

RESEARCH ARTICLE

Priced but not paid: The electoral penalty for undemocratic behaviour in candidate-choice experiments, and the geography of a missing literature

Richard Kyalo Mutuku¹

¹ Department of Political Science, University of Nairobi, Nairobi, Kenya

ORCID: <https://orcid.org/0000-0003-0919-6022>

Corresponding author: Richard Kyalo Mutuku, Department of Political Science, University of Nairobi, richardkyalo9@gmail.com

Submitted 13 March 2025 **Accepted** 27 April 2025
Revised 11 November 2025 **Published** 09 January 2026

ABSTRACT

Purpose. Experimental political science has converged on a reassuring finding: when citizens are shown a politician who would gerrymander a district, defy a court, or purge the civil service, they withdraw support. This article asks how large that penalty actually is once the estimates are collected and placed on a common scale, and where in the world the design that produces them has been run.

Methods. A systematic search of candidate-choice and candidate-evaluation experiments that randomly assign a democracy-violating attribute to a politician and measure the effect on electoral support identified twenty-six studies, of which nineteen qualify, across thirteen countries. Seven independent samples report an estimate on the probability scale together with a usable measure of precision. These were harmonised into percentage-point changes in the probability of selection and pooled using a DerSimonian–Laird random-effects estimator, with Knapp–Hartung intervals, leave-one-out diagnostics and a prediction interval.

Findings. The pooled penalty is 12.0 percentage points (95% CI 9.2 to 14.9; Knapp–Hartung CI 7.7 to 16.3). Every individual estimate is negative and every confidence interval excludes zero. Heterogeneity is substantial in magnitude but not in sign ($\tau = 3.0$ points; $I^2 = 80.7$ per cent), and the 95 per cent prediction interval, from 3.8 to 20.3 points, excludes zero. The same literature, however, reports that the penalty falls to roughly three points and loses statistical significance once the violating candidate is a co-partisan whose policy positions the respondent shares. Three of the seven pooled samples come from the United States. One qualifying study has been fielded in sub-Saharan Africa, and it measures electoral violence rather than institutional transgression. Where Global South cases are included in multi-country conjoints, the country-specific coefficient is generally not reported.

Implications. The literature has established with considerable precision what citizens will do when democratic principle is the only thing at stake, and almost nothing about what they will do when it is not. The evidence base sits in consolidated democracies at a moment when forty-four countries, twelve of them in sub-Saharan Africa, are recorded as autocratizing. The article specifies what an African replication would have to test, and why the existing design would have to change to test it.

Keywords: *democratic backsliding; conjoint experiments; vertical accountability; meta-analysis; external validity; African politics*

1. Introduction

In October 2010 the Nigerian National Assembly amended the electoral act; in 2015 an incumbent president lost and left. In 2015 the Burundian constitutional court found that Pierre Nkurunziza's first term did not count; he stood, won, and stayed. Both publics, asked in survey after survey, said they preferred democracy to any other form of government. Only one of them removed the man who did not.

That contrast is the empirical puzzle behind a decade of experimental work. Across thirty-nine African countries surveyed in Afrobarometer's ninth round, 66 per cent of respondents preferred democracy to any alternative, 80 per cent rejected one-man rule, and 72 per cent supported presidential term limits (Afrobarometer 2024). Comparable majorities appear in Europe and the Americas. Yet the V-Dem Institute records forty-four countries autocratizing in 2025 against eighteen democratizing, with 41 per cent of the world's population, some 3.4 billion people, living in a country moving away from democracy (Nord et al. 2026). If citizens hold the preferences they report, the arithmetic of elections should make executive aggrandisement expensive. It manifestly does not.

Experimental political science has offered a precise and influential answer to this puzzle. Beginning with Graham and Svobik's (2020) candidate-choice design, researchers have

placed voters in front of paired hypothetical politicians, randomly assigned one of them a democracy-violating position, and measured how much support that position costs. The results have been remarkably consistent. Citizens do punish. They punish by a lot. And yet, Graham and Svobik argued, the punishment is not nearly large enough to survive contact with partisanship and policy disagreement, so that in a polarised electorate the transgressor still wins.

The design has since been fielded in the United States, Turkey, Finland, Poland, Germany, Spain, the Czech Republic, the United Kingdom, Mexico and South Korea. It has produced, by my count, nineteen qualifying studies in thirteen countries. What it has not produced is a synthesis. Individual estimates are quoted across the literature as though they were interchangeable, but they are reported in at least four incommensurable metrics: percentage-point average marginal component effects, interaction-marginal component effects, recall-rate ratios, and differences on seven-point Likert scales. No one has put them on a common scale and asked what the pooled number actually is, how much of the variation between studies is real, and whether the design's findings travel.

This article does that, and then asks a second question that the first makes unavoidable. Where has the design been run? The answer, set against the geography of contemporary autocratization, is uncomfortable. The literature's inferential base sits almost entirely in consolidated democracies with independent electoral commissions, functioning courts and

low levels of clientelist exchange. Its central finding concerns the behaviour of voters in precisely the settings where democratic breakdown is least likely, and it is close to silent about the settings where breakdown is most common. Sub-Saharan Africa, with twelve autocratizing countries and the largest and longest-running public-opinion infrastructure in the developing world, contributes one qualifying study to the pooled evidence, and that study measures a candidate's reputation for electoral violence rather than a stated intention to weaken an institution.

