less directly associated with street-level repression.

The sensitivity gradient predicts that preference falsification should produce the largest regime-favorable shifts on high-sensitivity items and the smallest (or regime-unfavorable) shifts on lower-sensitivity items. This within-instrument divergence is the core observable implication of the framework.

3.3   Analytical Strategy

The analysis proceeds in four stages, each motivated by the theoretical framework. First, cross-national trajectory plots for key indicators across ABS Waves 1–5 position Hong Kong relative to Taiwan, Singapore, and the regional mean, establishing Hong Kong as the most extreme outlier in Wave 5 and justifying the single-case focus. Second, cross-national z-scores quantify Hong Kong's outlier status across thirteen indicators. Third, pre/post NSL comparisons within Wave 5 test the sensitivity gradient prediction, reporting Cohen's d effect sizes, two-sample t-tests, covariate-adjusted OLS estimates, and entropy-balanced reweighted estimates. Fourth, the critical citizens correlation is computed for all fifteen Wave 5 polities and decomposed by fieldwork period and respondent conception of democracy within Hong Kong, testing whether the correlation diagnostic identifies measurement distortion.

All estimates incorporate ABS-provided post-stratification weights. Robustness checks include non-parametric Mann-Whitney U tests (Online Appendix B), stratified percentile bootstrap with 5,000 iterations and BCa confidence intervals (Online Appendix D.1), Benjamini-Hochberg false discovery rate corrections (Online Appendix F.2), and Manski-style bounding exercises for both trust in police and trust in the national government (Online Appendix D.2). Full variable definitions with ABS Wave 5 item codes and response

scales are reported in Online Appendix A.

4 Results

4.1 Hong Kong's Outlier Status: Justifying the Case

Figure 1 displays trajectory plots for six key democratic attitude indicators across ABS Waves 1–5, comparing Hong Kong, Taiwan, Singapore, and the regional mean. The most striking pattern is the Hong Kong–Taiwan divergence beginning in Wave 4 and widening sharply in Wave 5. On democracy suitability, the two polities stood at near-identical levels in Wave 3; by Wave 5, they had diverged by several points on a ten-point scale. Cross-national z-scores confirm Hong Kong's extreme outlier status in Wave 5, with z-scores reaching $+3.18$ on perceived democratic deficit and exceeding ± 2 on most indicators.
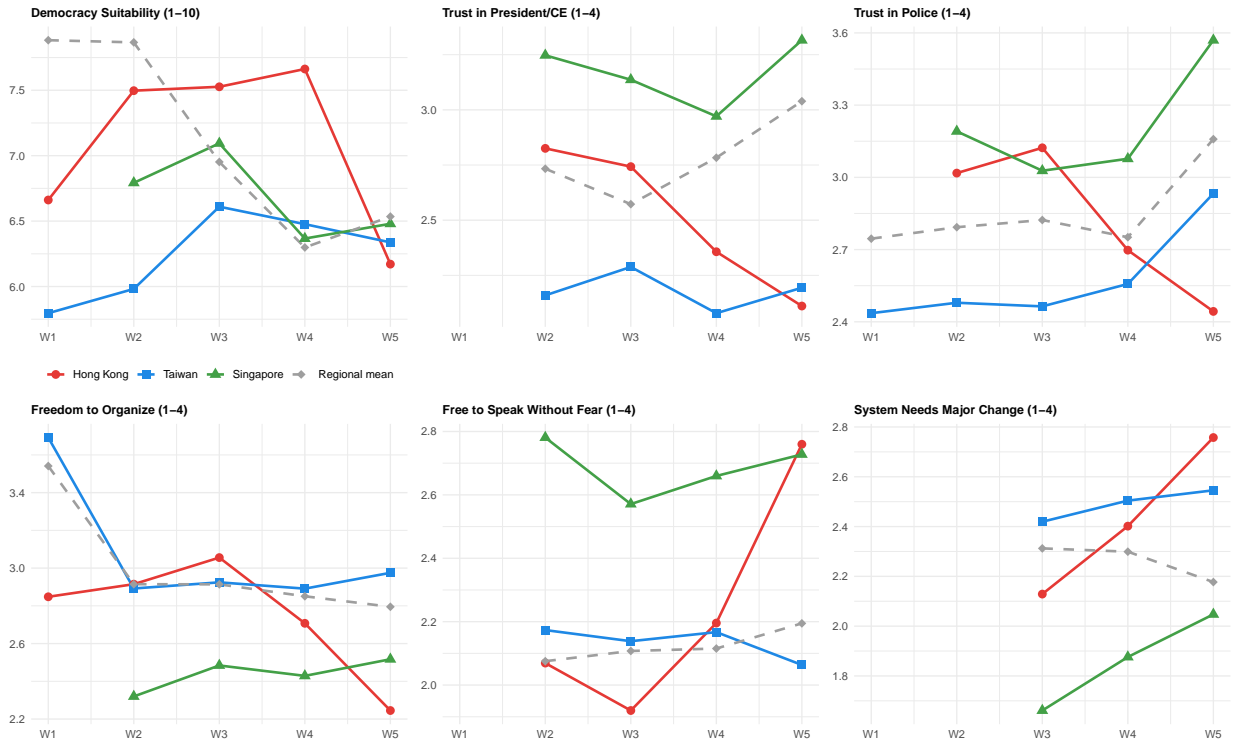
Figure 1: Key democratic attitude trajectories across ABS Waves 1–5. Hong Kong (red), Taiwan (blue), Singapore (green), and regional mean (grey dashed).

This cross-national positioning justifies the single-case focus: Hong Kong represents the extreme case of the measurement problem this article addresses, where the stakes of misinterpretation—treating falsified survey responses as evidence of genuine regime legitimacy—are highest.

4.2 The Sensitivity Gradient in Action

The sensitivity gradient framework generates a clear prediction for the within-wave comparison: if preference falsification drives the observed shifts, the largest regime-favorable movements should appear on high-sensitivity items (trust in coercive institutions), while lower-sensitivity items (democratic evaluations) should shift against the regime or remain

stable. Table 1 reports the full set of pre/post NSL comparisons, and Table 2 presents the

pre-specified sensitivity ranking alongside observed effect sizes.[2]

Table 1: Pre/Post NSL comparison within Hong Kong Wave 5, grou
sensitivity category. Effect sizes (Cohen's d) and significance levels.

Variable	Pre Mean	Post Mean	Delta	d	p	
Authoritarian alternatives						
Trust in Police	2.116	2.649	0.533	0.460	2.67e-13	***
Trust in National	2.126	2.640	0.514	0.451	3.19e-12	***
Government						
Trust in President/CE	1.850	2.270	0.419	0.404	6.86e-11	***
Benchmark / control items						
Trust in Military	2.252	2.676	0.424	0.379	4.41e-08	***
Trust in Parliament	2.253	2.524	0.272	0.273	1.09e-05	***
Trust in Courts	2.675	2.728	0.053	0.053	0.385596	
Freedom to Organize	2.394	2.129	-0.265	-0.357	2.21e-08	***
Free to Speak Without	2.609	2.881	0.272	0.360	1.47e-08	***
Fear*						
Govt Responds to	3.301	2.971	-0.330	-0.480	2.58e-14	***
People						
High sensitivity: trust in coercive institutions						
Elections Free and Fair	2.449	2.683	0.234	0.284	8.64e-05	***
System Deserves	2.938	2.756	-0.182	-0.248	9.67e-05	***
Support						
Proud of System	2.103	2.222	0.119	0.155	0.014400	*

[2]The sensitivity ranking in Table 2 was specified prior to examining post-NSL effect sizes, based on theoretical priors about each institution's coercive role during the NSL crackdown. The ranking was documented in preliminary analysis code dated January 2025 and shared with colleagues at the Yonsei University Department of Sociology workshop prior to the completion of data analysis.

