Testing Cross-National and Cross-Wave Predictive Transportability of a Three-Item Civic-Norm Score

Held-Out-Country Prediction and No-Refit Transfer in ISSP Citizenship 2004 and 2014

Keiji Yoshimura

Independent Researcher

keiji.yoshimura.research@gmail.com

Integrated Article v0.3.1 - 23 August 2026

Author note

This manuscript reports secondary analysis of de-identified International Social Survey Programme (ISSP) data. The 2014 data were used to generate the hypothesis. Before examining any 2004 individual-level outcome or model result, the 2004 variables, prediction procedure, decision criteria, and analysis sequence were finalized. Because that plan was not deposited in an external time-stamped registry, the manuscript describes the 2004 analysis as prospectively specified rather than preregistered.

Generative AI systems were used for theory stress-testing, code generation and review, reproducibility checks, literature-search assistance, drafting, and document preparation. AI-generated suggestions were not treated as empirical evidence. The author retains responsibility for study design, interpretation, and the final text.

Abstract

Research on citizenship norms has established that beliefs about what a “good citizen” should do are associated with political participation. Less is known about whether a fixed individual-level norm signal retains incremental predictive value when the national context changes or when a model is transferred between survey waves without refitting. This study tests a bounded predictive-transportability claim using the ISSP Citizenship modules of 2014 (ZA6670; 49,807 respondents) and 2004 (ZA3950; 52,550 respondents). Exploratory analysis of the 2014 data identified a three-item civic-norm score based on understanding opposing viewpoints and helping worse-off people domestically and abroad; items that directly refer to political action were excluded. Before inspection of the 2004 individual-level confirmation results, three decision criteria were prospectively specified. Across 38 eligible 2004 held-out country/sample units, adding the three-item score to a broad prospectively specified demographic and political-attitudinal baseline improved equal-unit mean log loss by 0.01023 (95% equal-unit bootstrap interval [0.00683, 0.01367]); its increment exceeded that of a specific two-item duty comparator by 0.01003 [0.00680, 0.01335]; and within-country permutation of the norm score degraded prediction by 0.01450 [0.01217, 0.01690]. A model estimated in 2014 and applied to 2004 without coefficient refitting retained a 0.00983 [0.00672, 0.01292] increment from the norm score. Predictive increments were substantially smaller for reported contentious behavior and institutional contact. An exploratory reverse transfer did not support universal performance in countries absent from the 2004 training domain. These results support only a narrow, non-causal conclusion: the fixed three-item score carries incremental respondent-linked predictive information across many of the tested ISSP country and wave shifts, with substantial heterogeneity and no claim of universal transportability or measurement invariance.

Keywords: citizenship norms; political participation; cross-national prediction; predictive transportability; leave-one-country-out validation; ISSP

1. Introduction

Political participation is connected not only to opportunities and resources but also to citizens' beliefs about what responsible citizenship requires. A prominent literature distinguishes duty-based citizenship from more engaged or self-expressive forms of citizenship. Dalton (2008), for example, argued that changing citizenship norms can alter the repertoire of participation rather than simply produce civic disengagement. Cross-national work subsequently showed that different "good citizen" norms are associated with different modes of participation across countries (Bolzendahl & Coffé, 2013), while later analysis of the ISSP Citizenship modules linked active and cosmopolitan norms to protest activity across 22 democracies represented in the 2004 and 2014 waves (Cenker-Özek et al., 2021).

Those studies establish a substantive relationship between citizenship norms and political participation. The present study asks a different question: does a civic-norm signal estimated in some national contexts improve prediction in an entirely held-out country, and does that signal retain predictive value when a model is transferred between survey waves without refitting its coefficients? Explanatory association and out-of-sample prediction are related but distinct scientific targets (Yarkoni & Westfall, 2017). The distinction is especially important in cross-national research because random respondent splits can allow country-specific patterns to appear in both training and test samples.

The candidate predictor examined here was identified after an earlier exploratory 2014 model failed its own pre-specified country-holdout criteria. During decomposition of that failed model, citizenship-norm variables showed unexpected incremental predictive performance. Because that observation was discovered after examining the 2014 data, all 2014 evidence is treated as exploratory. A separate 2004 confirmation plan was then finalized before any 2004 individual-level outcome or model result was inspected. The earlier model is not reinterpreted as successful and is not part of the present hypothesis.

The resulting Civic Norm Transportability Hypothesis is deliberately narrow. It proposes that three normative judgments—trying to understand people with opposing views, helping worse-off people in one's own country, and helping worse-off people elsewhere in the world—carry incremental predictive information about respondents' self-reported likelihood of being able to attempt counter-action against a law regarded as unjust or harmful, beyond demographic characteristics, political efficacy, institutional evaluation, and political and social trust. The hypothesis is predictive rather than causal and is not a general theory of political behavior.

Throughout this article, transportability is used in a predictive sense: retention of incremental predictive performance after a shift in country or survey wave. It does not refer to causal-effect transportability. The purpose is also not to claim that citizenship norms and participation are newly associated. The narrower question is whether a fixed norm signal continues to add information when the social context used for evaluation differs from the contexts used for model estimation.

The study proceeds in two stages. First, the 2014 ISSP Citizenship II data provide the exploratory origin of the three-item score and the source models for cross-wave transfer. Second, the 2004 ISSP Citizenship data provide the prospectively specified confirmation under leave-one-country-out prediction. Three prospectively specified tests evaluate whether the score (1) improves prediction beyond a broad prospectively specified demographic and political-attitudinal baseline, (2) contributes more than a specific two-item duty comparator that excludes voting, and (3) retains respondent-linked predictive information when the score is shuffled within each held-out country while preserving its country-level marginal distribution. Secondary analyses examine no-refit cross-wave transfer, reported political behavior, internal-consistency diagnostics, rival-explanation robustness, survey-weight sensitivity, and reverse transfer.

2. Methods

2.1 Data sources and study sequence

The study uses two official ISSP Citizenship integrated files distributed by GESIS. ISSP Citizenship 2004 (ZA3950, version 1.3.0) contains 52,550 respondents, and ISSP Citizenship II 2014 (ZA6670, version 2.0.0)

contains 49,807 respondents. The 2014 integrated file does not include Estonia or New Zealand because those country datasets reached the archive after the integrated file had been prepared; GESIS distributes them separately. Source microdata are not redistributed with the reproducibility materials for this study. Because the integrated files sometimes distinguish analytically separate national or subnational samples, the term "country/sample unit" is used for the held-out evaluation unit where appropriate.

The sequence of analysis is central to the interpretation. The 2014 data were used for exploratory discovery. Before any 2004 individual-level confirmation result was inspected, the 2004 construct mapping, outcome coding, model comparisons, country-eligibility rule, bootstrap rule, and decision criteria were finalized. The plan was not externally registered, so it is not described as a preregistration. Before any 2004 outcome/model results were viewed, two implementation corrections were made and documented: legacy ZA3950 variable names were mapped to the intended age, sex, education, and weight variables, and a bootstrap routine was corrected so that the bootstrap interval reported for a decision criterion was the same interval used for its classification. Neither correction changed the constructs, outcome, models, thresholds, or decision rules.

Table 1. Data sources and analytic roles. Eligibility for the 2004 country-holdout analysis required at least 100 valid test cases and at least 20 positive and 20 negative outcome cases.