I argue two things. First, the measured penalty is real, consistent in direction, and large: pooling seven independent samples yields 12.0 percentage points, with a prediction interval that excludes zero. Second, that number is a ceiling rather than an estimate of anything that happens in an election. It is what citizens do when democratic principle is the only consideration on the table, and the same literature demonstrates that it collapses by roughly three-quarters when the other considerations are restored. The gap between the ceiling and the floor is where democratic erosion actually occurs, and the literature has measured the ceiling in thirteen countries and the floor in two.

The article proceeds in seven parts. Section 2 sets out what the candidate-choice design is supposed to measure and the accountability theory it operationalises. Section 3 describes the search, the inclusion criteria, the harmonisation rules and the estimator, and states the hypotheses. Section 4 reports the pooled penalty and its heterogeneity. Section 5 reports what happens to that penalty under realistic choice conditions. Section 6 maps the geography of the evidence and documents a reporting asymmetry that makes the gap larger than a simple count of studies suggests. Section 7 specifies what an African replication would have to do differently, and concludes.

2. What the candidate-choice design measures

The design rests on a deterrence logic that political science inherited, largely without acknowledgement, from the economics of crime. In Becker's (1968) formulation an actor commits an offence when the expected gain exceeds the probability of detection multiplied by the penalty. Applied to executive aggrandisement, the gain is the durable advantage that comes from capturing a court or an electoral commission; the probability of detection is whether voters notice; and the penalty is the vote share lost at the next election. O'Donnell's (1994, 1998) distinction between vertical and horizontal accountability locates the electorate as the residual enforcer. When courts, legislatures and audit bodies have been captured, the voter is the only remaining check, and the whole edifice of delegative democracy turns on whether that check has teeth.

Candidate-choice experiments were designed to measure the size of the tooth. The method itself comes from Hainmueller, Hopkins and Yamamoto (2014), who showed that a fully randomised factorial profile design identifies the average marginal component effect of each attribute without the confounding that plagues observational studies of candidate evaluation. The application to democratic norms adds one attribute to an otherwise ordinary candidate profile: a stated position on gerrymandering, on obeying a court, on prosecuting journalists, on who appoints judges. The coefficient on that attribute is read as the price of transgression.

The reading is not innocent. Abramson, Koçak and Magazinnik (2022) point out that a positive average marginal

component effect does not establish that a majority prefers the attribute, because the average can be driven by a minority with intense preferences. Leeper, Hobolt and Tilley (2020) show that comparing average marginal component effects across respondent subgroups confounds preferences with the arbitrary choice of reference category, and recommend marginal means instead. Both objections matter here, because the substantive question in this literature is almost always a subgroup question: will co-partisans defect? Several of the studies synthesised below report estimates that are vulnerable to exactly these critiques, and I flag them where relevant.

There is a deeper problem, and Graham and Svolik (2020) raised it themselves. A conjoint task strips a vote down to a comparison of attributes with everything else held equal. Real electorates do not face that choice. They face a choice in which the transgressor is usually their own party's candidate, usually agrees with them on the issues they care about, and is usually running against someone they dislike for reasons that have nothing to do with democracy. The experiment measures willingness to punish when punishment is cheap. Democratic erosion happens when it is not.

Two hypotheses follow, and I state them formally because the rest of the article tests them.

Hypothesis 1. Across independent samples and countries, the randomly assigned undemocratic attribute reduces the probability that a candidate is chosen, and the pooled effect is distinguishable from zero.

Hypothesis 2. The pooled effect estimated under unconstrained choice conditions substantially exceeds the effect estimated in the same literature under conditions that restore partisan and policy stakes.

A third question is descriptive rather than hypothetical, and I treat it as such: what is the geographic distribution of the samples that produce these estimates, and how does it compare with the distribution of autocratization?

3. Data and method

3.1 Search and inclusion

I searched for studies that satisfy four conditions. The design must be a survey experiment. It must randomly assign, at the level of an individual politician or incumbent, an attribute that constitutes a violation of democratic procedure: manipulating district boundaries or polling access, defying or packing a court, restricting the press, purging or politicising the bureaucracy, refusing to accept an election result, or using or condoning electoral violence. The outcome must be electoral support, whether a forced choice between candidate profiles, a stated vote intention, or a stated probability of voting for the profile. And the report must be retrievable in full text.

Searches were run between 13 and 14 September 2026 across general web search and publisher platforms, following the reporting logic of PRISMA 2020 (Page et al. 2021) without claiming full PRISMA compliance, since no protocol was pre-registered and no database with a controlled vocabulary was available. Backward and forward citation chasing proceeded from Graham and Svolik (2020), Svolik (2023), Wunsch, Jacob and Derksen (2025) and Frederiksen (2022). Author bibliographies were checked for eighteen scholars associated with the literature, and I searched deliberately for African, Latin American, South Asian and Southeast Asian applications under the names of scholars who field candidate-choice experiments in those regions.

The search identified twenty-six studies. Nineteen satisfy all four conditions. Seven are excluded for stated reasons and are listed in Table 2, because the exclusions are themselves part of the finding: several of the most prominent Global South studies fail the outcome criterion rather than the design criterion, measuring whether respondents regard an action as democratic rather than whether they would vote against the person who took it (Krishnarajan 2023), whether they endorse a norm-eroding policy in the abstract (Simonovits et al. 2022), or choosing between country profiles rather than candidate profiles (Neundorff et al. 2026; Ridge 2024).

