What the Balkan Barometer Can and Cannot Rank: A Measurement Audit of the Western Balkans' Flagship Survey

Miloš Kankaraš

Center for Psychology, Cetinjski put 36, 81100 Podgorica, Montenegro

ORCID: 0000-0002-3190-7751

Correspondence: milos@centerforpsychology.me

Working paper, version 2, September 2026. Analysis package: Zenodo, https://doi.org/10.5281/zenodo.22919169.

Version 2 (29 September 2026) corrects the reference apparatus after a full citation and claim-support check. Two sentences claimed more than their sources show and were narrowed: the fit-index caution now attributed to Sass et al. (2014) is the one their paper makes, and the interviewer-effects sentence no longer attaches a magnitude to West and Blom (2017). Nineteen missing DOIs were added, three citations were corrected to the APA form for three or more authors, a citation to the World Development Indicators was added where the text used them, and one issue number was completed. No result, table, figure or conclusion changed.

Abstract

The Balkan Barometer, annual since 2015 with roughly 1,000 face-to-face interviews in each of six Western Balkan economies, is cited across the region as what its citizens think. We evaluate it as a measuring instrument from the Regional Cooperation Council's own released microdata for 2023 to 2025, within a total survey error frame. The published percentages reproduce from the microdata (440 of 444 cells within one point). Of the fifteen pairwise economy differences on each of thirty 2025 headline items, 71 percent are resolvable at a design effect of 2.0, but only 40 percent of adjacent-rank pairs are: six points is the least difference at which two economies of 1,000 can be told apart. The WB6 figure is an unweighted mean of six economies whose populations differ tenfold; under population weights it moves by 6 to 12 points on the most-quoted items, including EU membership (64 versus 58 percent in 2025). Of twenty-three headline series present in all three editions, four are comparable throughout; between 2024 and 2025 the changed series moved by 14 points on average against 6 for the stable ones. The 2025 institutional battery is scalar invariant once two cells are freed, police in Kosovo* and courts in North Macedonia; with them freed, Kosovo*'s raw-score lead disappears and the top three economies cannot be separated. The two published editions score 16 of 32 on a documentation checklist built from what the ESS, the Eurobarometer

and the AAPOR Transparency Initiative publish routinely; precision, comparability and limitations score zero. We report two database errors and a file anomaly for correction, and close with the four things a technical report would need to contain.

Keywords: cross-national surveys; Western Balkans; rank uncertainty; measurement invariance; survey transparency; Balkan Barometer

---

1. Introduction

Each year, most recently on 11 December 2025 in Brussels (RCC, 2025c), the Regional Cooperation Council (RCC) presents the Balkan Barometer, and the same sentences then travel through the region's ministries, delegations and newsrooms: support for EU membership has risen to a record, citizens trust religious organisations more than parliaments, rising prices are the region's first concern (RCC, 2025a). The Barometer is the only current survey fielded on one questionnaire in all six Western Balkan economies in the same weeks whose microdata are public. The Standard Eurobarometer's candidate-country waves cover five of the six on the EU's questionnaire, and the Gallup Balkan Monitor covered all six between 2006 and 2010. It has been run annually since 2015, and its 2025 round interviewed 6,012 citizens and 1,216 companies (RCC, 2025b). This position is why its numbers are treated as the region's opinion rather than as one survey's estimate of it.

An instrument that is used this way can be asked the ordinary questions of survey measurement. In the total survey error frame (Groves & Lyberg, 2010) each of the Barometer's public claims rests on a different error source. A claim about an economy's level or about the ordering of economies rests on sampling error and the design effect. The regional figure rests on a stated estimand and on weighting. A claim about change rests on the instrument having been held constant. A comparison of economies on a multi-item scale rests on measurement equivalence across languages and contexts. And every one of them rests, for the reader, on documentation, because the reader cannot see the fieldwork. These are the five questions we ask, in that order. We do not ask whether the region's citizens are right about anything. We follow the RCC's designation convention for Kosovo* throughout (see Declarations).