Component	ISSP 2014 (exploratory/source)	ISSP 2004 (confirmation)
Archive	ZA6670 v2.0.0	ZA3950 v1.3.0
Rows in source file	49,807	52,550
Role	Exploratory discovery; source models for cross-wave transfer	Prospectively specified confirmation
Primary outcome	V45: reported likelihood of counter-action against an unjust or harmful law	V40: corresponding 2004 item
Three-item civic-norm score	V10, V12, V13	V9, V11, V12 (mapped to the same item content)
Cross-national evaluation	Leave-one-country-out exploration across 34 integrated country/sample units	Leave-one-country-out confirmation across 38 eligible country/sample units
Summary unit	Held-out country/sample	Held-out country/sample

2.2 Three-item civic-norm score and comparators

The primary civic-norm score is the mean of three "good citizen" items, each recoded from the valid 1–7 response scale to 0–1 as $(x-1)/6$. The items ask whether a good citizen should (a) try to understand the reasoning of people with other opinions, (b) help people in one's own country who are worse off, and (c) help people elsewhere in the world who are worse off. A respondent-level score was calculated when at least two of the three items were valid; otherwise the score was set to missing. The score deliberately excludes engaged-citizenship items that directly refer to monitoring government, activity in political or social associations, or political/ethical consumption. This restriction was chosen before the 2004 confirmation to reduce direct content overlap with the primary survey response. If the resulting score was missing inside a prediction fold, the model pipeline imputed it with the median calculated from training countries only; no held-out-country value was used to determine that median.

Two comparison scores were retained. A six-item engaged-citizenship score uses the full engaged battery and was calculated when at least four of six items were valid. A two-item duty score averages the tax-compliance and law-obedience items, requires both items, and excludes voting because voting is itself a direct political act. The duty score is a specific comparator; results involving it are not interpreted as evidence about duty norms or moral-duty measures in general. Missing comparison scores entering a prediction model were imputed from the training-fold median only.

2.3 Primary and secondary outcomes

The primary outcome is the self-reported likelihood of being able to attempt counter-action against a law the respondent regarded as unjust or harmful (2014 V45; 2004 V40; archive label: "Unjust law: likeliness of counter-action"). Under the prospectively specified binary coding, valid response categories 1–2 ("very likely" or "fairly likely") were coded 1 and categories 3–4 ("not very likely" or "not at all likely") were coded 0; invalid responses were treated as missing. The prepared 2004 data contained 46,580 respondents with a valid primary outcome. This item does not isolate willingness, behavioral intention, perceived efficacy, opportunity, or any other latent psychological mechanism, so the manuscript refers to it descriptively as reported counter-action likelihood

rather than as a validated willingness or readiness construct. Secondary outcomes were (a) reported past-year contentious behavior, coded 1 if the respondent reported either boycotting products (2004/2014 V18) or taking part in a demonstration (V19) in the past year and 0 for other valid response categories, and (b) reported past-year institutional contact, coded 1 if the respondent reported contacting a politician (V21) in the past year and 0 for other valid categories. In 2004 the valid Ns were 51,748 for the contentious-behavior composite and 51,040 for institutional contact; 38 country/sample units met the same minimum test-set event-count rule used for the primary analysis.

2.4 Baseline prediction model

The baseline model was prospectively specified as a broad predictive comparator for the 2004 confirmation. It includes age, sex, educational degree, internal political efficacy, external political efficacy, political trust, social trust, and a five-item institutional-evaluation composite. The composite uses politician self-interest (2014 V50; 2004 V44), election honesty (2014 V58; 2004 V55), election fairness (2014 V59; 2004 V56), public-service commitment to serving people (2014 V60; 2004 V57), and public-service corruption (2014 V61; 2004 V59), all oriented so that higher values indicate a more negative institutional evaluation. Each component was mapped to 0–1 and the composite was calculated when at least three of five components were valid. Four interaction terms—B×I, B×E, I×E, and B×I×E, where B is the institutional-evaluation composite, I internal efficacy, and E external efficacy—were retained from the 2014 discovery-stage predictor set. The baseline is not interpreted as a validated causal model; it is a broad predictive comparator designed to ask whether the civic-norm score adds information beyond the specified demographic and political-attitudinal predictors.

2.5 Leave-one-country-out prediction

For each leave-one-country-out fold, one country/sample unit was held out in its entirety. Country fixed effects were not used because a completely held-out country has no estimated country-specific parameter. Base predictors and any missing score variables were imputed using medians calculated only from the training countries. Interaction terms were reconstructed after imputation from the imputed, unstandardized B, I, and E values so that they remained algebraically consistent. Predictors were then standardized using training-fold means and standard deviations only. L2-regularized logistic regression was fitted with scikit-learn LogisticRegression (solver="newton-cholesky", C=1.0, max_iter=2000) on the training countries and evaluated in the held-out country using binary log loss. Predictive improvement is reported as log loss of the baseline model minus log loss of the augmented model, so positive values favor the augmented specification.

Held-out country/sample-unit predictive improvements were given equal weight in the primary summary rather than weighted by national sample size. Uncertainty for the prospectively specified confirmation tests used 10,000 bootstrap resamples of the 38 eligible held-out country/sample-unit effects with equal sampling probability. These intervals summarize variation of the equal-unit predictive summary across the tested evaluation units; because leave-one-country-out folds share much of their training data, they should not be interpreted as design-based confidence intervals for a randomly sampled superpopulation of countries. The primary model fitting and evaluation were unweighted. In a post-confirmation sensitivity analysis, the archive-provided survey weight was normalized within each country as raw weight divided by the mean valid positive weight for that country. Missing normalized weights were assigned unit weight (1.0). The normalized weights were used in logistic-regression fitting and in held-out log-loss calculation; median imputation and standardization remained unweighted.

2.6 Prospectively specified 2004 confirmation tests

The 2004 confirmation required all three of the following criteria to be met. First, adding the three-item civic-norm score to the baseline had to improve equal-unit mean held-out log loss, with a 95% equal-unit bootstrap interval wholly above zero. Second, the incremental improvement from the three-item score had to exceed the incremental improvement from the specific two-item duty comparator, again with the interval wholly above zero. Third, the trained baseline-plus-norm model was retained while the civic-norm score was permuted 100 times within each held-out country. This preserves that country's marginal score distribution but breaks the original respondent-score correspondence and, consequently, its joint structure with respondent-level baseline variables. The observed assignment had to outperform the within-country permutations with an interval wholly above zero. This test can show that country-level marginal score distributions alone do not account for the

predictive advantage; it is not a causal demonstration of an individual-level mechanism. The six-item engaged-citizenship score was evaluated as a supportive comparison and could not compensate for failure of any required criterion.

2.7 No-refit cross-wave transfer

Before inspection of the 2004 individual-level confirmation results, three models were estimated on all eligible 2014 cases: the baseline model, the baseline plus the three-item civic-norm score, and the baseline plus the six-item engaged-citizenship score. The 2014 imputation medians, scaling parameters, feature order, intercepts, and coefficients were then applied to the mapped 2004 data without coefficient refitting. The main transfer quantity was the equal-unit mean improvement of the three-item norm model over the corresponding baseline model. A separate analysis evaluated 2004 countries absent from the normalized 2014 integrated country set. The chronological direction is backward only because the hypothesis was discovered in the 2014 data; this analysis is a model-transfer test, not a forecast of the past. The no-refit transfer claim concerns the incremental predictive contribution of the civic-norm score relative to an equally transferred baseline; it is not a claim that the full 2014 model retained invariant absolute calibration in 2004.