3.2 Harmonisation

Of the nineteen qualifying studies, seven independent samples report an effect on the probability scale together with a standard error or confidence interval. These form the pooled set. All were converted to percentage-point changes in the probability of a profile being selected, with the sign fixed so that a negative value denotes a penalty. Three conversions require explanation.

Carey et al. (2022) report the effect of the norm-*respecting* level of each attribute, so the sign is reversed here. They report separately for Democratic and Republican respondents in the public sample; I combined the two by inverse-variance weighting, which yields 10.0 points with a standard error of 1.05. Saikkonen and Christensen (2023) report three separate transgressions; I take disrespecting court decisions, because judicial defiance is the one transgression represented in every study in the pooled set, and report the other two in the sensitivity analysis. Their point estimates are printed in the text but the precision is not: the confidence intervals appear only in the article's Figure 2. The interval used here is read from that figure and cross-checked against a design-based calculation, which for 1,030 respondents completing six forced-choice tasks and a three-level attribute implies a standard error near two percentage points once clustering is allowed for. The two agree closely, and the leave-one-out estimate in Section 4 shows the pooled result does not turn on this entry. Wunsch, Jacob and Derksen (2025) report interaction-marginal component effects for respondent subgroups defined by their understanding of democracy rather than a sample-wide average marginal component effect. I take the estimate for respondents holding a liberal-democratic understanding under the authoritarian framing, which is the closest available analogue, and I re-estimate the pool without it as the principal robustness check, since it is the one entry in the set that is a conditional rather than an unconditional quantity.

Where a study reports multiple experiments on separate samples, each sample enters separately. Svolik (2023) therefore contributes two entries, from Turkish fieldwork in 2018 and 2019. Where a study reports multiple transgressions on the same sample, only one enters the primary pool, to avoid treating correlated estimates as independent evidence.

The remaining qualifying studies could not be pooled, for reasons recorded in Table 2. Yeung (2025) reports recall-rate ratios rather than differences in probability. Kaftan and Gessler (2025) report differences on a seven-point scale. Lewandowsky and Jankowski (2023) report predicted probabilities without a tabulated average marginal component effect. Gutiérrez-Romero and LeBas (2020) report the Kenyan penalty in the text as "nearly nine percentage points" but present the confidence intervals graphically only. Frederiksen (2025) reports a Czech effect moving from -0.13 to -0.18 under a vulnerability prime, without a tabulated standard error

in the accessible text. Several others report no extractable estimate at all: Jacob (2025) reports that Polish disapproval of electoral-integrity violations does not translate into revised vote choice, without a tabulated penalty, and Gidengil, Stolle and Bergeron-Boutin's (2022) American and Canadian average marginal component effects are not retrievable in any accessible version of the text. All are reported in Table 2 and discussed but not pooled.

3.3 Estimator

I use the DerSimonian and Laird (1986) random-effects estimator, which is appropriate because the studies differ in country, transgression type, survey mode and year, so a common underlying effect is not a credible assumption. Heterogeneity is summarised with Q , τ and I^2 following Higgins and Thompson (2002). Because k is small, I report both the conventional Wald interval and the Knapp and Hartung (2003) adjusted interval, which is more conservative and better calibrated at small k . Following Int'Hout et al. (2016) I report a 95 per cent prediction interval, which is the quantity a reader actually wants: the range within which the effect in a new, as-yet-unstudied setting is expected to fall. Leave-one-out estimates are reported for every entry. An Egger et al. (1997) regression test for small-study effects is reported for completeness and interpreted with the scepticism that $k = 7$ demands.

The two existing meta-analyses of candidate-choice and information experiments in political science set the benchmark for what a synthesis of this kind can claim. Schwarz and Coppock (2022) pooled sixty-seven factorial survey experiments on candidate gender and found an average advantage of roughly two percentage points for women. Incerti (2020) pooled corruption-information experiments and showed that the apparent effect of such information on vote share depends heavily on whether the study was a survey or a field experiment. Both had study sets an order of magnitude larger than the one available here, and both could therefore support meta-regression. The present synthesis cannot, and I do not attempt one. What follows is a pooled estimate with heterogeneity diagnostics, not a model of the sources of that heterogeneity.

4. The pooled penalty

Table 1 lists the seven pooled samples and Figure 1 displays them.

Table 1. Studies contributing to the pooled estimate

Study and country	Fieldwork	N	Effect (pp)	95% CI
Svolik (2023), 2019 wave — Turkey	Jun–Aug 2019	573	-23.16	-31.04, -15.28
Wunsch et al. (2025) — Poland	Jul–Aug 2021	1,979	-15.90	-26.48, -5.32
Saikkonen & Christensen (2023) — Finland	May–Jun 2020	1,030	-14.00	-18.00, -10.00
Svolik (2023), 2018 wave — Turkey	May–Jun 2018	n.r.	-13.53	-23.49, -3.56
Graham & Svolik (2020) — United States	Aug–Sep 2018	1,691	-11.70	-13.45, -9.95
Carey et al. (2022) — United States	Mar–Apr 2019	954	-10.00	-12.07, -7.93
Huang et al. (2026) — United States	2023–2024	1,837	-7.00	-8.75, -5.25

Notes: effects are changes in the probability of a profile being selected, in percentage points, sign fixed so that negative denotes a penalty. Transgressions coded: Svolik

2019, purging civil servants who voted against the party; Wunsch et al., executive control of judicial appointments; Saikkonen and Christensen, disrespecting court decisions; Svulik 2018, replacing judges prejudiced against the party; Graham and Svulik, six positions pooled; Carey et al., elected officials not bound by court decisions; Huang et al., antidemocratic platform in a single-race context. Carey et al. combine two partisan subgroups by inverse-variance weighting. Wunsch et al. is an interaction-marginal component effect for respondents holding a liberal-democratic understanding of democracy under the authoritarian framing. Svulik's 2019 estimate is the change in the AKP's two-party vote share among respondents who would turn out to vote; his 2018 sample size is not reported in the accessible text.