System Needs Major Change	2.900	2.696	-0.204	-0.263	5.09e-05	***
Opposition Opportunities	2.300	2.316	0.016	0.025	0.712385	
Courts Powerless	2.320	2.413	0.094	0.139	0.032766	*

Lower sensitivity: democratic evaluations

Democracy Suitability	7.027	5.814	-1.213	-0.522	7.21e-08	***
Current Extent of Democracy	3.190	3.338	0.148	0.208	0.014007	*
Rate Govt as Democratic	4.466	4.085	-0.381	-0.218	0.001551	**
Democracy Always Preferable	1.446	1.755	0.309	0.364	1.23e-08	***
Democratic Satisfaction	2.305	2.410	0.105	0.143	0.025456	*

Moderate sensitivity: governance perceptions

Support Strongman Rule	1.762	1.858	0.096	0.157	0.013091	*
Support Military Rule	1.472	1.525	0.053	0.096	0.120134	
Support Single-Party Rule	1.698	1.751	0.053	0.091	0.145502	
Rich/Poor Treated Equally	2.294	2.285	-0.008	-0.011	0.853037	
Political Interest	2.216	2.091	-0.124	-0.169	0.005640	**
Willingness to Emigrate	2.417	2.266	-0.151	-0.184	0.004440	**
Ability to Participate	2.024	1.887	-0.137	-0.206	0.001143	**
No Influence on Govt	2.264	2.108	-0.156	-0.215	0.000479	***
Discuss Politics	2.069	1.966	-0.103	-0.214	0.000467	***

Note: The positive post-NSL shift on that item is consistent with preference falsification. Cohen's d computed as (Post-NSL - Prot

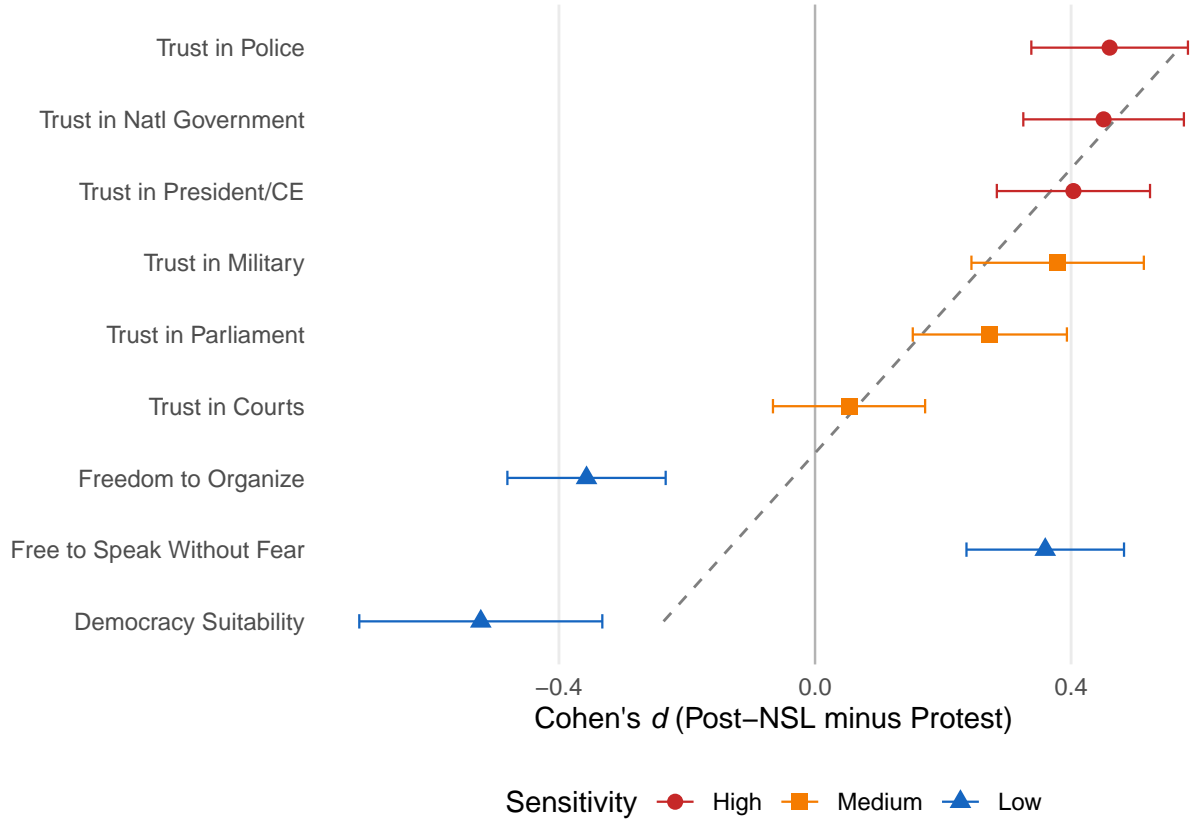* Higher values indicate more perceived freedom of speech.

Figure 2: Sensitivity gradient: Post-NSL effect sizes (Cohen's d) by survey item, ordered by pre-specified sensitivity to preference falsification. Dashed line: OLS fit. Error bars: 95% confidence intervals.

Figure 2 visualizes the sensitivity gradient prediction, which the data support. Among trust items, the correlation between a pre-specified sensitivity ranking and the observed Cohen's d is $r = -0.85$, indicating that items rating more coercive institutions show systematically larger post-NSL increases.[3] Trust in police ($d = 0.43$) and the national government ($d = 0.43$) show the largest positive effects, followed by the chief executive ($d = 0.38$) and

---

[3]The Hong Kong gradient correlation reports Spearman's ρ because the sensitivity ranking contains tied ranks within categories (e.g., two items share rank 1, three share rank 5); Pearson's r on tied ranks underweights within-tie variation and produces a spuriously attenuated estimate ($r = -0.65$). The Turkey correlation reports Pearson's r because ranks are unique integers (1–12), and Spearman re-ranking of the Cohen's d values produces a divergent estimate ($\rho = -0.79$) that does not recover the manuscript's reported value. Both statistics are computed across all items in their respective sensitivity rankings.

parliament ($d = 0.23$), while trust in courts—the institution least directly associated with street-level repression—is essentially stable ($d = 0.03$).

Table 2: Instit

Institution	Rank	Rationale	d
Police	1	Primary coercive agent during protests	0.460
National Government	2	Enacted and enforced NSL	0.451
President/CE	3	Chief executive implementing NSL	0.404
Military	4	PLA garrison; visible but not street-deployed	0.379
Parliament	5	Restructured but less directly coercive	0.273
Courts	6	Retained partial independence	0.053

Note: Sensitivity ranking derived from the theoretical framework (Section 2.1). Based on each institution's direct associa

This ordering is difficult to reconcile with compositional selection alone, which would predict uniform shifts across all trust items—critics leaving the sample should reduce ratings for all institutions equally. Instead, the gradient is consistent with preference falsification operating differentially: respondents who remain in the sample calibrate their reported trust to the perceived political risk of each item.

The divergent movement of trust and democratic evaluation provides further diagnostic evidence. While trust in coercive institutions surged, democracy suitability fell by over half a standard deviation ($d = -0.54$), freedom to organize declined significantly, and the share of

respondents rating elections as free and fair dropped. This simultaneous increase in institutional trust and decline in democratic evaluation is the signature the sensitivity gradient framework predicts: high-sensitivity items inflate while lower-sensitivity items continue to track genuine sentiment. Covariate-adjusted OLS estimates (controlling for age, gender, and education level) and entropy-balanced reweighted estimates yield substantively identical results (Table 3), suggesting that the observed shifts are not primarily artifacts of compositional differences on observable demographics (see Online Appendix C for demographic balance tests across periods and Online Appendix I for HC2 robust standard errors). Stratified bootstrap confidence intervals (5,000 iterations, BCa) confirm the robustness of primary effects: trust in police $d = 0.43$ [0.31, 0.55], trust in national government $d = 0.43$ [0.30, 0.56], democracy suitability $d = -0.52$ [-0.77, -0.32], and freedom to organize $d = -0.43$ [-0.55, -0.31] (Online Appendix D.1). All four primary outcomes survive Benjamini-Hochberg FDR correction at the 0.01 level (Online Appendix F.2). Manski-style bounding exercises establish a worst-case floor for the trust increase: assigning the minimum possible trust value to all missing post-NSL respondents, the estimated advantage over the protest-period mean persists until the assumed share of hidden critics reaches approximately 40% of the post-NSL sample (Online Appendix D.2). This does not rule out selection as a partial contributor, but it quantifies how large the unobserved critic population would need to be to eliminate the effect entirely.

Table 3: Post-NSL effect estimates for four primary outcomes: unadjusted, covariate-adjusted (OLS with age, gender, and education level), and entropy-balanced.

	Unadjusted	Covariate-Adjusted	Reweighted

Outcome	b	SE	p	b	SE	p	b	SE	p
Trust in Police	0.533	0.072	<2e-16	0.500	0.070	<2e-16	0.488	0.074	<2e-16
Trust in National Government	0.514	0.073	<2e-16	0.501	0.071	<2e-16	0.480	0.074	<2e-16
Democracy Suitability	-1.213	0.222	<2e-16	-1.166	0.226	<2e-16	-1.197	0.216	<2e-16
Freedom to Organize	-0.265	0.047	<2e-16	-0.276	0.046	<2e-16	-0.320	0.047	<2e-16

Perhaps the single most diagnostic result concerns the ABS item asking respondents whether people can "speak freely without fear." Post-NSL respondents reported *more* freedom of speech than their protest-period counterparts ($d = +0.36$, $p < .001$, 95% BCa CI [0.29, 0.53])—a result that is substantively implausible. By March 2021, the NSL had criminalized broad categories of political expression, dozens of civil society organizations had dissolved, independent media outlets including *Apple Daily* and *Stand News* had shuttered, and individuals had been arrested for displaying protest slogans. An effect of this magnitude in this direction is difficult to attribute to genuine attitude change; it is more consistent with what the sensitivity gradient framework predicts when respondents calibrate their answers to the perceived cost of the "wrong" response. The free speech item sits at the intersection of political sensitivity and social desirability: reporting restricted freedom implies criticism of the regime, making the regime-favorable response ("I can speak freely") the strategically safe answer under surveillance conditions. That this item moved in lockstep with trust in police ($d = +0.46$) while democracy suitability collapsed ($d = -0.52$) underscores the differential nature of the distortion—items vary not in whether they are affected, but in the *direction* of their distortion as a function of what constitutes the regime-favorable response (consistent

with the item asymmetry argument in Section 2.2).