2.8 Secondary and post-confirmation analyses

Secondary analyses evaluated the two reported behavioral outcomes and internal consistency of the three-item score. After the prospectively specified 2004 result was known, additional robustness analyses tested simple rival explanations. These added political interest (2014 V47; 2004 V42), response efficacy (2014 V46; 2004 V41), and prior measured political behavior to the baseline; decomposed the three score items; compared log loss, Brier score, and area under the receiver-operating-characteristic curve (AUC); repeated the primary comparison using normalized survey weights; and performed an exploratory reverse no-refit transfer from 2004 to 2014. These post-confirmation analyses are treated as robustness or boundary-setting evidence and cannot retroactively alter the prospectively specified classification.

3. Results

3.1 Prospectively specified 2004 confirmation

All three prospectively specified confirmation criteria were met. Across 38 eligible held-out country/sample units, adding the three-item civic-norm score to the baseline improved equal-unit mean log loss by 0.01023 (95% equal-unit bootstrap interval [0.00683, 0.01367]). The held-out-unit improvement was positive in 29 of 38 units. Thus the three-item score added predictive information beyond demographics, political efficacy, institutional evaluation, political and social trust, and the retained interaction terms.

The comparison with the specific two-item duty score was also positive. The incremental improvement from the three-item civic-norm score exceeded the increment from that duty comparator by 0.01003 (95% equal-unit bootstrap interval [0.00680, 0.01335]); this contrast was positive in 31 of 38 held-out units. The result shows that this particular tax-compliance/law-obedience comparator did not produce a similar increment; it does not rule out duty norms or other moral-norm scores in general.

The within-country permutation test also met its criterion. Shuffling the three-item score within each held-out country while preserving that country's marginal score distribution increased log loss by an equal-unit mean of 0.01450 (95% equal-unit bootstrap interval [0.01217, 0.01690]). The observed respondent-score correspondence outperformed the shuffled correspondence in 36 of 38 units. The result is consistent with respondent-linked predictive information beyond country-level score distributions alone; it does not by itself identify a causal or psychological mechanism.

The supportive six-item engaged-citizenship score improved prediction by 0.01711 (95% equal-unit bootstrap interval [0.01293, 0.02161]). Under the decision rule fixed before inspection of the 2004 confirmation results, all prospectively specified criteria were met.

Table 2. Prospectively specified ISSP 2004 confirmation results. Intervals are 95% equal-unit bootstrap intervals. Positive values indicate better held-out-country prediction after adding the civic-norm score, a larger increment than the specific duty comparator, or worse prediction after within-country shuffling.

Prospectively specified test	**Quantity**	**Mean difference**	95% equal-unit bootstrap interval	Interpretation
Increment beyond baseline	Baseline log loss − norm-augmented log loss	0.01023	[0.00683, 0.01367]	Criterion met under prospectively specified rule
Comparison with duty score	Civic-norm increment − duty-score increment	0.01003	[0.00680, 0.01335]	Criterion met under prospectively specified rule
Within-country permutation	Shuffled-score log loss − observed-score log loss	0.01450	[0.01217, 0.01690]	Criterion met under prospectively specified rule
Six-item engaged score (supportive)	Baseline log loss − six-item-score model log loss	0.01711	[0.01293, 0.02161]	Supportive

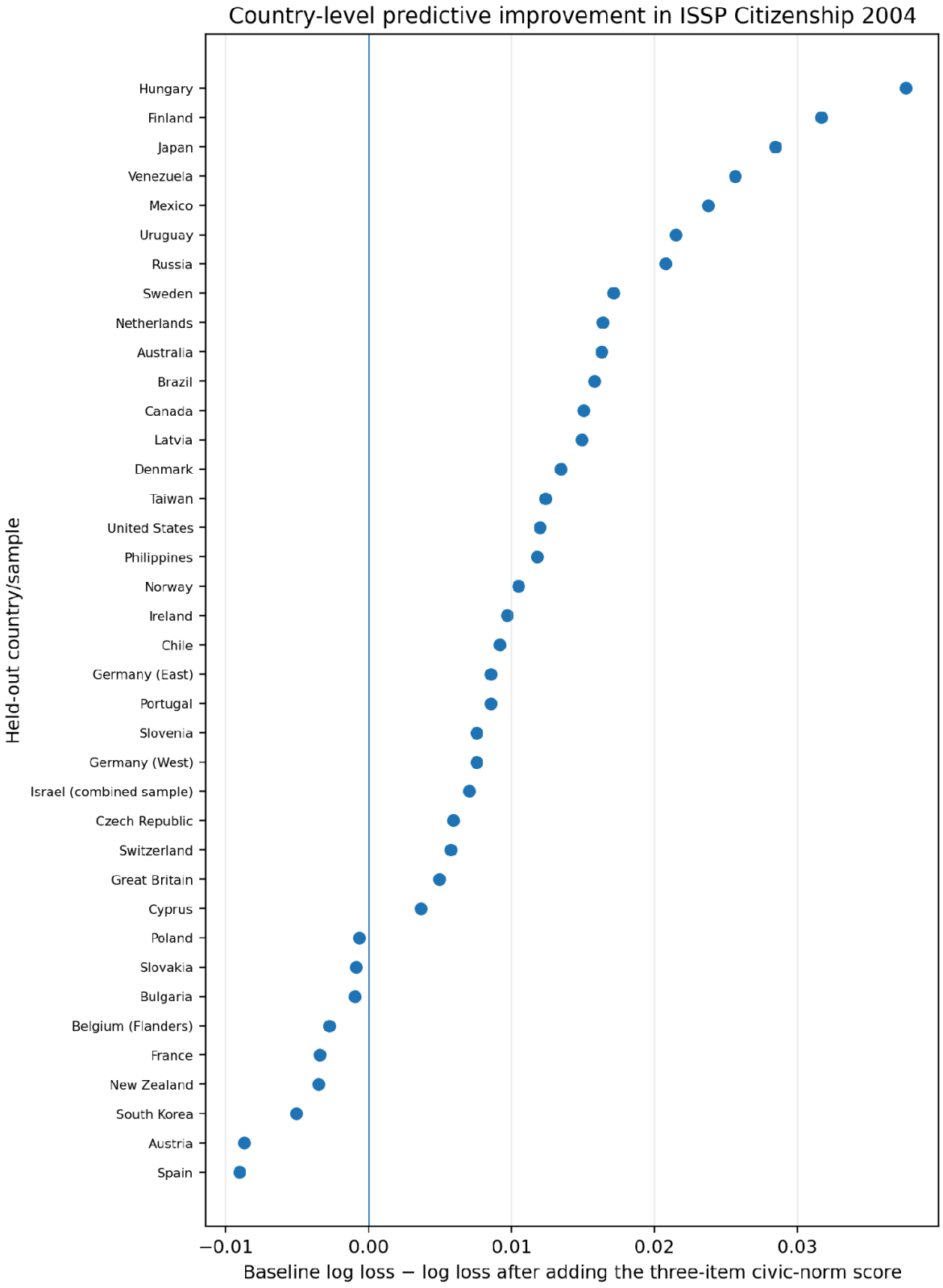

Figure 1. Held-out country/sample-unit predictive improvement from adding the three-item civic-norm score in ISSP Citizenship 2004. Positive values indicate lower held-out-country log loss after adding the score to the baseline. The average improvement is positive, but nine of 38 country/sample units are non-positive, demonstrating heterogeneity rather than universal country-level success.