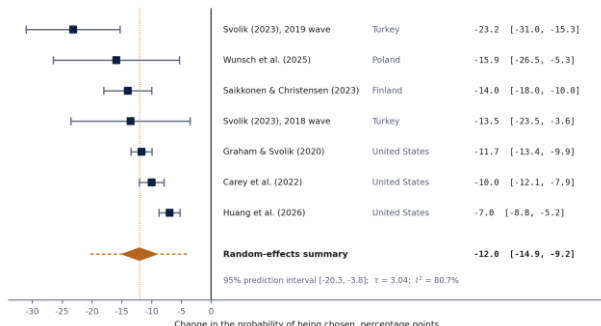


Figure 1. The electoral penalty for undemocratic behaviour across seven independent samples. Squares are point estimates, horizontal lines 95 per cent confidence intervals, the diamond the random-effects summary and the dashed line the 95 per cent prediction interval.

The pooled random-effects estimate is **-12.02 percentage points**, with a Wald interval of -14.90 to -9.15 and a Knapp-Hartung interval of -16.31 to -7.74. Hypothesis 1 is supported without qualification. Every one of the seven estimates is negative; every one of the seven confidence intervals excludes zero; and the fixed-effect estimate, -10.13 with a standard error of 0.51, tells the same story more precisely and less honestly.

Heterogeneity is substantial. $Q = 31.08$ on six degrees of freedom ($p < 0.001$), $\tau = 3.04$ percentage points, and $I^2 = 80.7$ per cent. Four-fifths of the observed variation between studies is not sampling error. But the heterogeneity is entirely in magnitude, not in sign. The 95 per cent prediction interval runs from -20.27 to -3.78, so a new study of this design in a country not yet studied would be expected to find a penalty of between four and twenty points. It would not be expected to find nothing, and it would not be expected to find a reward.

Leave-one-out estimates range from -10.80, when Svulik's 2019 Turkish wave is dropped, to -12.96, when Carey et al. are dropped. No single study drives the result. Removing the one conditional estimate in the set, the Polish interaction-marginal component effect, moves the pool to -11.80 (-14.77 to -8.83) and leaves I^2 essentially unchanged at 83.3 per cent. Substituting Saikkonen and Christensen's other two transgressions changes the pooled estimate by less than two points in either direction: inciting violence, at -27.0, is the largest single estimate in the whole literature, and failing to condemn violence, at -13.0, sits almost exactly on the pooled mean.

Two subgroup contrasts are worth reporting, with the caveat that $k = 3$ and $k = 4$ respectively make them descriptive rather than inferential. The three United States samples pool to -9.56 (-12.40 to -6.72). The four samples fielded elsewhere pool to -16.25 (-20.66 to -11.84). The direction of that difference is the opposite of what a sceptic would predict. If the penalty were an artefact of American respondents' familiarity with the vocabulary of democratic norms, it should be largest in the United States. It is smallest there. The Egger regression intercept is -2.40 ($p = 0.010$), which would ordinarily flag small-study effects, but with seven observations, three of them from a single research programme, the test has almost no

power to distinguish publication bias from genuine between-country variation, and I decline to interpret it as evidence of either.

There is a temporal pattern in the American estimates that deserves a sentence, though it cannot be tested here. Graham and Svulik's 2018 fieldwork yields 11.7 points, Carey et al.'s 2019 fieldwork 10.0, and Huang, Lovett and Helmke's 2023–2024 fieldwork 7.0. The three studies differ in design as well as date, so the decline is not identified. It is, however, consistent with the possibility that the penalty has eroded in the United States over six years of intense partisan conflict over electoral procedure, and it is a question a properly powered replication programme could answer.

5. The penalty under realistic conditions

The pooled estimate answers a question that no voter is ever asked. In a forced-choice conjoint, the respondent compares two strangers who differ on a randomised list of attributes, and the partisan and policy stakes that dominate real voting are either absent or independently randomised. Hypothesis 2 holds that the penalty measured this way substantially exceeds the penalty measured when those stakes are restored, and the literature tests this within its own designs.

Graham and Svulik (2020) report the decisive comparison. Their headline penalty of 11.7 points is estimated across all candidate pairings. When they restrict attention to the condition that resembles an actual general election, in which the violating candidate is the respondent's co-partisan, the platforms diverge, and the respondent's own policy preferences align with their party, the penalty falls to **3.46 points with a confidence interval from -2.64 to 9.54**. It is no longer distinguishable from zero. The same paper reports that support for one's own party falls from 63.1 per cent to 54.8 per cent when that party's candidate violates a democratic principle, and that only **13.1 per cent of respondents are willing to punish a co-partisan** by voting for the other party.

Huang, Lovett and Helmke (2026) reach a compatible conclusion from the opposite direction. Their central contribution is that the choice context itself conditions the penalty: an antidemocratic candidate loses 7.0 points in a single-race contest but 11.2 points when the respondent is voting in multiple races at once and can register disapproval in one without abandoning their party in all. The penalty is not a fixed property of the voter. It is a property of the structure of the ballot.