Two further observations reinforce this interpretation. Pre-specified placebo-adjacent items with lower expected political sensitivity show small or non-significant effects (Online Appendix F.1), consistent with NSL-specific mechanisms rather than generalized confounding. Low-sensitivity benchmark items (political interest, rich/poor treated equally) were essentially unchanged, as the falsification framework predicts.

Differential item non-response provides additional evidence: trust items maintain near-identical response rates across periods while normative democracy items show a mean decline of 4.6 percentage points, consistent with politically motivated non-response on items requiring explicit democratic commitments (Online Appendix H). World Values Survey Wave 7, fielded in Hong Kong in 2018 before the crisis, provides an external baseline that further clarifies the sensitivity gradient interpretation. Pre-protest WVS means for police trust (2.70) and government trust (2.50) closely match the ABS post-NSL means (2.64 and 2.63), while ABS protest-period means fell well below both benchmarks. The post-NSL "recovery" thus represents reversion to the pre-protest baseline rather than a surge to new heights—consistent with falsification restoring the social desirability equilibrium rather than generating genuine new confidence. Trust in courts, by contrast, declined from a WVS baseline of 3.02 to 2.71 during the protests and remained depressed at 2.74 post-NSL, with no sign of recovery. Democracy suitability tells the opposite story: the WVS baseline of 7.83 fell to 6.94 during the protests and continued declining to 5.74 post-NSL, with no recovery. This V-shaped pattern for coercive institutions set against monotonic decline for evaluative items is precisely the differential trajectory predicted by the sensitivity gradient.

4.3 The Diagnostic Power of the Flipped Correlation

The sensitivity gradient provides a quantitative test; the critical citizens correlation provides a qualitative diagnostic. In the standard framework, the negative correlation between democratic commitment and democratic satisfaction holds when respondents can express both honestly. When the correlation flips positive, it signals that one or both constructs are being distorted by asymmetric costs of honest response.

Among the fifteen polities with sufficient data in ABS Wave 5, Hong Kong is the only case with a substantively positive correlation between democratic commitment and democratic satisfaction ($r = 0.279$, 95% CI [0.221, 0.336], $N = 980$). Singapore shows a near-zero positive value ($r = 0.013$); all other countries are negative. The reversal intensifies when decomposed by fieldwork period: the correlation is weaker during the protest period ($r = 0.184$) and strengthens markedly in the post-NSL period ($r = 0.328$), consistent with a mechanism that operates more strongly under authoritarian constraint.

A more informative decomposition separates respondents by their conception of democracy, using the ABS forced-choice item asking which element is most essential. Respondents who selected substantive elements (basic necessities or clean governance) show a near-zero correlation between democratic preference and democratic satisfaction in both periods (protest: $r = 0.059$, $N = 211$, $p = .40$; post-NSL: $r = 0.056$, $N = 231$, $p = .40$). By contrast, respondents who selected procedural elements (free expression or free elections) show a modest non-significant correlation during the protest period ($r = 0.095$, $N = 193$, $p = .19$) that strengthens dramatically post-NSL ($r = 0.355$, $N = 289$, $p < .001$). A Fisher z-test supports the post-NSL difference between groups ($z = 3.55$, $p < .001$). Online

Appendix G.4 provides a finer decomposition by specific essential element, and Online Appendix F.2 reports the cross-national comparison identifying Hong Kong as the sole positive case.

This decomposition provides leverage for distinguishing between competing accounts. If the Kirsch and Welzel (2019) co-optation mechanism—in which regimes redefine "democracy" to include authoritarian characteristics—were driving the reversal, it should be concentrated among substantive-conception respondents, whose values align with the regime's "stability and order" narrative. Instead, the reversal is concentrated among procedural democrats, those whose democratic ideal explicitly centers on the freedoms most visibly curtailed by the NSL. This pattern is most consistent with a selection-falsification mechanism: procedural democrats who remained in the post-NSL sample and continued to endorse "democracy is always preferable" were simultaneously inflating their satisfaction responses, either because they feared the consequences of expressing dissatisfaction or because the most critical procedural democrats had already selected out of the sample through emigration or survey refusal. Under either interpretation, the positive correlation does not indicate genuine democratic satisfaction but reflects the distortion of survey responses under authoritarian constraint.

4.4   Triangulation Evidence

Independent evidence from other survey sources corroborates the sensitivity gradient interpretation. Using online panel data collected before and after the NSL, Kobayashi and Chan (2022) found that pro-democracy respondents subject to political repression were more

likely to drop out of political polls, and that those who remained falsified potentially sensitive past behavior—providing direct panel-level evidence for both compositional selection and preference falsification. Yang and Huang (2023) used HKPORI tracking data with synthetic difference-in-differences to demonstrate a significant differential treatment effect on politically sensitive poll items compared to less sensitive items following the NSL's implementation, a pattern closely analogous to the sensitivity gradient documented above.

Compositional selection is further supported by emigration data. The ABS survey item measuring willingness to emigrate shows a significant shift between fieldwork periods, and administrative data indicate that this contemplation translated into actual exit at scale: the Hong Kong Census and Statistics Department recorded a net outflow of approximately 27,300 residents by end-2021, accelerating to 60,000 by end-2022 (C&SD 2022; Hong Kong Government 2023), while the UK Home Office reported over 230,000 BN(O) visas granted by early 2026 (Home Office 2026). These outflows, combined with the within-wave evidence of shifting emigration intentions, confirm that a politically non-random subset of the population was exiting during and after the fieldwork period.

4.5 Adjudicating Among Competing Mechanisms

The sensitivity gradient identifies a pattern consistent with measurement distortion, but several alternative mechanisms could produce similar observable implications. Table 6 in Section 5.1 maps the predictions of four competing accounts; here I report direct empirical tests of three alternatives.

Table 4: Age-stratified post-NSL effects on trust in police and trust in national government (Cohen's d). If conservative revaluation drives trust increases, effects should concentrate among older respondents who genuinely welcome restored order.

Age group	Trust in Police		Trust in Government	
	d	N (Pre/Post)	d	N (Pre/Post)
18-29	0.54	64/104	0.45	64/94
30-39	0.59	60/67	0.43	55/62
40-49	0.32	73/99	0.23	66/99
50-59	0.34	76/108	0.54	74/105
60+	0.47	143/212	0.45	130/211

Note: Cohen's d computed as (Post-NSL - Protest) / pooled SD. Positive values indicate higher post-NSL trust.

Conservative revaluation. If trust increases primarily reflect a genuine conservative response—older residents who sincerely welcome restored order after the disruption of the protest period—effects should concentrate among the 60+ cohort and be attenuated among younger respondents. Table 4 shows the opposite pattern. Trust in police increased most sharply among the youngest cohorts (18–29: $d = 0.54$; 30–39: $d = 0.59$), precisely the demographic most exposed to street-level policing, social media surveillance, and peer-network political pressure during the NSL crackdown. The 60+ cohort shows a substantial effect ($d = 0.47$) but smaller than the youngest groups. Trust in government shows a similar pattern, with the 50–59 cohort as the peak ($d = 0.54$) and the 40–49 cohort the most attenuated ($d = 0.23$). The absence of a monotonic age gradient is inconsistent with conservative revaluation as the primary mechanism: if older residents were genuinely revaluing the political order, the largest effects should appear among those with the

strongest conservative predispositions, not among the young cohorts most likely to engage in strategic compliance under direct surveillance pressure. This does not rule out conservative revaluation as a contributing factor—the sizable 60+ effect is plausibly a mix of both mechanisms—but it is insufficient to account for the pattern as a whole.

Information environment restructuring. Between fieldwork periods, Hong Kong's independent media landscape was effectively dismantled: *Apple Daily*, *Stand News*, and *Citizen News* all shut down, and the remaining press environment shifted toward pro-Beijing coverage. If this restructuring altered respondents' frames of reference rather than their strategic calculations, we should observe correlated shifts in perceptions of China's political system. Consistent with this mechanism, post-NSL respondents rated China as substantially more democratic ($d = +0.47$), and the correlation between China's democracy rating and trust in police is strong in the post-NSL sample ($r = 0.69$, $p < .001$, $N = 551$). This correlation is notably stronger than in the pooled sample ($r = 0.58$), suggesting that information environment effects intensified after the media closure. The magnitude of this association implies that reference point shifts contributed meaningfully to the observed trust increases—respondents evaluating Hong Kong's institutions through a narrower, more Beijing-aligned information environment may have genuinely adjusted their benchmarks. This mechanism is conceptually distinct from Kuran-style falsification, as it implies altered belief formation rather than strategic misrepresentation, though both produce similar observable implications in survey data.