3.2 Country heterogeneity and sensitivity of the summary mean

The confirmation result was heterogeneous. The baseline-to-norm improvement was positive in 29/38 units, the contrast with the specific duty comparator in 31/38, and the within-country permutation advantage in 36/38. Recomputing the equal-unit summary after omitting each held-out evaluation result in turn left the mean baseline-to-norm improvement between approximately 0.00949 and 0.01075. Thus the positive equal-unit mean was not attributable to a single held-out evaluation result. This leave-one-summary-unit-out diagnostic does not remove that unit from the training data of all other folds and therefore is not interpreted as a full training-influence analysis.

3.3 No-refit 2014→2004 cross-wave transfer

The cross-wave transfer also retained the three-item norm signal. A model estimated in 2014 and applied to 2004 without coefficient refitting improved 2004 prediction by an equal-unit mean of 0.00983 (95% equal-unit bootstrap interval [0.00672, 0.01292]) relative to the corresponding transferred baseline, with positive improvement in 32/38 units. Among the ten 2004 countries absent from the normalized 2014 integrated country set, the mean improvement was 0.01052 (95% interval [0.00462, 0.01637]), positive in 8/10 countries. Because the target-wave coefficients, training medians, scales, and feature order were not refitted, this provides evidence that the predictive norm increment was not limited to within-wave re-estimation.

3.4 Internal-consistency diagnostics

In 2004, the three-item civic-norm score had pooled Cronbach alpha = 0.708; the median country alpha was 0.688, and the minimum was 0.474. No country had alpha below 0.40. These diagnostics indicate that the three-item score was not grossly incoherent in the tested samples, but they do not establish configural, metric, or scalar measurement invariance across countries or waves.

3.5 Outcome specificity

The incremental signal was substantially smaller for the two reported political-behavior outcomes. For past-year contentious behavior, adding the three-item score improved log loss by 0.00209 (95% equal-unit bootstrap interval [0.00008, 0.00414]). For institutional contact, the improvement was 0.000878 (95% interval [0.000184, 0.001603]). Both estimates were positive but much smaller than the 0.01023 improvement for the primary reported counter-action likelihood. This attenuation argues against interpreting the finding as a general-purpose predictor of realized political participation.

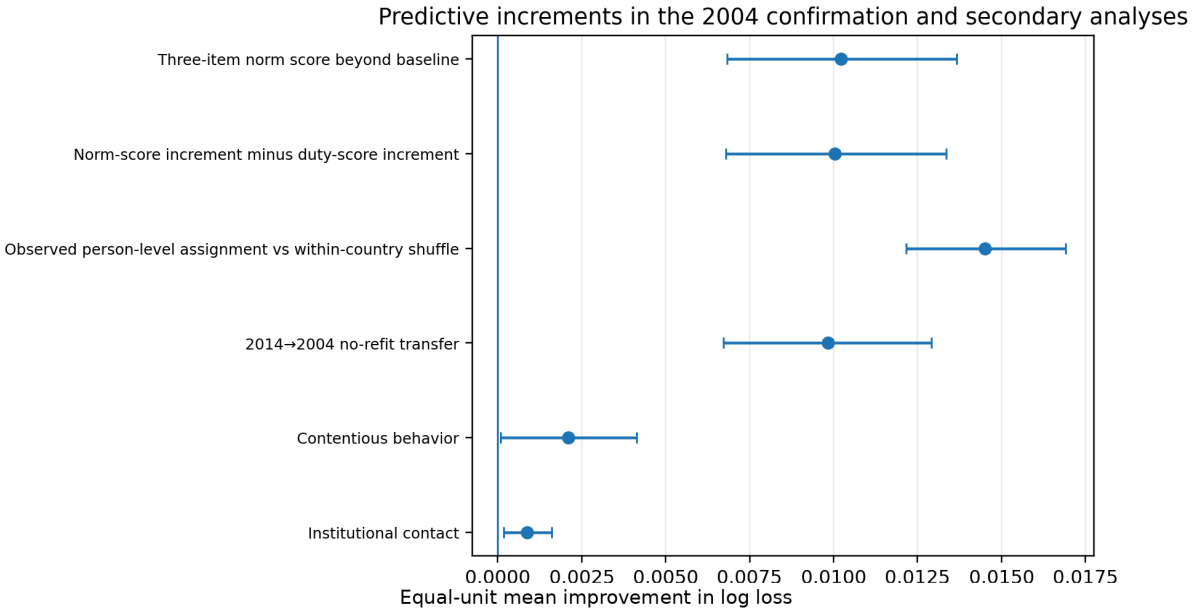

Figure 2. Equal-unit predictive differences for the three prospectively specified 2004 confirmation tests, the no-refit 2014→2004 transfer, and the two secondary behavioral outcomes. Error bars show 95% equal-unit bootstrap intervals. Positive values favor the civic-norm score or, for the permutation test, the observed respondent-score correspondence.

3.6 Post-confirmation robustness analyses

Several post-confirmation analyses tested simple rival explanations without changing the prospectively specified 2004 classification. Adding political interest to the baseline reduced but did not eliminate the three-item score's increment: 0.00869 [0.00539, 0.01211] in 2004 and 0.00727 [0.00348, 0.01100] in 2014. Adding response efficacy as well produced increments of 0.00652 [0.00374, 0.00925] and 0.00582 [0.00339, 0.00848], respectively. A further augmented post-confirmation baseline that incorporated prior contentious behavior and institutional contact retained positive increments of 0.00571 [0.00314, 0.00821] in 2004 and 0.00492 [0.00287, 0.00719] in 2014. These results show that the measured signal was not fully absorbed by the included indicators of political interest, perceived governmental responsiveness, or prior measured activism/contact; unmeasured proxy explanations remain possible.

The 2004 result was also insensitive to use of the archive-provided survey weights in a post-confirmation analysis. After normalizing positive weights within country, assigning unit weight to missing normalized values, using those weights in model fitting, and using them again in held-out log-loss calculation, the equal-unit mean improvement from adding the three-item score was 0.01003 (95% equal-unit bootstrap interval [0.00653, 0.01356]), positive in 29 of 38 units. Training-fold median imputation and feature standardization remained unweighted. This estimate is very close to the unweighted confirmation estimate of 0.01023. Appendix E documents the variable lineage and complete country-level weighted results.

Metric robustness was similar. In 2004, the three-item score improved log loss by 0.01023, Brier score by 0.00459, and AUC by 0.01687; the corresponding 95% intervals were all positive. The same qualitative pattern held in 2014, including an average AUC increment of 0.01958. The result is therefore not solely a log-loss calibration artifact.