The non-poolable studies point the same way. Wunsch, Jacob and Derksen (2025) find that the Polish penalty is concentrated among respondents who hold a liberal-democratic understanding of democracy; among respondents holding a majoritarian understanding, the coefficient on executive control of judicial appointments is positive, at 0.127 with a standard error of 0.054, and among those holding an authoritarian understanding it is indistinguishable from zero. The transgression is only a transgression to those who already read it that way. Jacob (2025) pushes the point further. Running a simulated Polish runoff between Law and Justice and the Civic Coalition, he finds that voters disapprove of their own candidate's violations of electoral integrity and that neither the electorate as a whole nor any identifiable segment of it converts that disapproval into a changed vote. Disapproval and defection are different quantities, and only the first of them is what a standard conjoint measures. Kaftan and Gessler (2025) show the same asymmetry across Germany and Poland: opposition to backsliding depends on

which conception of democracy the respondent holds, with a difference of 0.47 points on a seven-point scale ($p < 0.0001$) between violations of electoral and of majoritarian principles. Simonovits, McCoy and Littvay (2022) demonstrate that the same respondents endorse norm-eroding actions 6.7 points more often when their own party would be the one doing the eroding, and that the gap is larger among Republicans (8.8 points) than Democrats (4.9 points). Krishnarajan (2023), though excluded from the pool because his outcome is perceived democraticness rather than vote choice, supplies the mechanism: when confronted with an identical undemocratic act, 62 per cent of left-leaning respondents call it highly undemocratic if a right-wing politician does it and 36 per cent do so if a left-wing politician does.

Hypothesis 2 is supported. The pooled 12-point penalty is what citizens charge when democratic principle is the only item on the invoice. When partisanship and policy are added to the bill, the charge falls by roughly three-quarters and, in the one study that estimates it directly, ceases to be statistically distinguishable from nothing. This is not a criticism of the designs. It is what their authors have been saying. It is, however, routinely lost when the headline figure is quoted, and the distinction matters more outside the settings where these studies were run than inside them, for reasons the next section sets out.

6. The geography of the evidence

Table 2 maps the full set of studies identified, including those excluded from the pool and those excluded from the review.

Table 2. Evidence map: where the design has been fielded

Region	Country	Studies	Status in this synthesis
North America	United States	Graham & Svolik (2020); Carey et al. (2022); Gidengil et al. (2022); Frederiksen (2022, 2024, 2025); Yeung (2025); Huang et al. (2026)	3 pooled; 5 qualifying but non-poolable or pooled-reporting
North America	Canada	Gidengil et al. (2022)	Qualifying; AMCEs not retrievable in full text
Western Europe	Finland	Saikkonen & Christensen (2023)	Pooled
Western Europe	Germany	Lewandowsky & Jankowski (2023); Kaftan & Gessler (2025)	Qualifying; non-poolable metrics
Western Europe	Spain	Muñoz et al. (2026)	Qualifying; marginal means reported graphically
Western Europe	United Kingdom	Frederiksen (2022, 2024, 2025)	Qualifying; country-specific estimates not reported
Central Europe	Poland	Wunsch et al. (2025); Jacob (2025); Kaftan & Gessler (2025); Chiopris et al. (2025)	1 pooled; 3 qualifying, non-poolable
Central Europe	Czech Republic	Frederiksen (2022, 2024, 2025)	Qualifying; effect reported in text without tabulated SE
Eastern Europe	Ukraine	Mazepus & Toshkov (2022)	Qualifying; seven-point scale, vignette
Middle East / Eurasia	Turkey	Svolik (2023); Şaşmaz et al. (2022)	2 pooled; 1 qualifying, full text inaccessible
Middle East / North Africa	Egypt, Morocco	Ridge (2024)	Excluded: country-profile, not candidate-profile, conjoint
Latin America	Mexico	Frederiksen (2022, 2024, 2025)	Qualifying; country-specific estimate not reported

Latin America	Brazil	Cohen et al. (2023)	Excluded: panel study, not experiment
Latin America	Colombia, Peru	Lupu, Rau & Zechmeister (n.d.)	Excluded: outcome is perceived justification, not vote choice
East Asia	South Korea	Frederiksen (2022, 2024, 2025)	Qualifying; country-specific estimate not reported
Southeast Asia	Indonesia, Malaysia, Philippines, Singapore, Thailand	Fossati (2025)	Excluded: no democracy-violation attribute tested
Sub-Saharan Africa	Kenya	Gutiérrez-Romero & LeBas (2020)	Qualifying; CIs graphical only
North America	United States	Simonovits et al. (2022)	Excluded: outcome is endorsement of a norm-eroding policy, not candidate support
North America	United States	Krishnarajan (2023)	Excluded: outcome is perceived democraticness, not candidate support
Cross-national	32 countries	Neundorff et al. (2026)	Excluded: country-profile conjoint

Three features of this map matter.

The base is narrow and it is Northern. Thirteen countries have hosted a qualifying study. Eleven are OECD members; the exceptions are Ukraine and Kenya. Three of the seven pooled samples, and eight of the nineteen qualifying studies, are American. The design has been fielded once in sub-Saharan Africa, never in South Asia, and never in a country classified as a closed autocracy. Set that against the regional distribution of autocratization: twelve autocratizing countries in sub-Saharan Africa, eleven in Eastern Europe, six in Latin America, five in East Asia and the Pacific, four in South and Central Asia (Nord et al. 2026). The single largest concentration of the phenomenon has the thinnest experimental evidence about the mechanism that is supposed to stop it.