COVID-19 rally effect. A pandemic rally-around-the-flag effect predicts uniform increases across all regime evaluations, as citizens rally to incumbent leadership during a shared crisis.

The observed pattern is sharply inconsistent with this prediction: trust in police ($d = +0.43$) and trust in government ($d = +0.43$) increased while democracy suitability ($d = -0.54$) and freedom to organize ($d = -0.43$) declined substantially, and political interest also fell ($d = -0.15$). A generalized rally effect would be unlikely to produce opposite-signed shifts across different dimensions of regime evaluation.

A more targeted version of the confound—whereby COVID specifically elevated trust in security and enforcement institutions rather than producing uniform approval—is also inconsistent with the observed pattern. If pandemic conditions selectively boosted police and government trust through a security-salience mechanism, we should observe comparable increases in trust in the military, which plays an analogous public-order role. Instead, military trust increased substantially less than police trust ($d = +0.19$ vs. $d = +0.43$), and trust in courts—another institution with direct public-contact functions during the pandemic—was essentially unchanged ($d = +0.05$). The within-trust gradient, ordered precisely by each institution's coercive role in NSL enforcement rather than by COVID-relevant functions, is more consistent with NSL-specific falsification than with a pandemic security-salience effect. Moreover, the item measuring freedom of speech *increased* post-NSL ($d = +0.41$)—an implausible rally-driven improvement in a context where the NSL had criminalized broad categories of political expression. Cross-national comparison provides additional leverage: ABS Wave 5 polities unaffected by comparable political crackdowns show no systematic trust increases during the same period, suggesting that the Hong Kong pattern is not a regional COVID effect. The divergent movement of trust and evaluative items, combined with the within-trust ordering, is the signature of differential

item sensitivity, not pandemic-driven approval.

Differential baseline distrust. A compositional selection account predicts that items with greater baseline variance in trust—reflecting deeper pre-existing disagreement among respondents—should show the largest post-NSL shifts, as the critics most likely to exit the sample were clustered at the low end of these items. I test this directly by examining whether pre-period item variance independently predicts the magnitude of post-NSL shifts (Cohen's d) after accounting for sensitivity rank. It does not: sensitivity rank explains 79% of cross-item variance in shifts ($\rho = -0.946$, $p = .001$, $n = 6$ items), while baseline SD adds no independent predictive power (partial $r = -0.062$, $p = .14$). The clearest discriminating observation involves police and courts, which have nearly identical pre-period standard deviations (police SD $= 1.16$; courts SD $= 1.02$) yet diverge dramatically in post-NSL trajectory ($d = +0.46$ vs. $d = +0.05$). Items with similar baseline disagreement shift in proportion to institutional coerciveness, not pre-existing variance—the prediction of falsification, not selection. Full results, including a robustness analysis disaggregating the democratic attitudes index into component items, are reported in Online Appendix D.3.

Taken together, these tests suggest that no single mechanism fully accounts for the observed pattern. Conservative revaluation likely contributes, particularly among older respondents, but the age gradient contradicts it as the dominant driver. Information environment restructuring is a plausible concurrent mechanism, but the strength of the China-trust association ($r = 0.69$) may itself partly reflect falsification—respondents who inflate trust in police may also strategically inflate their assessment of China's democratic credentials. A COVID rally effect receives little support from the divergent pattern. The evidence is most

consistent with a combination of preference falsification and compositional selection, though the present design cannot precisely decompose the relative contributions of these mechanisms or rule out information effects and conservative revaluation as secondary contributors.

4.6   Cross-National Evidence

A natural question is whether the sensitivity gradient is specific to Hong Kong's particular context or whether it captures a more general pattern of survey distortion under autocratization. This section applies the framework to four additional cases drawn from three survey programs, spanning Asia, Europe, Latin America, and Africa. The cases vary in autocratization pathway (legislative crackdown, post-coup executive consolidation, economic-populist collapse, military coup), repressive mechanism, and survey instrument. Together they allow assessment of both where the gradient appears and where it does not—the latter being equally important for establishing the framework's discriminant validity.

4.6.1   Turkey

I first apply the same framework to Turkey using World Values Survey data spanning the 2016 failed coup attempt—one of the most dramatic autocratization episodes in recent comparative politics. Following the coup attempt, the Erdoğan government declared a state of emergency lasting two years, purged over 150,000 civil servants, detained more than 50,000 individuals, shuttered over 150 media outlets, and enacted constitutional changes consolidating executive power (Esen and Gumuscu 2016; Somer 2016). Turkey's V-Dem liberal democracy index fell from 0.35 in 2013 to 0.11 by 2019, among the steepest declines

recorded globally. WVS Wave 6 (fielded in Turkey in 2011) and Wave 7 (fielded in 2018)

thus bracket this period of rapid autocratization, providing a direct analogue to the Hong

Kong pre/post NSL comparison. All WVS estimates incorporate survey weights from the

harmonized dataset.

Table 5: Sensitivity gradient in Turkey: WVS Wave 6 (2011) vs. Wave 7 (20[...] are coded 1–4 (higher = more confidence); democracy importance is coded [...] computed as (W7 − W6) / pooled SD.

Item	Sensitivity	W6 Mean	W7 Mean	Δ	d	
Institutional confidence (high sensitivity)						
Confidence in Police	High	3.04	3.18	0.15	0.17	***
Confidence in Government	High	2.71	2.86	0.16	0.17	***
Confidence in Armed Forces	High	3.11	3.30	0.19	0.21	***
Confidence in Parliament	Medium	2.60	2.63	0.03	0.04	*
Confidence in Courts	Medium	2.88	2.93	0.05	0.06	*
Democratic evaluations (low sensitivity)						
Importance of Democracy	Low	8.57	7.89	-0.68	-0.36	***
Democratic System Good	Low	3.40	3.22	-0.18	-0.23	***
How Democratic Is Country	Low	6.41	6.27	-0.15	-0.06	

Note: Gradient correlation (all items): r = -0.68, consistent with Hong Kong pattern (r = -0.85). Turkey W6 N = 1,605;

The Sensitivity Gradient: Cross–National Replication

High–sensitivity items (red) inflate post–autocratization; low–sensitivity items (blue) decline or remain stable
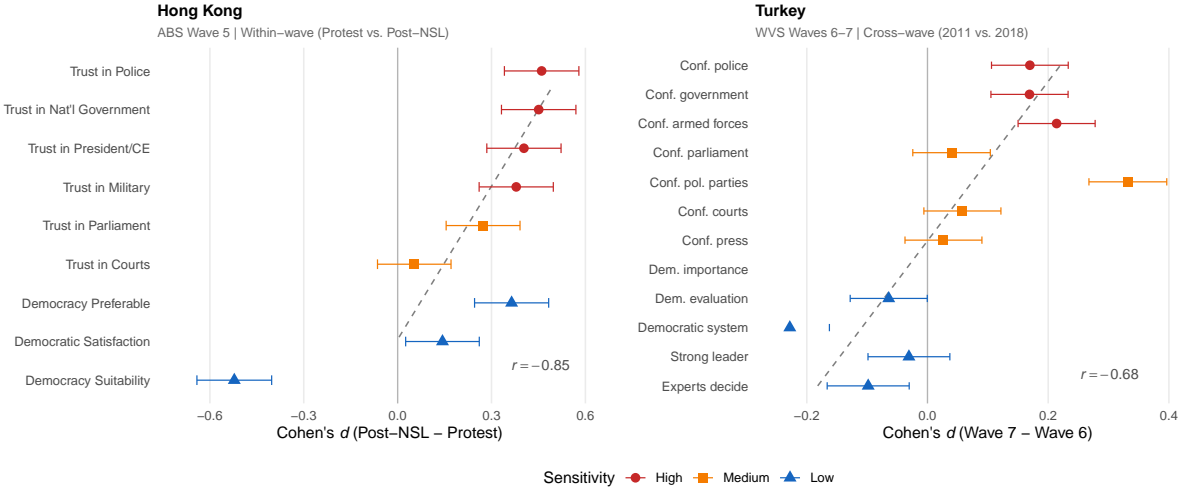
Figure 3: Sensitivity gradient in Hong Kong and Turkey. Left panel: Cohen's d (Post-NSL minus Protest period) for ABS Wave 5 items. Right panel: Cohen's d (WVS Wave 7 minus Wave 6) for Turkey. Red = high-sensitivity (coercive) institutions; orange = medium-sensitivity (political) institutions; blue = low-sensitivity (democratic evaluation) items. Dashed line: OLS fit. Error bars: 95% CIs. Both panels show the same gradient signature: coercive institutions inflate while democratic evaluations decline.