3.7 Exploratory reverse transfer and limits of domain invariance

After the confirmation result was known, the transfer direction was reversed as an exploratory boundary test. A model estimated in 2004 and applied to 2014 without refitting showed a positive mean increment from the three-item score across all 34 2014 countries (+0.00826, 95% bootstrap interval [+0.00286, +0.01287]). However, among the six 2014 countries absent from the normalized 2004 country set, the mean was −0.00567 and the 95% bootstrap interval spanned zero [−0.02675, +0.00844]. The evidence therefore does not support a claim that the same fixed predictor will improve performance in every wholly novel national domain.

4. Discussion

4.1 What the prospectively specified confirmation supports

The primary finding is not that citizenship norms correlate with political participation; that proposition is already established. Rather, the present analysis shows that a fixed three-item civic-norm score carried incremental predictive information into national contexts excluded from model fitting, and that the same score met all three prospectively specified criteria in the 2004 dataset after being identified exploratorily in 2014. A model estimated in 2014 also retained the norm increment when transferred to 2004 without target-wave refitting. Together, these results are consistent with positive average predictive transportability within the tested ISSP Citizenship domains.

The specific duty-score comparison and the within-country permutation test sharpen that interpretation. The three-item score added more predictive information than the particular duty comparator based on tax compliance and law obedience. The permutation test preserved each held-out country's score distribution while destroying which respondent had which score and its joint relation with baseline variables; the resulting performance loss shows that country-level differences in the score's marginal distribution alone cannot account for the predictive advantage. The test does not establish a causal respondent-level mechanism.

4.2 Relation to prior research and novelty boundary

Dalton's distinction between duty-based and engaged citizenship and subsequent cross-national studies provide the substantive background for this analysis. Bolzendahl and Coffé (2013) used ISSP 2004 to relate specific "good citizen" norms to voting, party membership, and activism across 25 nations. Cenker-Özek et al. (2021) used ISSP 2004 and 2014 to relate active, cosmopolitan, dutiful, and rights-based norms to protest across 22 democracies. The present study therefore does not claim novelty for an association between engaged citizenship norms and participation.

Nor is country-holdout prediction itself a new method in cross-cultural research. Recent work in a different social-norm domain has used predictions for societies excluded from estimation to evaluate cross-cultural generalization (Eriksson et al., 2026). The candidate contribution here is narrower: applying country-holdout incremental prediction, within-country permutation, and no-refit cross-wave transfer to a fixed individual-level citizenship-norm signal. A targeted literature search did not identify a directly matching ISSP citizenship-norm analysis that combines these tests, but the search was not exhaustive and no priority claim is made.

This predictive framing complements explanatory modeling rather than replacing it. As Yarkoni and Westfall (2017) emphasized, theoretically interpretable associations can have uncertain predictive value outside the observations used to estimate them. In cross-national survey data, country-level holdout makes that issue explicit because the national context itself changes between estimation and evaluation.

4.3 A bounded interpretation of reported counter-action likelihood versus reported behavior

The outcome-specificity analyses caution against converting the result into a general theory of participation. The three-item score had a larger predictive increment for the primary survey response—the reported likelihood of attempting counter-action against a law regarded as unjust or harmful—than for reported past-year contentious behavior or institutional contact. The survey item does not separately identify willingness, perceived efficacy, opportunity, behavioral intention, or action capacity, so the mechanism behind this difference cannot be inferred from the present data. The bounded empirical statement is only that the score predicted the primary survey response more strongly than the two reported behavioral outcomes under the tested models.

4.4 Why the result is bounded rather than universal

The result is positive on average but heterogeneous. Nine of 38 2004 country/sample units did not show a positive incremental prediction from the three-item score. More importantly, the exploratory reverse transfer from 2004 to 2014 was positive across all 2014 countries on average but was not positive in the subset of countries absent from the 2004 training domain. The evidence therefore favors a reproducible average signal across the tested domains, not a country-invariant law. This distinction is consistent with explicit constraints on generality (Simons et al., 2017): national context is part of the target domain rather than incidental noise.

4.5 Limitations

- The study is observational and cross-sectional within each wave. It cannot establish that civic norms cause reported counter-action likelihood or political participation, nor that norms are psychologically prior to efficacy, trust, identity, or political socialization.
- The three-item score shows acceptable internal consistency, but formal configural, metric, and scalar measurement invariance across countries and waves was not established. Cross-national response styles, translation differences, or survey-mode differences could contribute to the observed transport pattern.
- All evidence comes from two waves of the same ISSP module family. Replication with a non-ISSP instrument and different item wording is necessary before claiming broader predictive portability of the score.
- Unmeasured stable traits, civic socialization, political identity, and social-desirability responding may account for part of the incremental signal. The post-confirmation rival-variable analyses reduce several simple proxy explanations but cannot eliminate unmeasured confounding.
- The primary confirmation dichotomizes a four-category counter-action likelihood item. An ordinal analysis would retain more response information and should be evaluated in future work without redefining the present binary-prediction result.
- Transport performance is heterogeneous. The exploratory reverse-transfer result for countries absent from the 2004 training domain directly argues against universal domain invariance.
- The within-country permutation test does not isolate whether the predictive information arises from the score values themselves, the pattern of imputed score values induced by missingness, or the respondent-level joint structure of the score with measured and unmeasured covariates. It therefore supports a respondent-linked predictive-information claim, not an identified individual mechanism.

4.6 Implications and next tests

The next tests should increase the domain shift rather than enlarge the hypothesis. Priorities include formal multi-group or item-response analyses of measurement comparability; replication in an independent survey instrument; prospectively specified prediction in other civic-attitude datasets or waves; explicit modeling of country-level response styles; and pre-specified analysis of national contexts in which transferred prediction fails. Failure under those tests should narrow or reject the hypothesis rather than trigger post hoc expansion.

5. Conclusion

Across ISSP Citizenship 2004 and 2014, a fixed three-item civic-norm score based on understanding opposing viewpoints and helping worse-off people domestically and abroad provided incremental predictive information for respondents' reported likelihood of attempting counter-action against a law regarded as unjust or harmful beyond a broad demographic and political-attitudinal baseline. All three prospectively specified 2004 criteria were met; the observed respondent-score correspondence outperformed within-country score permutations; and the predictive increment remained positive in a 2014→2004 transfer without target-wave coefficient refitting. The increment was substantially smaller for two reported political-behavior outcomes and was not positive in every wholly novel national domain. The evidence is therefore consistent with a bounded predictive-transportability claim for this fixed score across many of the tested country and wave shifts. It does not establish a causal mechanism, measurement invariance, universal cross-cultural portability, or a general theory of participation.

Declarations

Data availability

The ISSP respondent-level microdata are not redistributed with this article or its public-facing reproducibility materials. Users must obtain the official files directly from GESIS under the applicable archive access terms: ISSP Citizenship 2004, ZA3950 v1.3.0, https://doi.org/10.4232/1.11372; and ISSP Citizenship II 2014, ZA6670 v2.0.0, https://doi.org/10.4232/1.12590. Appendix F records the expected input filenames, row counts, and SHA-256 values used to identify the exact source files for reproduction.