The one African study tests a different thing. Gutiérrez-Romero and LeBas (2020) interviewed 1,210 Kenyans nine months after the 2013 general election, asking each respondent to cast two ballots between candidate profiles that varied on a violence attribute and a performance attribute. A candidate rumoured to have used electoral violence lost close to nine percentage points relative to a candidate about whom nothing was said, and a candidate explicitly described as never having used violence gained about eight. This is a genuine and valuable result, and it is the only African estimate in the literature. But electoral violence is a campaign behaviour, not an institutional transgression. Nothing in the African evidence base tells us what Kenyan, Nigerian or Zambian voters would do with a candidate who promised to appoint judges sympathetic to his party, to extend a term limit, or to bring the electoral commission under executive control. Those are the transgressions that actually occur, and the Afrobarometer instrument asks about attitudes toward them without ever putting them in front of a voter as a choice.

Where the Global South is sampled, the estimate is not reported. This is the sharpest finding in the map and it is easy to miss. Frederiksen's five-country conjoint, fielded in September and October 2020 with between 2,350 and 2,999 respondents per country, includes Mexico and South Korea alongside the United States, the United Kingdom and the Czech Republic (Frederiksen 2022, 2025). The data exist. The country-specific coefficient on undemocratic behaviour for

Mexico and for South Korea is not reported in the published article text, and the Cambridge supplementary appendix for the 2022 article contains attribute distributions and a note on deviations from pre-registration but no such table. The 2025 article reports the Czech coefficient numerically, moving from -0.13 to -0.18 under a vulnerability prime, and describes Mexico and South Korea only by the sign and non-significance of an interaction term. Neundorff et al. (2026) field a conjoint in thirty-two countries including Ghana, Kenya, Nigeria, South Africa, Tanzania, Uganda and Zambia, but the profiles are countries rather than candidates, so the design cannot speak to what a voter would do at a ballot box.

The pattern is not a conspiracy, and it is not unusual in comparative work. Multi-country conjoints are typically powered for the pooled estimate and report country-level results in appendices or replication archives. But the consequence is that a reader who wants to know the size of the democratic penalty in Mexico must obtain and re-analyse a replication dataset, while the same quantity for the United States is in the abstract of three separate articles. Over time that asymmetry becomes self-reinforcing: the American estimate is the one that gets quoted, meta-analysed, and taught.

There is a further reason why the missing evidence is not merely a gap to be filled by replication. The design assumes a choice structure that does not obtain in much of the world it omits. A forced choice between two candidate profiles presupposes that the election is the mechanism through which power changes hands, that the ballot is secret, and that the respondent's vote is not embedded in a chain of obligation running through a local broker. Where clientelist exchange structures the vote, as it does across much of sub-Saharan Africa (Bratton and Mattes 2001; Weghorst and Lindberg 2013), the relevant counterfactual is not "would you choose the other candidate?" but "would you forgo what your candidate's network provides?" Where the electoral commission's independence is itself in doubt, as Ferree, Gibson and Long (2014) document for Kenya in 2013, a respondent's stated willingness to punish a transgressor is conditioned on a belief about whether the punishment could be delivered. And where an incumbent's capture of institutions has already altered the media environment, as Levitsky and Way (2010) describe for competitive authoritarian regimes generally, the assumption that the respondent has accurately perceived the transgression, which Krishnarajan (2023) shows is unsafe even in the United States, becomes heroic.

Cheeseman and Klaas (2018) catalogue the repertoire that actually threatens African elections: gerrymandering of constituency boundaries, control of the voter register, the timing and location of polling stations, the deployment of state media, and the selective use of the courts. Posner and Young (2007) documented the institutionalisation of term limits as a constraint that African presidents increasingly respect, and Bermeo (2016) situates the incremental, legalistic character of contemporary backsliding, which is precisely the form that makes it hard for voters to perceive. Each of these is a testable candidate attribute. None has been tested.

7. What an African replication would have to do

The obvious response to Section 6 is to run the Graham and Svobik design in Nairobi, Accra and Lagos and see what comes out. That would be worth doing, and the prediction interval from Section 4 gives it a clear benchmark: a penalty between four and twenty percentage points. But a straight

replication would import the design's central limitation into a setting where that limitation bites harder.

Three modifications follow from the argument above.

First, the transgression must be locally live. "Ignore a court ruling" is a plausible transgression in Warsaw and Washington. In Nairobi, where the Supreme Court annulled a presidential election in 2017, it is the most salient one available, and it should be tested. But it should be tested alongside the transgressions that Cheeseman and Klaas (2018) identify as the practical repertoire, and alongside the term-limit question, which Afrobarometer (2024) shows commands 72 per cent support across thirty countries and which is the single most common site of African constitutional manipulation. A design that tested only the North Atlantic transgression list would produce an estimate that is technically valid and substantively beside the point.

Second, the co-partisan condition must be the headline, not the subgroup. In a two-party American electorate, the co-partisan condition is a moderator. In an electorate where the party system is organised around ethnic or regional coalitions, it is the whole question. A design should therefore randomise the candidate's coalition alignment relative to the respondent's own and report that interaction as the primary estimand, using marginal means rather than subgroup average marginal component effects to avoid the reference-category problem that Leeper, Hobolt and Tilley (2020) identify. Gutiérrez-Romero and LeBas (2020) show the value of this: their violence penalty roughly doubles in magnitude among co-ethnic and co-partisan respondents, which is the reverse of the American pattern and is the kind of finding that a pooled multi-country design would erase.