The questions are askable because the RCC has made them so. The public presentation of the Barometer is a set of reports, an interactive database and infographics. A reader of the reports finds no margin of error, no weighting statement, no questionnaire and no methodological change log. But the results page of the interactive database carries a link, labelled "Balkan Barometer Micro Data", to the respondent-level SPSS files for every edition from 2017 to 2025. They are the basis of our audit, and everything we report can be reproduced from them by anyone.

The paper belongs to our series that audits published international rankings as measuring instruments (Kankaraš, 2026). It borrows the series' rank-inference logic from Goldstein and Spiegelhalter (1996): a published ordering is a set of estimates, each with a sampling distribution, and the question is which orderings the reader draws from it are supported. Two things are new here in method. Because the Barometer is a probability survey, its precision is a matter of arithmetic rather than assumption, and the paper shows how to proxy the unpublished design effect from the geography a released file does carry. Because its microdata are public, the cross-language

equivalence that composite indices can only be presumed to have can here be tested, and probed against context using a nationality variable that one edition records.

The constructive frame comes first. Section 2 describes the data and methods. Section 3 reports what the Barometer can rank: its published figures reproduce, its large contrasts are real, and its within-edition tests are interpretable. Section 4 reports what it cannot: adjacent positions, the regional figure as an estimate of regional opinion without a stated estimand, year-on-year change on items that changed, and economy comparisons on batteries where two items work differently, together with the errors the audit found in the database and the files. Section 5 sets out the four things a technical report would need to contain, scored against three reference programmes. Section 6 states the limitations and concludes.

2. Data and method

2.1 Sources

Three respondent-level files were downloaded from rcc.int on 18 September 2026: the 2023 edition (6,027 respondents, 282 variables, with a weight variable), the 2024 edition (6,000 respondents, 220 variables, no weight) and the 2025 edition (6,012 respondents, 172 variables, no weight). None carries a primary sampling unit, interviewer or language-of-interview identifier. The interactive database was read in full: 63 public-opinion indicators presented in 95 tables covering 2015 to 2025, each table giving per-economy percentages, the per-economy n and, in a hidden column, each economy's share of the regional figure. The 2023 and 2024 reports were read for their methodology notes (RCC, 2023, pp. 102-105; RCC, 2024, pp. 111-116). The 2025 report was listed as "coming soon" on the publications page on 18 September 2026, so its methodology is taken from the site's About page and its key-findings page (RCC, 2025a, 2025b). Population figures are the World Bank's 2024 mid-year estimates (World Bank, 2026).

The three editions were fielded by three contractors: an Indago-led consortium in March and April 2023, the Data Centrum Research Institute in March and April 2024, and CeSID with IDRA Research and Consulting in June 2025 (the About page gives 21 May to 30 June; the file's interview dates run from 3 to 30 June). The questionnaire had 77, 43 and 39 questions respectively. Each edition describes a stratified multi-stage cluster design with random-route household selection, a birthday rule for respondent selection and about ten interviews per sampling point. The 2024 note names the languages used in Kosovo* (Albanian and Serbian) and North Macedonia (Macedonian and Albanian) and reports a response rate of "22% on average" without definition. The 2023 and 2025 notes say only that the English questionnaire was "translated into related local languages". No edition states how many versions were fielded, whether Albania and Kosovo* received the same Albanian version, or whether Bosnia and Herzegovina, Montenegro and Serbia received one version or three. No edition reports a response rate except 2024. No file records the language of interview.