Code and reproducibility

This single-file article includes integrated technical appendices documenting the prospectively specified 2004 plan, wave-specific variable mapping, score construction, missing-data handling, survey-weight sensitivity, numerical verification, input-file hashes, environment information, and reproducible analysis logic. Respondent-level ISSP source microdata are not redistributed. No separate public code repository accompanies this version. The integrated appendices provide the technical record needed to understand and reconstruct the analyses from legally obtained GESIS source files; no external supplement is required to understand the analyses reported here.

Ethics statement

This study is a secondary analysis of de-identified survey data distributed by GESIS. No new participant recruitment, intervention, or collection of identifiable human-subject data was conducted for this project.

Funding

No external funding was used for this analysis.

Competing interests

The author declares no competing interests.

AI assistance disclosure

Generative AI systems were used throughout the research workflow for theory stress-testing, code generation and review, reproducibility auditing, literature-search assistance, manuscript drafting, and document preparation. AI-generated suggestions were not treated as empirical evidence. Numerical results were produced from versioned analysis code and checked against archived outputs. The author is responsible for all final scientific claims and interpretations.

References

Bolzendahl, C., & Coffé, H. (2013). Are “Good” Citizens “Good” Participants? Testing Citizenship Norms and Political Participation across 25 Nations. Political Studies, 61(S1), 45–65. https://doi.org/10.1111/1467-9248.12010

Cenker-Özek, C. I., Çakmaklı, D., & Karakoç, E. (2021). Rights and responsibilities: Citizenship norms and protest activity in a cross-country analysis. Social Science Quarterly, 102(4), 1394–1407. https://doi.org/10.1111/ssqu.13041

Dalton, R. J. (2008). Citizenship Norms and the Expansion of Political Participation. Political Studies, 56(1), 76–98. https://doi.org/10.1111/j.1467-9248.2007.00718.x

Eriksson, K., Strimling, P., Vartanova, I., & Simpson, B. (2026). Same flavours, different taste buds: a theory for predicting social norms for specific behaviours across cultures. Journal of the Royal Society Interface, 23(237), 20251122. https://doi.org/10.1098/rsif.2025.1122

ISSP Research Group. (2012). International Social Survey Programme: Citizenship — ISSP 2004. GESIS Data Archive, Cologne. ZA3950 Data file Version 1.3.0. https://doi.org/10.4232/1.11372

ISSP Research Group. (2016). International Social Survey Programme: Citizenship II — ISSP 2014. GESIS Data Archive, Cologne. ZA6670 Data file Version 2.0.0. https://doi.org/10.4232/1.12590

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. Perspectives on Psychological Science, 12(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Appendix A. Analysis provenance

The three-item score was discovered during decomposition of an earlier exploratory 2014 model that failed its own pre-specified out-of-country predictive criteria. The failed model is not part of the present hypothesis and its failure was not reclassified as success. The civic-norm observation was instead treated as a new exploratory finding, after which the 2004 confirmation variables, comparisons, prediction procedure, and decision criteria were finalized before inspection of the 2004 individual-level confirmation results.

The prospectively specified 2004 classification depended only on the three required tests described in Section 2.6. Subsequent analyses—behavioral outcomes, item decomposition, rival-variable baselines, alternative predictive metrics, survey-weight sensitivity, and reverse cross-wave transfer—are reported as secondary, robustness, or boundary-setting analyses and cannot retroactively change the prospectively specified classification.

Appendix B. Constraints on generality

The directly supported target domain is the set of national survey contexts represented by the 2004 and 2014 ISSP Citizenship modules under the mapped variables and coding used here. The evidence supports positive average predictive transportability of the fixed score across the tested country/sample units and one strong 2014→2004 no-refit transfer. It does not imply that every country benefits from the predictor, that the score is measurement-invariant across cultures, or that the signal will survive different survey instruments, languages, modes, political eras, or outcome definitions. The negative exploratory reverse-transfer result in 2014 countries absent from the 2004 training domain is direct evidence against universal domain invariance.

Appendix C. Cross-wave variable mapping and analysis construction

Table C1 makes the 2004↔2014 harmonization explicit. "Harmonized field" refers to the analysis representation after wave-specific source variables were mapped to a common construct. Valid ordinal items were rescaled to 0–1 in the direction shown; invalid archive codes were treated as missing before score construction. The complete analysis code uses these same mappings.

Construct / role	ISSP 2014 source	ISSP 2004 source	Harmonized field	Coding / rule
Civic norm: understand opposing views	V10	V9	V10_01	1–7 → (x−1)/6; higher = stronger norm
Civic norm: help worse-off people domestically	V12	V11	V12_01	1–7 → (x−1)/6; higher = stronger norm
Civic norm: help worse-off people abroad	V13	V12	V13_01	1–7 → (x−1)/6; higher = stronger norm
Six-item engaged-citizenship score	V8–V13	V7–V12	six_item_engaged_score (code alias G)	Mean of six mapped items; ≥4 valid required
Two-item duty comparator (no voting)	V6, V7	V5, V6	two_item_duty_score	Tax compliance + law obedience; both valid required
Primary outcome	V45	V40	binary_counter_action_outcome (code alias voice_likely)	Valid 1–4; 1–2=1, 3–4=0
Response efficacy	V46	V41	response_efficacy	Valid 1–4 → 0–1; higher = greater expected response
External political efficacy	V41, V42	V36, V37	external_efficacy (E)	Two items mapped to 0–1; both valid required
Internal political efficacy	V43, V44	V38, V39	internal_efficacy (I)	Two items oriented high=greater efficacy; both valid required
Political trust	V49	V43	political_trust	Valid 1–5 → 0–1; higher = trust
Social trust	V52	V46	social_trust	Valid 1–4 → 0–1; higher = trust
Institutional evaluation: politician self-interest	V50	V44	B50	Valid 1–5 → 0–1; higher = more negative evaluation
Institutional evaluation: election honesty	V58	V55	B58	Valid 1–5 → 0–1; higher = more negative evaluation
Institutional evaluation: election fairness	V59	V56	B59	Valid 1–5 → 0–1; higher = more negative evaluation
Institutional evaluation: public-service commitment	V60	V57	B60	Valid 1–4 → 0–1; higher = more negative evaluation
Institutional evaluation: public-service corruption	V61	V59	B61	Valid 1–5 → 0–1; higher = more negative evaluation
Institutional-evaluation composite	above five	above five	institutional_evaluation (B)	Mean; ≥3 of 5 valid required
Political interest	V47	V42	political_interest	Wave-mapped robustness predictor
Boycott (past year)	V18	V18	V18_pastyear	Archive response 1=1; other valid 2–4=0
Demonstration (past year)	V19	V19	V19_pastyear	Archive response 1=1; other valid 2–4=0

Construct / role	ISSP 2014 source	ISSP 2004 source	Harmonized field	Coding / rule
Contentious-behavior composite	V18 or V19	V18 or V19	contentious_voice	OR/max of the two valid binary indicators
Contact politician (past year)	V21	V21	institutional_contact	Archive response 1=1; other valid 2–4=0
Age	AGE	v201	age	Valid 15–110
Sex	SEX	v200	sex	Valid archive categories 1 or 2
Education degree	DEGREE	v205	degree	Valid 0–6
Survey weight	WEIGHT	v381	weight / weight_norm	Finite positive values; normalized within country for sensitivity
Country/sample identifier	C_ALPHAN	C_ALPHAN	country	Archive ISO/sample prefix

Appendix C.1 Score missingness, imputation, interactions, and scaling

The three-item score was computed only when at least two component items were valid; the six-item engaged score required at least four valid items; and the two-item duty comparator required both items. The five-item institutional-evaluation composite required at least three valid components. A score or predictor that remained missing then entered fold-specific median imputation. Medians were estimated exclusively from the training countries in each leave-one-country-out fold. Held-out-country observations did not affect imputation parameters. B×I, B×E, I×E, and B×I×E were constructed after imputation from the unstandardized imputed component variables, after which the full feature set was standardized using training-fold means and standard deviations. This order prevents algebraically inconsistent interaction values after missing-data handling.