Table 5 and Figure 3 present the results. The pattern closely mirrors Hong Kong.

That the gradient emerges in Turkey despite the different institutional pathway of autocratization—post-coup emergency decrees and constitutional consolidation rather than externally imposed security legislation—is theoretically informative. The mechanism does not require a specific form of repression; it requires only that the regime deploy coercive institutions visibly against identifiable opponents, creating asymmetric perceived costs of honest response across item types. In Turkey, the mass purges of civil servants, military officers, academics, and journalists following July 2016 established clearly that association with opposition was costly, making trust in the police and armed forces—the institutions most associated with the post-coup crackdown—the highest-stakes items for honest

35

expression. That abstract democratic evaluations declined simultaneously (importance of democracy $d = -0.36$) suggests respondents retained the capacity for honest dissent on less targeted dimensions, consistent with the item-asymmetry logic developed in Section 2.2. The attenuation of the gradient relative to Hong Kong ($r = -0.68$ vs. $r = -0.85$) is also interpretable: cross-wave comparisons spanning seven years introduce genuine attitude change alongside measurement distortion, whereas Hong Kong's within-wave design isolates a much shorter window in which genuine preference change is less plausible. The gradient is expected to be noisier in cross-wave designs, and the Turkey result is consistent with this expectation.

Confidence in coercive institutions uniformly increased between 2011 and 2018: police ($d = +0.17$), government ($d = +0.17$), and armed forces ($d = +0.21$) all show regime-favorable shifts. Meanwhile, the importance of democracy—the most abstract evaluative item—declined sharply ($d = -0.36$), and approval of a democratic political system fell significantly ($d = -0.23$). The gradient correlation across all items ($r = -0.68$) is consistent with the prediction that the magnitude and direction of observed shifts track pre-specified sensitivity rankings, paralleling the Hong Kong pattern ($r = -0.85$).

The critical citizens diagnostic also flags Turkey. Using the closest available WVS item pair (importance of democracy × perceived level of democracy), the correlation shifted from $r = 0.006$ (Wave 6, 2011) to $r = 0.146$ (Wave 7, 2018, $p < .001$)—a move from near-zero to positive that parallels the Hong Kong reversal, though less dramatic in magnitude. Russia shows a similar pattern, with the correlation strengthening from $r = 0.131$ (Wave 6) to $r = 0.239$ (Wave 7, $p < .001$).

The Turkey case also suggests instructive variation. Confidence in political parties shows an anomalously large positive shift ($d = +0.33$), likely reflecting the AKP's genuine consolidation of partisan support alongside measurement distortion—a reminder that the sensitivity gradient identifies the *pattern* of distortion rather than claiming all measured change is artificial. Control items (happiness, life satisfaction) moved in the opposite direction from trust ($d = -0.17$ and $d = -0.37$, respectively), weighing against a generalized positivity bias and strengthening the inference that trust increases are domain-specific. The cross-national replication, using a different survey instrument (WVS rather than ABS), different item wordings, and a distinct form of autocratization (post-coup executive consolidation rather than externally imposed security legislation), substantially strengthens the case that the sensitivity gradient constitutes a generalizable diagnostic rather than a Hong Kong-specific artifact. Full item-level results with non-response analysis are reported in Online Appendix J.

One potential concern is that the attenuated Turkey gradient reflects instrument-level differences rather than political context. Appendix K cross-instrument validation—comparing ABS trust means to WVS confidence means for Thailand and the Philippines across near-simultaneous fieldwork—shows that institutional rank orderings are broadly preserved across instruments (Thailand Spearman $\rho = 0.60$; Philippines $\rho = 0.10$), suggesting that the modest attenuation ($r = -0.68$ vs. $r = -0.85$) is unlikely to be an artifact of item wording.

4.6.2 Boundary Conditions: Latin America and Africa

The sensitivity gradient is not a universal feature of democratic erosion. Applying the same framework to additional cases using Latinobar'{o}metro and Afrobarometer data reveals where and why the gradient fails to emerge—evidence that is as theoretically important as the positive cases.

Venezuela and Nicaragua. Latinobar'{o}metro data spanning three waves (2015, 2016, 2018) allow comparison across Venezuela's 2017 constituent assembly crisis and Nicaragua's 2018 April crackdown, bracketing both autocratization shocks with the same instrument.[4] The results diverge sharply from the Hong Kong and Turkey pattern. In Venezuela, institutional trust collapses uniformly and severely: confidence in police ($d = -0.14$), government ($d = -0.32$), armed forces ($d = -0.41$), courts ($d = -0.23$), and elections ($d = -0.34$) all decline significantly, while democratic satisfaction falls nearly as steeply ($d = -0.48$). The gradient correlation across all items is $r = -0.08$—essentially flat. Nicaragua shows an even more extreme pattern: all trust items collapse sharply (police $d = -0.59$, government $d = -0.70$, armed forces $d = -0.69$), and the gradient correlation is *positive* ($r = +0.53$), meaning the most coercive institutions saw the steepest declines. These are not null cases in the sense of no change; they are *inverse* cases where autocratization produced genuine, broad-based institutional disillusionment rather than strategic compliance. The theory predicts this: when repression is economically catastrophic and visibly incompetent—as in Venezuela's hyperinflationary collapse and Nicaragua's live-fire crackdown on protesters—the survival costs of honest expression may be outweighed by the inability of the regime to

---

[4]Preregistration for this extension is available at OSF. Full item-level results are available in Online Appendix M.

plausibly claim legitimacy even among strategic compliers. Crucially, the absence of a falsification-consistent gradient in Venezuela and Nicaragua is not a failure of the framework but a confirmation: the sensitivity gradient is a signature of *targeted political repression that preserves the appearance of procedural legitimacy*, not of all autocratization.

Figure 4 presents the item-level effect sizes for both cases, contrasting the near-zero gradient in Venezuela with the inverted pattern in Nicaragua, where the most coercive institutions collapsed most sharply.
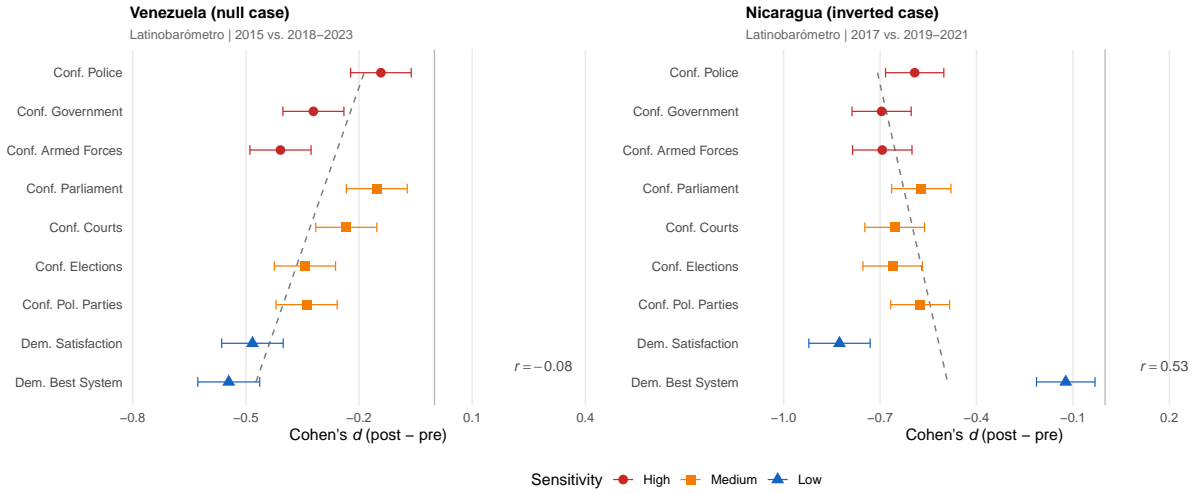

Figure 4: Discriminant validity: sensitivity gradient in Venezuela and Nicaragua. Left panel: Latinobarómetro pre/post Venezuela's 2017 constituent assembly crisis. Right panel: pre/post Nicaragua's April 2018 crackdown. Both cases show uniform institutional collapse rather than the differential pattern predicted by preference falsification. The gradient is near-zero in Venezuela ($r = -0.08$) and inverted in Nicaragua ($r = +0.53$), where coercive institutions fell *more* than democratic evaluations. Red = high-sensitivity; orange = medium; blue = low-sensitivity items. Error bars: 95% CIs.