2.2 The headline crosswalk

We defined thirty-two headline indicators, each a percentage of respondents giving a stated answer, and mapped to the variable that carries it in each edition (Appendix A). Twenty-four are present in all three editions, and eight are 2025-only items used for the rank-resolution analysis. For each indicator the wording, response scale, base and option set were compared across editions from the SPSS labels and the database tables, and each adjacent pair of editions was graded as identical,

trivially reworded, changed in scale or base, changed in option set, or changed in construct; an indicator counts as comparable across a pair if it is identical or trivially reworded, and across all three editions if both pairs are (Appendix A). The published percentage for each indicator, edition and economy was read from the database table that the RCC itself uses for that wording, and compared with the microdata value on the base the RCC used: all respondents where don't-know is a coded category (2023, 2024, and some 2025 items), and valid responses where don't-know is stored as missing (most 2025 attitude items).

2.3 Precision and rank resolution

We computed the standard error of each economy estimate as the binomial standard error at the valid-response n, multiplied by the square root of a design effect. No design effect is published. Five values are used: 1.0, the simple-random-sampling floor, and 1.5, 2.0, 2.5 and 3.0. The 2023 weights imply a Kish weighting design effect of 1.00 to 1.02 by economy (Kish, 1992), so weighting contributes nothing. Clustering is the unknown, and the files record no sampling point.

Two proxies indicate where it lies. The 2025 file records the municipality (243 in all, a median of 24 respondents each) and the 2024 file the sub-national region (51). The intra-class correlation of the headline indicators across municipalities, with the region means removed because the design is stratified by region, has a median of .073 across items. Combined with the ten interviews per sampling point that the methodology notes describe, it implies a design effect of 1.7 (1.3 at five per point). The same correlation without removing the region means is .13, implying 2.2, and the region-level correlation in 2024 is .02, implying 1.2. Under nesting of sampling points within municipalities, the municipality correlation understates the sampling-point correlation, which puts the true design effect above these proxies; stratification below the region, if the contractors used it, removes between-municipality variance from the sampling variance and puts it below the raw proxy. The proxies are therefore indicative rather than bounds, and they sit between 1.2 and 2.2. We headline results at a design effect of 2.0 and report the full range.

For each indicator and edition the fifteen pairwise economy differences were tested at the 5 percent level under each design effect and counted, both unadjusted and with a Holm adjustment within the item. The five differences between adjacent ranks were counted separately. The unadjusted count is an upper bound on resolvable pairs. Rank intervals were then formed for each economy by simulation: 4,000 draws of the six estimates from their sampling distributions, ranked, with the 2.5th and 97.5th percentiles of each economy's rank taken as its interval, which is the construction Goldstein and Spiegelhalter (1996) describe. For two economies of 1,000 respondents at a proportion near one half, the least significant difference at the 5 percent level is 4.4 points under simple random sampling and 5.4, 6.2, 6.9 and 7.6 points at design effects of 1.5 to 3.0. The difference detectable with 80 percent power is 6.3, 7.7 and 8.9 points at 1.0, 1.5 and 2.0. Year-on-year and 2023-to-2025 movements per economy were tested the same way.

2.4 The regional figure

We recomputed each headline WB6 percentage as the population-weighted mean of the six economy percentages, with weights from the 2024 mid-year population (Albania 14.7, Bosnia and Herzegovina 19.6, Kosovo* 9.9, Montenegro 3.9, North Macedonia 11.3, Serbia 40.7 percent), and set it beside the published equal-weight figure. Weighting by the population aged 15 and over, the nearest the World Bank series (World Bank, 2026) comes to the survey's 18-and-over population,

changes no weight by more than 0.61 points (Serbia). Both figures take each economy's unweighted sample proportion as its estimate, which presumes each economy sample is self-weighting. Neither is a design-based estimate of a population parameter, and the exercise shows how far the two candidate estimands differ, not which is right.