Appendix C.2 Secondary-outcome construction

For both waves, the political-action items use four valid timing/propensity categories. “Have done it in the past year” was coded 1 and the other valid categories were coded 0. The contentious-behavior outcome is an OR rule: it equals 1 if either the boycott item (V18) or demonstration item (V19) indicates past-year action; if both component items are missing the composite is missing. Institutional contact uses the politician-contact item (V21) with the same past-year binary rule. In 2004, valid N=51,748 for the contentious-behavior composite and N=51,040 for institutional contact; 38 held-out units satisfied the minimum of 100 cases with at least 20 positive and 20 negative outcomes.

Appendix D. Numerical verification of manuscript values

Major reported values were checked against machine-readable outputs. The four primary 2004 confirmation quantities were independently recomputed from the raw ZA3950 v1.3.0 file. The 2014→2004 no-refit transfer was independently recomputed from raw ZA6670 v2.0.0 and ZA3950 v1.3.0. The refitted 2014 frozen-model parameter values matched the archived parameter file to floating-point precision; in an independent reproduction, the maximum absolute difference across the 127 numeric entries was below 10−14 (approximately 8.85×10−15). The weighted sensitivity was independently rerun from raw ZA3950 using semantic aliases for the three norm items, and the resulting 38-row country/sample-unit file was byte-identical to the archived sensitivity output. Other archived secondary/boundary results were independently machine-read and re-summarized for consistency. Differences below reflect manuscript rounding.

ID	Reported	Recomputed	Abs. difference	Status	Verification basis
C01 - Primary norm increment	0.01023	0.0102269541	3.05e-06	PASS	end-to-end raw-data rerun
C02 - Duty contrast	0.01003	0.0100316042	1.6e-06	PASS	end-to-end raw-data rerun
C03 - Within-country permutation	0.0145	0.0145013802	1.38e-06	PASS	end-to-end raw-data rerun
C04 - Six-item supportive result	0.01711	0.0171077322	2.27e-06	PASS	end-to-end raw-data rerun
C05 - 2014→2004 no-refit transfer	0.00983	0.00983057816	5.78e-07	PASS	end-to-end raw-data rerun; frozen model parameters reproduced to floating-point precision
C06 - 2004 units absent from 2014 training domain	0.01052	0.0105216285	1.63e-06	PASS	subset recomputation from raw transfer rerun
C07 - Pooled Cronbach alpha	0.708	0.708126533	0.000127	PASS	direct recomputation from complete three-item cases
C08 - Contentious behavior	0.00209	0.00209142147	1.42e-06	PASS	independent machine-read summary check
C09 - Institutional contact	0.000878	0.000877997037	2.96e-09	PASS	independent machine-read summary check
C10 - Weighted sensitivity	0.01003	0.0100295475	4.52e-07	PASS	end-to-end raw-data rerun using semantic norm-item aliases
C11 - Reverse transfer, all 2014 units	0.00826	0.00826277885	2.78e-06	PASS	independent machine-read summary check

C12 - Reverse transfer, novel-domain subset	-0.00567	-0.00567157898	1.58e-06	PASS	independent machine-read summary check

Verification note: bootstrap intervals and positive-unit counts were checked against the corresponding machine-readable records. The published decimal values are rounded summaries; the classification does not depend on any discrepancy shown above.

Appendix E. Survey-weight sensitivity and variable-lineage audit

A pre-publication audit examined an apparent discrepancy between the 2004 raw variable numbering reported in the manuscript (V9, V11, V12) and the harmonized columns used in the weighted sensitivity code (V10_01, V12_01, V13_01). The discrepancy was nomenclatural, not substantive. In the canonicalization step, raw 2004 V9 is mapped to the common 2014-name V10, raw V11 to V12, and raw V12 to V13; 0–1 recoding then yields V10_01, V12_01, and V13_01. The weighted analysis therefore used the intended three items. An independent semantic-name rerun produced a country-level CSV identical to the archived result; no numerical correction was required.

Semantic analysis name	Harmonized column	ZA3950 raw variable	Archive item label	Recoding
norm_understand_others	V10_01	V9	Good citizen: understand other opinions	(x−1)/6; high=stronger norm
norm_help_domestic	V12_01	V11	Good citizen: help less privileged people in country	(x−1)/6; high=stronger norm
norm_help_global	V13_01	V12	Good citizen: help less privileged people in world	(x−1)/6; high=stronger norm

Weighting procedure: valid positive ZA3950 weight v381 was divided by the mean valid positive weight within the respondent's country/sample unit. The normalized weight was supplied as sample_weight during model fitting and as sample_weight in held-out log-loss calculation. Missing normalized weights were assigned 1.0. Median imputation and StandardScaler estimation were unweighted. This analysis was post-confirmation sensitivity evidence and did not alter the prospectively specified 2004 classification.

Held-out unit	N	Weighted baseline log loss	Weighted norm-model log loss	Improvement
AT	931	0.654840	0.662185	-0.007345
AU	1799	0.674263	0.657794	0.016469
BE-FLA	1220	0.660811	0.665385	-0.004574
BG	1017	0.532137	0.533307	-0.001170
BR	1843	0.704923	0.688771	0.016152
CA	1143	0.695128	0.683154	0.011974
CH	1037	0.651508	0.644979	0.006530
CL	1401	0.664327	0.652821	0.011506
CY	995	0.545779	0.543020	0.002759
CZ	1247	0.555455	0.556300	-0.000846
DE-E	370	0.577684	0.569386	0.008298
DE-W	786	0.592463	0.585214	0.007249
DK	1117	0.613249	0.600378	0.012870
ES	2286	0.635435	0.644140	-0.008704
FI	1249	0.538872	0.506582	0.032290
FR	1186	0.714072	0.718809	-0.004737
GB-GBN	801	0.670237	0.662794	0.007443
HU	974	0.467989	0.429246	0.038743
IE	1029	0.629616	0.618629	0.010988
IL (J+A)	1136	0.623878	0.616891	0.006987
JP	1145	0.563373	0.535070	0.028303
KR	1277	0.692609	0.697402	-0.004793
LV	954	0.551686	0.536888	0.014798
MX	1176	0.723505	0.699568	0.023937
NL	1708	0.599265	0.586770	0.012495
NO	1306	0.627269	0.617050	0.010219
NZ	1305	0.720455	0.723007	-0.002552
PH	1084	0.840601	0.833009	0.007592
PL	1211	0.550909	0.551661	-0.000752
PT	1500	0.634033	0.628020	0.006013
RU	1671	0.473144	0.453654	0.019490
SE	1214	0.555050	0.538410	0.016640
SI	1014	0.594679	0.587111	0.007568
SK	1021	0.580833	0.580814	0.000019
TW	1740	0.563238	0.549804	0.013434
US	1451	0.724701	0.705837	0.018863
UY	1081	0.693536	0.671669	0.021867
VE	1155	0.745038	0.719937	0.025100

Equal-unit summary of the weighted sensitivity: mean improvement = 0.01002955; 95% equal-unit bootstrap interval [0.00653388, 0.01355721]; positive improvement in 29/38 units. The unweighted prospectively specified mean was 0.01022695.