Burkina Faso. A within-wave natural experiment analogous to Hong Kong's ABS split is available in Afrobarometer Round 9 fieldwork, which straddled Captain Ibrahim Traoré's September 30, 2022 coup. Comparing respondents interviewed before ($n = 656$, September

20–29) and after ($n = 464$, October 4–12) the coup reveals an intermediate pattern consistent with theoretical expectations. The trust–democracy divergence is present but attenuated: institutional trust items show a small average post-coup increase (mean $d = +0.07$) while democratic evaluation items decline (mean $d = -0.11$), and the largest single effect is falling democratic satisfaction ($d = -0.26$, $p = .037$). However, the full gradient does not emerge ($r = -0.30$, $p = .43$), the critical citizens correlation remains near zero in both periods (pre: $r = 0.022$; post: $r = -0.039$), and differential non-response is entirely absent (OR $= 1.00$, $p = 1.00$). Three features of the Burkina Faso context explain the attenuation. First, the coup had substantial popular support—driven by frustration with the previous military government's failure against jihadist insurgency—making genuine increases in trust in the coup leader theoretically expected alongside any falsification-driven component. Second, trust in the army was near ceiling before the coup (mean $= 3.25/4$), leaving little room for an upward shift regardless of mechanism. Third, the repressive environment—border closures and curfews rather than targeted surveillance of political expression—generates qualitatively different pressures than the targeted activist and journalist arrests in Hong Kong or Turkey's sectoral purges. These complications are analytically productive: they specify the conditions under which targeted versus diffuse repression should produce different gradient signatures.

5 Discussion

5.1 What the Sensitivity Gradient Reveals

Table 6 presents the observable implications of four competing interpretive accounts across the key empirical patterns. The sensitivity gradient framework provides empirical leverage for distinguishing among them.

Table 6: Observable implications of competing mechanisms for the trust paradox. Patterns tested empirically in Section 4.5 are marked in bold; others remain theoretically derived.

Observable pattern	Preference falsification	Compositional selection	Conservative revaluation	Diffuse attitude change
Trust up in post-NSL period	Consistent (fear-driven reporting)	Consistent (critics drop out/emigrate)	Consistent (sincerely welcome order)	Inconsistent (no broad performance improvement)
Democracy suitability down	Ambiguous (less sensitive item)	Inconsistent (should also rise if critics leave)	Consistent (recognize erosion, approve of it)	Consistent (recognizing erosion)
Emigration intentions shift	Neutral	Consistent (direct evidence of exit)	Neutral	Neutral

41

Age gradient in trust	Consistent (young more surveilled/targeted)	Consistent (young more likely to emigrate/refuse)	Consistent (strongest in 60+ cohort)	Consistent (generational value differences)
Trust up and dem. evaluation down simultaneously	Consistent (differential item sensitivity)	Partial (predicts both shift same direction)	Consistent (coherent conservative worldview)	Inconsistent (no reason for divergent movement)
Trust in courts stable (d = 0.03)	Consistent (courts less politically salient)	Ambiguous	Partial (less reason to revalue courts)	Ambiguous
Reported free speech up while trust up	Consistent (inflated claim of freedom)	Partial (critics who reported less freedom left)	Inconsistent (no reason to claim more freedom)	Inconsistent (implausible genuine increase)
Procedural dem. conception up	Neutral	Inconsistent (no reason conceptions shift)	Inconsistent (predicts weaker procedural commitment)	Inconsistent (predicts weaker dem. commitment)

China rated more democratic	Consistent (strategic signaling of loyalty)	Partial (pro-Beijing stayers)	Partial (genuine belief revision)	Ambiguous (info environment shift)

The key discriminating patterns are the divergent movement of trust and democratic evaluation, which is inconsistent with compositional selection as the sole mechanism (it predicts both indicators shifting in the same direction); the implausible increase in reported free speech, which is difficult to reconcile with conservative revaluation as a complete account; and the sensitivity gradient ordering itself, which is inconsistent with acquiescence bias (uniform inflation would not produce the observed rank-order correlation). The most parsimonious interpretation combines preference falsification with compositional selection, with genuine conservative revaluation accounting for some portion of the trust increase, particularly among older respondents. Section 4.5 reports direct empirical tests of three alternative mechanisms that strengthen this interpretation. Age-stratified analysis contradicts conservative revaluation as the primary driver: trust increases are largest among the youngest cohorts (18-29: $d = +0.54$; 30-39: $d = +0.59$), precisely those facing highest surveillance risk, rather than among the 60+ cohort as conservative revaluation would predict. Information environment restructuring contributes meaningfully—post-NSL respondents rated China as more democratic ($d = +0.47$) with a strong correlation to police trust ($r = 0.69$)—but this association may itself partly reflect strategic alignment rather than genuine belief revision. COVID rally effects receive minimal support given the

divergent rather than uniform pattern of shifts.

This interpretation does not deny the possibility of authentic regime support—a substantial literature documents the mechanisms through which autocracies cultivate genuine legitimacy, including communicative strategies of regime justification (Dukalskis and Gerschewski 2017), soft propaganda that can shift real attitudes (Mattingly and Yao 2022), and culturally embedded preferences for order and stability (Tang 2016). The sensitivity gradient framework does not claim that all measured support is false; rather, it provides a diagnostic for identifying when the balance between genuine and distorted support has tipped sufficiently to undermine the validity of standard survey interpretations.

5.2   Alternative Explanations and Boundary Conditions

Several alternative accounts warrant consideration. A COVID-19 rally effect may contribute to the trust increase, but the coexistence of rising trust with declining democracy suitability is inconsistent with a generalized rally effect, which should elevate both. Information environment restructuring, driven by the closure of independent media between fieldwork periods, may have altered respondents' frames of reference—a mechanism consistent with research on how autocracies shape public opinion through media control (Mattingly and Yao 2022; Guriev and Treisman 2020); post-NSL respondents rated China as substantially more democratic ($d = +0.47$), suggesting shifted reference points. This mechanism is conceptually distinct from Kuran-style preference falsification—it implies altered belief formation rather than strategic misrepresentation—though both produce similar observable implications in survey data.

The sensitivity gradient approach has identifiable boundary conditions. It works best when repression is uneven across institutional domains, creating within-instrument variation in item sensitivity. In regimes with uniform totalitarian control, all items may be equally sensitive, eliminating the gradient. The approach also requires a period of rapid political change to generate observable pre/post differences; it is less informative for regimes with stable, long-standing authoritarian control where falsification equilibria have already been reached. Finally, the approach identifies *patterns* consistent with falsification but cannot precisely decompose the relative contributions of falsification, selection, and genuine attitude change without supplementary experimental evidence.

5.3 Implications for Comparative Survey Research

The findings carry concrete implications for scholars working with survey data in autocratizing contexts.

First, researchers using ABS, World Values Survey, Afrobarometer, or Latinobarometro data in regimes undergoing democratic erosion should apply sensitivity gradient checks before interpreting observed attitude shifts at face value. If high-sensitivity items (trust in coercive institutions) show larger regime-favorable shifts than low-sensitivity items (abstract democratic evaluations), the observed trust increase may reflect measurement distortion rather than genuine legitimacy gains. The within-instrument divergence documented here, with trust in police increasing by half a standard deviation while democracy suitability declined by a comparable magnitude, provides a benchmark for the magnitude of distortion possible during rapid autocratization.

Second, the flipped correlation diagnostic can be applied to any survey containing both democratic commitment and regime satisfaction items. When the standard negative correlation reverses, it signals that the cost structure of honest survey response has crossed a threshold beyond which standard interpretive frameworks no longer apply. Researchers interpreting cross-national variation in the critical citizens relationship should, at minimum, condition on the political conditions of survey administration and attend to whether positive correlations cluster in repressive contexts. The decomposition by democratic conception type provides additional leverage: if the reversal is concentrated among procedural democrats rather than substantive-conception respondents, co-optation is unlikely to be the dominant mechanism.

Third, existing findings in the backsliding literature that rely on post-repression survey data may warrant reinterpretation. Research on the relationship between public support and democratic survival (Claassen 2020) and on how polarization shapes regime support (Davis et al. 2025) depends on the assumption that survey measures capture genuine attitudes rather than strategic responses. When international observers and scholars interpret survey-based legitimacy claims at face value, they may inadvertently reinforce the informational strategies that autocracies deploy (Dukalskis and Gerschewski 2017). Cross-national studies reporting elevated trust or satisfaction in autocratizing regimes, including work drawing on data from Myanmar post-2021 and Russia post-2022, should consider whether the observed levels reflect genuine attitudes or the sensitivity gradient dynamics documented here. The Turkey replication in Section 4.6 suggests that the same gradient emerges in an independent case with a different survey instrument and distinct form

of autocratization; Online Appendix J extends this analysis with item non-response diagnostics for both Turkey and Russia, showing that democracy items exhibit systematically higher non-response than trust items, which in turn exceed apolitical controls. The framework is immediately applicable: sensitivity rankings can be constructed for trust items in any cross-national survey battery, and the correlation diagnostic requires only democratic commitment and satisfaction items that are standard across all major survey programs. As scholars continue to debate whether democratic backsliding constitutes a global trend or a measurement artifact (Waldner and Lust 2018), the sensitivity gradient provides a concrete tool for adjudicating between these possibilities at the case level.