2.5 Cross-language equivalence

We tested five multi-item batteries for measurement invariance across the six economies: the institutional battery in each edition, the 2025 trust-in-information battery (six items, ten-point scale) and the 2025 regional-cooperation battery (four items). The institutional batteries are seven items each but not the same seven. Judiciary, parliament, government, political parties and police are common to all three editions. The 2023 and 2024 batteries add local authorities and the military (in 2024 "military and law enforcement"), and 2025 adds the civil service and elections. Each edition is therefore a separate audit of that edition's battery, not a test of stability across editions, and a five-item version common to all three is reported alongside. Before any multi-group sequence is read, each economy's single-group fit is checked against one gate, CFI of at least .95 and SRMR of at most .08; an economy that fails it is reported as such and the sequence is rerun without it, because a pooled fit index can sit at ceiling while one group's items do not form the factor at all. Because the files record neither language of interview nor interviewer, economies are the primary groups. A comparison across economies is a comparison across whatever language versions were fielded, confounded with context. The 2024 note names Albanian and Serbian for Kosovo* and Macedonian and Albanian for North Macedonia, and no edition says whether Albania and Kosovo* used one Albanian version or whether Bosnia and Herzegovina, Montenegro and Serbia used one version or three. The 2023 file records self-reported nationality, which in the two economies that fielded two versions is the available proxy for the version received. We use it to define economy-by-nationality groups with at least 200 respondents, a conventional floor for a single-group factor model (Kline, 2016; the smallest cell is Albanians in North Macedonia, 204), and rerun the 2023 battery on them. "Nationality" here is the file's self-identification item, not citizenship.

Ordinal items were fitted by multiple-group confirmatory factor analysis with the WLSMV estimator, delta parameterisation and standardised factors, under the Wu and Estabrook (2016) identification, in the configural, threshold, threshold-plus-loading (metric) and threshold-plus-loading-plus-intercept (scalar) sequence, using lavaan and semTools (Rosseel, 2012; Jorgensen et al., 2022); the ten-point items were fitted with robust maximum likelihood. Don't-know responses are treated as missing and cases deleted listwise, which removes between 2 and 30 percent of an economy's respondents depending on the battery (most often 5 to 10 percent). A pairwise-available sensitivity is reported. All fit indices are the scaled versions, with the 90 percent interval for RMSEA, and every nested step carries a scaled chi-square difference test (Satorra, 2000, as implemented in lavaan). Because approximate fit indexes behave differently under WLSMV with ordered items, and changes in them warrant caution under that estimator (Sass et al., 2014), fit was judged on CFI, SRMR and the difference tests together, and invariance at each step on the change in CFI, reported against both the .01 tolerance of Chen (2007) and the .02 tolerance that Rutkowski and Svetina (2014) derive for many groups. Verdicts that differ between the two are marked. Where full scalar invariance failed in 2025, we located the offending item-by-economy cells from the modification indices and the alignment and fitted two partial scalar models: one that frees the thresholds of the two items in every economy (the five-anchor model) and one that frees them only in the two cells; the latent

means with standard errors from both are reported beside the full-scalar means. The 2025 institutional battery was further examined by alignment (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2018), using sirt (Robitzsch, 2024) on loadings and intercepts from single-group maximum-likelihood factor analyses that treat the four-point items as continuous, with bootstrap standard errors for the aligned means from 200 resamples within economy, under the package defaults and under a tuned setting (scale .2 and .4, power .25) that gave the same ordering. It was also checked with a two-factor configural model separating elected institutions from police and courts, with a sensitivity excluding straightliners, and with ordinal logistic regression DIF with economy as the grouping variable (lordif; Choi et al., 2011), which flags an item when the McFadden pseudo-R-squared change exceeds .02 after iterative anchor purification (seven iterations to convergence), with the .035 benchmark also reported. Response quality was summarised per economy and edition by the don't-know rate across single-response items, by straightlining on each battery (an identical answer to every item) and by the extreme-response rate (Krosnick, 1991; Kim et al., 2019; Baumgartner & Steenkamp, 2001).