Third, the cost of defection must be made explicit. The experimental penalty is large because defecting from a profile costs nothing. A design that specifies what the respondent gives up by defecting, whether a development project, a local patron's protection, or simply the chance that their coalition holds the presidency, would estimate something closer to the quantity that matters. This is harder to do cleanly than a standard conjoint, and it introduces deception risks that require careful handling. It is also the only way to move from the ceiling to the floor.

Three limitations of the present synthesis should be stated plainly before the conclusion. The pooled set contains seven samples, which is small enough that the heterogeneity estimate is imprecise and the test for small-study effects is uninformative. Access to full texts was uneven, and five qualifying studies were excluded from the pool because the quantities they report could not be converted, which means the pooled estimate rests on the subset of the literature that reports in the most conventional metric, and that subset may not be representative. No protocol was pre-registered, and a second coder did not independently verify the inclusion decisions, so the coding reflects one reader's judgement about what constitutes a democracy-violating attribute. Each of these limitations argues for treating the 12-point figure as a well-supported order of magnitude rather than a precise parameter.

Conclusion

Citizens punish politicians who say they will violate democratic procedure. Pooling seven independent samples from four countries gives a penalty of 12.0 percentage points, with a prediction interval from 3.8 to 20.3 points that excludes zero and a direction that is consistent across every study yet conducted. That is a genuine finding, and it is not a trivial one.

It means that the electorate is not indifferent, that the raw material of vertical accountability exists, and that a politician who announces a transgression in advance pays for the announcement.

The finding does not, however, support the inference that is usually drawn from it. The same literature shows that the penalty falls to roughly three points, with a confidence interval covering zero, once the violating candidate is a co-partisan whose policies the respondent shares; that only 13 per cent of American respondents will cross party lines to punish one; and that in Poland respondents holding a majoritarian conception of democracy reward the same transgression that liberal democrats punish. The 12-point figure is a measurement of what voters will do when nothing else is at stake. Erosion happens precisely when something else is.

The geography compounds the problem. Thirteen countries have hosted this design, eleven of them OECD members and one of them in sub-Saharan Africa. Forty-four countries are autocratizing, twelve of them in sub-Saharan Africa. Where Global South cases are included in multi-country conjoint, the country-specific estimate is generally not published. The discipline has therefore built a precise instrument, calibrated it carefully, and used it almost exclusively in the places where the disease it measures is rarest.

Three directions follow. The first is a candidate-choice experiment fielded in African electorates on locally live transgressions, with coalition alignment as the primary moderator rather than a subgroup cut. The second is the disaggregation and publication of country-level estimates from the multi-country conjoint already fielded in Mexico, South Korea and elsewhere, which requires no new data collection and would immediately double the geographic reach of the pooled evidence. The third, and the hardest, is a design that prices defection rather than assuming it is free.

Afrobarometer's ninth round found that 66 per cent of Africans prefer democracy, that 72 per cent want term limits enforced, and that 53 per cent would nonetheless accept a military takeover if elected leaders abused power for their own ends (Afrobarometer 2024). The first two figures describe a public that would punish. The third describes a public that has concluded punishment through the ballot does not work. Which of those descriptions is right is an empirical question, and the design that could answer it has been run almost everywhere except where the question is being asked.

References

- Abramson, S. F., Koçak, K., & Magazinnik, A. (2022). What do we learn about voter preferences from conjoint experiments? *American Journal of Political Science*, 66(4), 1008–1020.
- Afrobarometer. (2024). *African insights 2024: Democracy at risk — the people's perspective*. Accra: Afrobarometer.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2), 169–217.
- Bermeo, N. (2016). On democratic backsliding. *Journal of Democracy*, 27(1), 5–19.
- Bratton, M., & Mattes, R. (2001). Support for democracy in Africa: Intrinsic or instrumental? *British Journal of Political Science*, 31(3), 447–474.
- Carey, J., Clayton, K., Helmke, G., Nyhan, B., Sanders, M., & Stokes, S. (2022). Who will defend democracy? Evaluating tradeoffs in candidate support among partisan donors and voters. *Journal of Elections, Public Opinion and Parties*, 32(1), 230–245.
- Cheeseman, N., & Klaas, B. (2018). *How to rig an election*. New Haven: Yale University Press.
- Chiopris, C., Nalepa, M., & Vanberg, G. (2025). A wolf in sheep's clothing: Citizen uncertainty and democratic backsliding. *Journal of Politics*, 87(4), 1272–1287.
- Cohen, M. J., Smith, A. E., Moseley, M. W., & Layton, M. L. (2023). Winners' consent? Citizen commitment to democracy when illiberal candidates win elections. *American Journal of Political Science*, 67(2), 261–276.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188.
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.
- Ferree, K. E., Gibson, C. C., & Long, J. D. (2014). Voting behavior and electoral irregularities in Kenya's 2013 election. *Journal of Eastern African Studies*, 8(1), 153–172.
- Fossati, D. (2025). What drives voting behavior in Southeast Asia? Evidence from candidate choice experiments in five countries. *Journal of Current Southeast Asian Affairs*, 44(3), 436–460.
- Frederiksen, K. V. S. (2022). Does competence make citizens tolerate undemocratic behavior? *American Political Science Review*, 116(3), 1147–1153.
- Frederiksen, K. V. S. (2024). Do partisanship and policy agreement make citizens tolerate undemocratic behavior? *Journal of Politics*, 86(2), 766–781.
- Frederiksen, K. V. S. (2025). Does priming democratic vulnerability make citizens punish undemocratic behavior? *Political Behavior*, 48(2), 635–656.
- Gidengil, E., Stolle, D., & Bergeron-Boutin, O. (2022). The partisan nature of support for democratic backsliding: A comparative perspective. *European Journal of Political Research*, 61(4), 901–929.
- Graham, M. H., & Svobik, M. W. (2020). Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *American Political Science Review*, 114(2), 392–409.
- Gutiérrez-Romero, R., & LeBas, A. (2020). Does electoral violence affect vote choice and willingness to vote? Conjoint analysis of a vignette experiment. *Journal of Peace Research*, 57(1), 77–92.
- Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1), 1–30.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558.
- Huang, K., Lovett, M., & Helmke, G. (2026). The electoral choice context and support for democratic norms. *Proceedings of the National Academy of Sciences*, 123(19), e2521380123.