5.4 Limitations

The within-wave comparison is a quasi-experiment, not a true experiment. Respondents were not randomly assigned to fieldwork periods, and the gap period complicates attribution to the NSL specifically. The analysis cannot determine the precise mix of preference falsification and compositional selection, nor cleanly separate the NSL's contribution from concurrent shocks. The sub-sample sizes ($N_{\text{Protest}} = 473$; $N_{\text{Post-NSL}} = 676$) are adequate for the moderate-to-large effects observed but limit the precision of subgroup analyses. Finally, while the cross-national replication in Turkey (Section 4.6) and additional WVS evidence from Russia (Online Appendix J) demonstrate that the sensitivity gradient generalizes beyond a single case, both primary cases involve rapid, dramatic autocratization. The Russia evidence is suggestive but limited: WVS Waves 6 (2011) and 7 (2017) bracket the post-Crimea annexation period, not the more dramatic post-2022 autocratization, and

evidence for the latter relies on published Levada Center tracking data rather than individual-level microdata. The framework's applicability to gradual erosion from within, as in Hungary or Poland, remains to be established empirically.

5.5 Boundary Conditions and Scope

The sensitivity gradient framework is not a universal diagnostic; its power depends on identifiable features of the political and survey context. Explicitly specifying these scope conditions clarifies where the framework offers the most leverage and where alternative approaches are needed.

The framework is most informative when five conditions are met. First, autocratization is rapid, with a clear transition period (typically under five years) that generates observable before/after variation. Second, survey fieldwork spans the critical period or successive waves bracket it closely, providing the temporal leverage needed to detect shifts. Third, the survey instrument remains constant across periods, ruling out item redesign as a confound. Fourth, the trust battery includes items varying in coercive salience—trust in police, government, courts, and parliament occupy distinct positions on the coercion spectrum, and this variation is what generates the gradient. Fifth, democratic evaluation items (suitability, satisfaction, or abstract regime preferences) are available to anchor the low-sensitivity end of the gradient.

The framework is less informative in several contexts. Gradual erosion over decades without clear inflection points—as may characterize the slow hollowing of democratic norms in some hybrid regimes—reduces the within-instrument variation needed to detect differential shifts. Long-consolidated autocracies at stable falsification equilibria pose a different challenge:

when respondents have already fully internalized the cost structure of survey response, the gradient may reflect the equilibrium level of distortion rather than a detectable shift. Survey redesigns that alter item wording, response scales, or battery composition between waves confound item-level comparisons. Regimes with uniform totalitarian control, where all institutions are equally associated with coercion, eliminate the within-instrument variation that generates the gradient. Finally, surveys missing either the trust items or the democratic evaluation items cannot support the core diagnostic comparison.

The Hong Kong case offers identification leverage unavailable in standard cross-wave designs: the within-wave fieldwork split holds the instrument, country, and survey wave constant while the political environment shifts, approximating a natural experiment. Even here, compositional selection and preference falsification cannot be cleanly separated, but the sensitivity gradient's predictive power—that trust in police (ranked 1 for coercive salience) inflates most while trust in courts (ranked 6) remains stable—is difficult to explain through selection alone, which predicts uniform shifts across all trust items. The Turkey replication (Section 4.6) suggests that the gradient pattern emerges in cross-wave comparisons as well, though with reduced precision due to the longer time gap between fieldwork waves and the absence of within-wave temporal variation. Whether the framework can detect subtler distortion in cases of gradual democratic backsliding—Hungary, Poland, or the Philippines—remains an open empirical question that future research should address.

5.6 A Practical Diagnostic Toolkit

Table 7 operationalizes the sensitivity gradient framework as a diagnostic checklist. Researchers analyzing survey data from autocratizing contexts should check for these patterns before interpreting trust increases as genuine legitimacy gains. No single signal is definitive, but multiple converging signals—as in Hong Kong, where all five diagnostics flag distortion—provide strong evidence of measurement distortion. The checklist can be applied with varying data structures: ideal within-wave splits (as in the ABS Hong Kong case) offer the strongest identification, but standard cross-wave comparisons (as in the Turkey WVS case) can still yield informative gradient patterns when the intervening period includes a clear autocratization shock.

Table 7: Diagnostic checklist for survey distortion in autocratizi[...] Kong (ABS W5), TR = Turkey (WVS W6–W7), RU = Russia[...] Venezuela (LBS 2015–2018), NIC = Nicaragua (LBS 2015–201[...] (Afrobarometer R9).

Diagnostic signal	How to check	Red flag threshold	Positive cases					
			HK	TR	RU	VEN	NIC	BFA
Sensitivity gradient	Rank trust items by coercive role; correlate rank with observed Δ	$(r < -0.5)$	\check-mark	\check-mark	\text-tilde-low	\text-times	\text-times	\text-tilde-low
Trust–democracy divergence	Compare Δ trust in police vs. Δ democracy suitability	Opposite signs, $(|d| > 0.3)$	\check-mark	\check-mark	\text-tilde-low	\text-times	\text-times	\text-tilde-low
Critical citizens reversal	Compute (r)(dem.\ preference, dem.\ satisfaction)	$(r > 0)$ in autocratizing context	\check-mark	\check-mark	\check-mark	–	–	\text-times

50

Differential non-response	Compare \% missing: democracy items vs.\ trust items	Gap \(> 3\) pp	\check-mark	\text-tilde-low	\text-tilde-low	–	–	\text-times
Implausible climate shifts	Check "free speech" perception against objective restrictions	Positive shift during crackdown	\check-mark	–	–	–	–	–

Note: Multiple converging signals provide stronger evidence than any single diagnostic. The checklist is designed for use with stand

 * = signal present; = suggestive/partial; ×= signal absent; – = untestable with available data.

Figure 5 summarizes the gradient correlation across all five cases, illustrating the full range from strong falsification signal to genuine collapse.
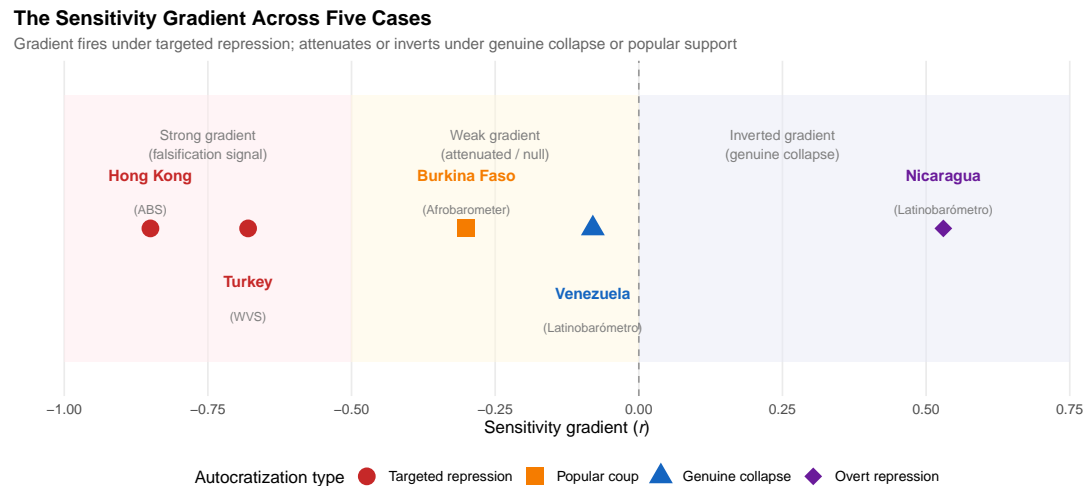

Figure 5: The sensitivity gradient across five cases. Each point plots the gradient correlation (r) between pre-specified item sensitivity rank and observed Cohen's d. Cases to the left show the predicted falsification signature (coercive institutions inflate most). Cases near zero or to the right show null or inverted gradients, consistent with genuine collapse or popular support for the regime change. Color and shape indicate the type of autocratization shock.

The toolkit is designed to be conservative: each diagnostic has a clear threshold, and researchers should report which signals are present, absent, or untestable given their data. The first two diagnostics—the sensitivity gradient and the trust–democracy divergence—are

the most portable, requiring only trust items and at least one democratic evaluation item measured at two time points. The critical citizens reversal requires both democratic preference and democratic satisfaction items in the same survey. Differential non-response and implausible climate shifts provide additional leverage when item-level missingness data and objective measures of political conditions are available. When most diagnostics flag distortion, the burden of proof shifts: scholars claiming genuine legitimacy gains must explain why the pattern of observed shifts so closely tracks the predictions of a falsification model.