Appendix E.1 Portable weighted-sensitivity implementation

The listing below is the portable equivalent of the weighted sensitivity used for the reported result. It uses semantic aliases for the three 2004 norm items and relative/project-supplied paths rather than machine-specific absolute paths. The full main-analysis logic is described in Sections 2.4–2.8 and Appendix C.

```
from pathlib import Path
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import log_loss

prepared = Path("data/prepared_2004.csv")
out_csv = Path("results/weighted_sensitivity_2004.csv")
d = pd.read_csv(prepared)

# These harmonized columns originate from ZA3950 V9, V11, and V12.
d["norm_understand_others"] = d["V10_01"]
d["norm_help_domestic"] = d["V12_01"]
d["norm_help_global"] = d["V13_01"]
items = ["norm_understand_others", "norm_help_domestic", "norm_help_global"]
d["norm3"] = d[items].mean(axis=1, skipna=True).where(d[items].notna().sum(axis=1) >= 2)
d = d[d["voice_likely"].notna() & d["country"].notna()].copy()
base = ["age", "sex", "degree", "I", "E", "B", "political_trust", "social_trust"]

def predict_fold(train, test, use_norm):
    raw = base + (["norm3"] if use_norm else [])
    imp = SimpleImputer(strategy="median")
    a = pd.DataFrame(imp.fit_transform(train[raw]), columns=raw, index=train.index)
    b = pd.DataFrame(imp.transform(test[raw]), columns=raw, index=test.index)
    for z in (a, b):
        z["B_I"] = z.B * z.I; z["B_E"] = z.B * z.E
        z["I_E"] = z.I * z.E; z["B_I_E"] = z.B * z.I * z.E
    cols = base + ["B_I", "B_E", "I_E", "B_I_E"] + (["norm3"] if use_norm else [])
    sc = StandardScaler()
    X = sc.fit_transform(a[cols]); Y = sc.transform(b[cols])
    model = LogisticRegression(max_iter=2000, C=1.0, solver="newton-cholesky")
    model.fit(X, train.voice_likely.astype(int), sample_weight=train.weight_norm.fillna(1.0))
    return model.predict_proba(Y)[:, 1]

rows = []
for country in sorted(d.country.unique()):
    train = d[d.country != country]; test = d[d.country == country]
    y = test.voice_likely.astype(int).to_numpy()
    if len(test) < 100 or y.sum() < 20 or len(y) - y.sum() < 20:
        continue
    w = test.weight_norm.fillna(1.0).to_numpy()
    p0 = predict_fold(train, test, False); p1 = predict_fold(train, test, True)
    loss0 = log_loss(y, p0, labels=[0, 1], sample_weight=w)
    loss1 = log_loss(y, p1, labels=[0, 1], sample_weight=w)
    rows.append((country, len(test), loss0, loss1, loss0 - loss1))

pd.DataFrame(rows, columns=["country", "n", "baseline_loss", "norm_loss", "improvement"]).to_csv(out_csv, index=False)
```

Appendix F. Reproduction environment, inputs, and workflow

The following information is sufficient to reconstruct the data-preparation and model-evaluation environment from legally obtained GESIS files. Source respondent-level microdata are not embedded in this article.

Dataset	Expected input filename	Rows	SHA-256	DOI
ISSP Citizenship 2004	ZA3950_v1-3-0.dta	52,550	ba1511d671bfb5d8e9952f084fc3e980eb5f466b2f7b928ff1722747b3c92f5f	https://doi.org/10.4232/1.11372
ISSP Citizenship II 2014	ZA6670_v2-0-0.dta	49,807	34aa98148b2f41e4a7265c57df4d4a0d542812b8198a8561c1bef71b40b8d07b	https://doi.org/10.4232/1.12590

Software environment used for final verification:

Software	Version used for final verification
Python	3.13.5
numpy	2.3.5
pandas	2.2.3

Software	Version used for final verification
scikit-learn	1.8.0
scipy	1.17.0

Exact archived implementation settings: scikit-learn LogisticRegression solver="newton-cholesky", C=1.0, max_iter=2000. The NumPy RNG seed used for the archived 2004 confirmation bootstrap/permutation workflow was 20260823. The confirmation used 100 within-country permutations per eligible held-out unit and 10,000 equal-unit bootstrap resamples.

Reproduction sequence:

1. Obtain ZA3950 v1.3.0 and ZA6670 v2.0.0 directly from GESIS under the archive's access terms and verify the input hashes shown above.
2. Prepare each wave using the mappings and coding rules in Appendix C. For 2004, map the legacy control variables v200/v201/v205/v381 to sex/age/degree/weight before preparation.
3. Recreate the 2014 exploratory/source-model features; compute the fixed three-item score using the mapped item content; and fit the 2014 source models using L2 logistic regression with solver="newton-cholesky", C=1.0, and max_iter=2000.
4. Freeze the 2014 training medians, feature order, scaling parameters, intercepts, and coefficients before applying the no-refit model to 2004.
5. Run the 2004 leave-one-country-out confirmation. Retain units with at least 100 valid cases and at least 20 positive and 20 negative outcomes. Recompute all interaction terms after training-fold median imputation and before standardization.
6. Run 100 within-country permutations of the three-item score for each eligible held-out unit using the already fitted norm-augmented model; preserve the country's score distribution while breaking the original respondent-score correspondence.
7. Summarize country/sample-unit effects with equal unit weight and 10,000 bootstrap resamples of the unit effects.
8. Run the reported secondary, robustness, weighted-sensitivity, and reverse-transfer analyses without allowing them to retroactively change the prospectively specified 2004 classification.
9. Compare recomputed major manuscript values with Appendix D and verify final document/package hashes.

Appendix F.1 Audit trail for pre-result execution corrections

Before any 2004 individual-level outcome/model result was inspected, two execution issues were corrected: (1) ZA3950 legacy control names v200/v201/v205/v381 were mapped to sex/age/degree/weight, and (2) the bootstrap routine was changed so that each confirmation criterion used the same single bootstrap interval for both reporting and classification. Neither correction changed the civic-norm items, outcome coding, model family, leave-one-country-out design, regularization, eligibility criteria, bootstrap threshold, or decision rule. The later pre-publication weighted-variable-lineage audit found no numerical error and required documentation/semantic naming only.

Appendix F.2 Interpretation of bootstrap intervals

All intervals labeled "equal-unit bootstrap" are nonparametric intervals obtained by resampling the held-out country/sample-unit effect estimates with equal probability. Because the leave-one-country-out models share overlapping training data, the unit effects are not independent in the sense assumed by a design-based sample of countries. The intervals are used as descriptive uncertainty summaries for the tested set of evaluation units and as the prospectively specified decision device; they are not population-sampling confidence intervals for a hypothetical superpopulation of countries.