- Incerti, T. (2020). Corruption information and vote share: A meta-analysis and lessons for experimental design. *American Political Science Review*, 114(3), 761–774.
- IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7), e010247.
- Jacob, M. S. (2025). Citizens as a democratic safeguard? The sequence of sanctioning elite attacks on democracy. *American Journal of Political Science*, 69(2), 455–470.
- Kaftan, L., & Gessler, T. (2025). The democracy I like: Perceptions of democracy and opposition to democratic backsliding. *Government and Opposition*, 60(2), 358–381.
- Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17), 2693–2710.
- Krishnarajan, S. (2023). Rationalizing democracy: The perceptual bias and (un)democratic behavior. *American Political Science Review*, 117(2), 474–496.
- Leeper, T. J., Hobolt, S. B., & Tilley, J. (2020). Measuring subgroup preferences in conjoint experiments. *Political Analysis*, 28(2), 207–221.
- Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge: Cambridge University Press.
- Lewandowsky, M., & Jankowski, M. (2023). Sympathy for the devil? Voter support for illiberal politicians. *European Political Science Review*, 15(1), 39–56.
- Lupu, N., Rau, E. G., & Zechmeister, E. J. (n.d.). *Public tolerance for anti-democratic behavior*. Unpublished manuscript.
- Mazepus, H., & Toshkov, D. (2022). Standing up for democracy? Explaining citizens' support for democratic checks and balances. *Comparative Political Studies*, 55(8), 1271–1297.
- Muñoz, J., Amat, F., Falcó-Gimeno, A., Guinjoan, M., & Rodon, T. (2026). Affective, but not ideological, polarization increases tolerance for undemocratic behavior. *European Political Science Review*. Advance online publication.
- Neundorff, A., Dahlum, S., Frederiksen, K. V. S., & Öztürk, A. (2026). Elections without constraints? The appeal of electoral autocracy across the world. *British Journal of Political Science*, 56, e1.
- Nord, M., Altman, D., Fernandes, T., Good God, A., & Lindberg, S. I. (2026). *Democracy report 2026: Unraveling the democratic era?* (10th ed.). Gothenburg: V-Dem Institute, University of Gothenburg.
- O'Donnell, G. A. (1994). Delegative democracy. *Journal of Democracy*, 5(1), 55–69.
- O'Donnell, G. A. (1998). Horizontal accountability in new democracies. *Journal of Democracy*, 9(3), 112–126.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- Posner, D. N., & Young, D. J. (2007). The institutionalization of political power in Africa. *Journal of Democracy*, 18(3), 126–140.
- Ridge, H. M. (2024). Democratic commitment in the Middle East: A conjoint analysis. *Political Science Research and Methods*, 12(2), 285–300.
- Saikkonen, I. A.-L., & Christensen, H. S. (2023). Guardians of democracy or passive bystanders? A conjoint experiment on elite transgressions of democratic norms. *Political Research Quarterly*, 76(1), 127–142.
- Şaşmaz, A., Yağcı, A. H., & Ziblatt, D. (2022). How voters respond to presidential assaults on checks and balances: Evidence from a survey experiment in Turkey. *Comparative Political Studies*, 55(11), 1947–1980.
- Schwarz, S., & Coppock, A. (2022). What have we learned about gender from candidate choice experiments? A meta-analysis of sixty-seven factorial survey experiments. *Journal of Politics*, 84(2), 655–668.
- Simonovits, G., McCoy, J., & Littvay, L. (2022). Democratic hypocrisy and out-group threat: Explaining citizen support for democratic erosion. *Journal of Politics*, 84(3), 1806–1811.
- Svolik, M. W. (2023). Voting against autocracy. *World Politics*, 75(4), 647–691.
- Weghorst, K. R., & Lindberg, S. I. (2013). What drives the swing voter in Africa? *American Journal of Political Science*, 57(3), 717–734.
- Wunsch, N., Jacob, M. S., & Derksen, L. (2025). The demand side of democratic backsliding: How divergent understandings of democracy shape political choice. *British Journal of Political Science*, 55, e39.
- Yeung, E. S. F. (2025). Dynamic democratic backsliding. *British Journal of Political Science*, 55, e164.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The author declares none.

Data availability. The coded study set, the harmonisation decisions, and the analysis script that produces every estimate reported in Section 4 are available from the author on request and will be deposited in a public repository on acceptance.