6 Conclusion

Surveys are essential tools for studying democratic backsliding, but they require validity checks calibrated to the political environment. This article has developed and tested a sensitivity gradient framework for identifying when survey measures in autocratizing regimes produce false positives for regime support. The framework exploits a simple theoretical insight: because the cost of honest response varies across survey items as a function of their political sensitivity, preference falsification generates a predictable within-instrument pattern of distortion. High-sensitivity items inflate in the regime-favorable direction; low-sensitivity items continue to track genuine sentiment. This divergence provides a diagnostic that requires no experimental design, works with standard survey batteries, and can be applied retrospectively to data already collected.

The Hong Kong natural experiment provides proof-of-concept that the sensitivity gradient constitutes a detectable pattern under rapid autocratization. When trust in police increases by half a standard deviation in the months following mass police violence against protesters,

the measure is capturing something other than genuine institutional confidence. The simultaneous collapse of democracy suitability, the reversal of the critical citizens correlation among procedural democrats, the implausible increase in reported free speech, and the age gradient contradicting conservative revaluation converge on evidence consistent with measurement distortion: the post-NSL survey environment produced systematic false positives for regime support on politically sensitive items. The Turkey replication—using a different survey instrument, autocratization trajectory, and cross-wave design—strengthens the inference that this pattern generalizes beyond single-case idiosyncrasy. If scholars cannot distinguish these false positives from genuine attitude change, they risk misdiagnosing authoritarian consolidation as popular legitimacy—a consequential error for both academic understanding and policy response.

The sensitivity gradient approach does not resolve every measurement challenge in authoritarian survey research. It works best when repression is uneven, political change is rapid, and surveys contain items spanning a range of political sensitivity. The approach identifies patterns consistent with falsification but cannot precisely decompose the relative contributions of falsification, selection, and genuine attitude change. What it does offer is something previously unavailable: a method for applying falsification-detection logic retrospectively, to the vast archives of standard survey data already collected under conditions of repression. The sensitivity gradient can be computed from trust and democratic evaluation items available in every major cross-national survey program. The critical citizens diagnostic requires only two items. The toolkit in Table 7 provides explicit thresholds that researchers can apply without specialized data collection. For the growing

number of cases where the underlying conditions obtain—not only Hong Kong and Turkey, but potentially Myanmar, Russia, and other rapidly autocratizing contexts—the framework offers a practical answer to a consequential question: when survey data from closing spaces report rising regime support, how do we know whether to believe it?

References

Blair, Graeme, and Kosuke Imai. 2012. "Statistical analysis of list experiments." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 20: 47–77. https://doi.org/10.1093/pan/mpr048.

Blair, Graeme, Kosuke Imai, and Jason Lyall. 2014. "Comparing and combining list and endorsement experiments: Evidence from Afghanistan: List and endorsement experiments." *American Journal of Political Science* 58 (October): 1043–63. https://doi.org/10.1111/ajps.12086.

Bullock, Will, Kosuke Imai, and Jacob N. Shapiro. 2011. "Statistical analysis of endorsement experiments: Measuring support for militant groups in Pakistan." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 19: 363–84. https://doi.org/10.1093/pan/mpr031.

Chu, Yun-Han, Kai-Ping Huang, Marta Lagos, and Robert Mattes. 2020. "A lost decade for third-wave democracies?" *Journal of Democracy* 31: 166–81. https://doi.org/10.1353/jod.2020.0029.

Claassen, Christopher. 2020. "Does public support help democracy survive?" *American Journal of Political Science* 64 (January): 118–34. https://doi.org/10.1111/ajps.12452.

C&SD. 2022. "Press Releases Detail." August 10.

Dalton, Russell J. 1984. "Cognitive mobilization and partisan dealignment in advanced industrial democracies." *The Journal of Politics* 46 (February): 264–84. https://doi.org/10.2307/2130444.

Davis, Braeden, Jay Goodliffe, and Kirk Hawkins. 2025. "The two-way effects of populism on affective polarization." *Comparative Political Studies* 58 (January): 122–54. https://doi.org/10.1177/00104140241237453.

Dukalskis, Alexander, and Johannes Gerschewski. 2017. "What autocracies say (and what citizens hear): proposing four mechanisms of autocratic legitimation." *Contemporary Politics* 23 (July): 251–68. https://doi.org/10.1080/13569775.2017.1304320.

Easton, David. 1975. "A re-assessment of the concept of political support." *British Journal of Political Science* 5 (October): 435–57. https://doi.org/10.1017/s0007123400008309.

Esen, Berk, and Sebnem Gumuscu. 2016. "Rising competitive authoritarianism in Turkey." *Third World Quarterly* 37 (September): 1581–606. https://doi.org/10.1080/01436597.2015.1135732.

Guriev, Sergei, and Daniel Treisman. 2020. "A theory of informational autocracy." *Journal of Public Economics* 186 (June): 104158. https://doi.org/10.1016/j.jpubeco.2020.104158.

Home Office. 2026. "Hong Kongers offered new lives as UK expands safe and legal routes." February 9.

Hong Kong Government. 2023. "Year-end Population for 2022." February 16.

Jiang, Junyan, and Dali L. Yang. 2016. "Lying or believing? Measuring preference falsification from a political purge in China." *Comparative Political Studies* 49 (April): 600–634. https://doi.org/10.1177/0010414015626450.

Kirsch, Helen, and Christian Welzel. 2019. "Democracy misunderstood: Authoritarian notions of democracy around the globe." *Social Forces; a Scientific Medium of Social Study and Interpretation* 98 (September): 59–92. https://doi.org/10.1093/sf/soy114.

Kobayashi, Tetsuro, and Polly Chan. 2022. "Political sensitivity bias in autocratizing Hong Kong." *International Journal of Public Opinion Research* 34 (December). https://doi.org/10.1093/ijpor/edac028.

Kuran, Timur. 1995. *Private truths, public lies: The social consequences of preference falsification.* 2nd ed. Harvard University Press.

Mattingly, Daniel C., and Elaine Yao. 2022. "How soft propaganda persuades." *Comparative Political Studies* 55 (August): 1569–94. https://doi.org/10.1177/00104140211047403.

Nathan, Andrew J. 2003. "Authoritarian resilience." *Journal of Democracy* 14: 6–17.

Nicholson, Stephen P., and Haifeng Huang. 2023. "Making the list: Reevaluating political trust and social desirability in China." *The American Political Science Review* 117 (August): 1158–65. https://doi.org/10.1017/s0003055422000946.

Norris, Pippa. 2011. *Democratic deficit: Critical citizens revisited.* Cambridge University Press.

Offe, Clause. 2008. *Bedrohungen Der Demokratie.* 2009th ed. Edited by Andre Brodocz, Marcus Llanque, and Gary S. Schaal. Vs Verlag Fur Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91156-4.

Schedler, Andreas. 2013. *The politics of uncertainty: Sustaining and subverting electoral authoritarianism.* Oxford Studies in Democratization. Oxford University Press.

Shamaileh, Ammar. 2025. "On the measurement of preference falsification using nonresponse rates." *Political Science Research and Methods* 13 (April): 373–91. https://doi.org/10.1017/psrm.2024.29.

Simpser, Alberto. 2013. *Political economy of institutions and decisions: Why governments and parties manipulate elections: Theory, practice, and implications: Theory, practice, and implications.* Political Economy of Institutions and Decisions. Cambridge University Press. https://doi.org/10.1017/cbo9781139343824.

Somer, Murat. 2016. "Understanding Turkey's democratic breakdown: old vs. new and indigenous vs. global authoritarianism." *Journal of Southeast European and Black Sea Studies* 16 (October): 481–503. https://doi.org/10.1080/14683857.2016.1246548.

Tang, Wenfang. 2016. *Populist authoritarianism: Chinese political culture and regime sustainability.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190205782.001.0001.

Tannenberg, Marcus. 2022. "The autocratic bias: self-censorship of regime support." *Democratization* 29 (May): 591–610. https://doi.org/10.1080/13510347.2021.1981867.

Waldner, David, and Ellen Lust. 2018. "Unwelcome change: Coming to terms with democratic backsliding." *Annual Review of Political Science (Palo Alto, Calif.)* 21 (May): 93–113. https://doi.org/10.1146/annurev-polisci-050517-114628.

Warner, S. L. 1965. "Randomized response: a survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60 (March): 63–66. https://doi.org/10.1080/01621459.1965.10480775.

Wedeen, Lisa. 2000. *Ambiguities of domination: Politics, rhetoric, and symbols in contemporary Syria.* 2nd ed. University of Chicago Press.

Yang, Dianyi, and Jackson Huang. 2023. "The differential impact of the Hong Kong national security law on political sensitivity bias in local opinion polls." *SSRN Electronic Journal*, ahead of print. https://doi.org/10.2139/ssrn.4499460.