2.6 Transparency

We scored sixteen documentation elements 0, 1 or 2 for each edition, one rater against a written rubric with quoted evidence, against what three programmes publish as a matter of course: the twelve disclosure elements of the AAPOR Transparency Initiative (AAPOR, 2026), the technical note annexed to every Standard Eurobarometer wave, which gives for each country the fieldwork institute, the fieldwork dates, the number of interviews and the population covered, and states the confidence intervals (European Commission, 2017, 2025), and the per-round documentation of the European Social Survey (ESS, 2026). Each score is backed by a quotation or a page reference (Appendix C). "Not found" means not found in the sources named, on the date named. The 2025 score is provisional because the 2025 report had not been published when the audit was run. The Cross-Cultural Survey Guidelines (Survey Research Center, 2016) and the translation documentation standard of Harkness et al. (2010) inform the translation element.

All computations are in the analysis package (Python 3.13.7 with pandas 2.3.3 and scipy 1.17.1; R 4.5.3 with lavaan 0.6-21, semTools 0.5-8, sirt 4.2-133 and lordif 0.4-2; seeds and estimator settings in its session-information file), and every number in this paper is listed in its results register. The package contains code and derived results only. The microdata are obtained from rcc.int.

3. What the Barometer can rank

3.1 The published figures reproduce

Of 444 economy-by-item cells compared on the unweighted microdata, 440 are within one point of the published value. The mean absolute difference is 0.35 points in 2023, 0.56 in 2024 and 0.30 in 2025. The four cells further than one point away are two Kosovo* cells in 2023 (relations better, 1.24; brings together, 1.05) and the exchanged pair of economy rows discussed in Section 4.5. The two Kosovo* cells are consistent with the 2023 report having used the file's weight: Kosovo* is the economy where the weight matters most (Kish design effect 1.019), and with the weight applied both cells, and every other 2023 cell, fall within a point (mean absolute difference 0.30). The 2024 and 2025 files carry no weight, and their published figures are unweighted counts. This is, in fact, the creditable finding of the audit. A reader who takes the published percentage as the survey's estimate

for that economy is not being misled about the number, and the RCC's release of the files is what allows the statement to be made.

One further item, the digital-worries question, reproduces only once its 2023 base is set to Internet users rather than all respondents, a base the report states in a figure note and the database does not.

3.2 Large contrasts are real

Two economies of 1,000 respondents each can be told apart, at the 5 percent level, when their percentages differ by 4.4 points under simple random sampling and by 6.2 points at a design effect of 2.0, for a proportion near one half. A single estimate carries a margin of 3.1 or 4.4 points. Most of the contrasts the Barometer reports are far larger than this, and the arithmetic confirms them. Across the 77 item-editions in the crosswalk, 79 percent of the 1,155 pairwise economy differences are resolvable under simple random sampling, 70 percent at a design effect of 2.0 and 64 percent at 3.0. With the Holm adjustment within items the figures are 72, 60 and 52 percent. For the thirty 2025 items the shares are 80, 71 and 66 percent unadjusted and 73, 61 and 54 percent adjusted (Table 1). Excluding the economic-situation item, whose 2025 coding is under query (Section 4.5), raises each by about one point.

The EU-membership item is the worked example, because it is the headline and because its wording is unchanged across the three editions. A second example on a domestic item follows in Section 4.1. In 2025, 86 percent of respondents in Albania, 78 percent in Kosovo*, 62 percent in North Macedonia, 59 percent in Bosnia and Herzegovina, 59 percent in Montenegro and 42 percent in Serbia said membership would be good. Albania's first place, Kosovo*'s second and Serbia's sixth are supported at every design effect considered, by the pairwise tests and by the simulated rank intervals (Figure 1). The Barometer can say, and its readers can repeat, that support is highest in Albania and Kosovo* and lowest in Serbia, that it exceeds half in five economies, and that the gap between the top and the bottom is more than forty points.

Figure 1. Support for EU membership by economy, 2025: point estimates with 95 percent intervals under simple random sampling (dark) and a design effect of 2.0 (grey), and the simulated 95 percent rank interval of each economy at a design effect of 2.0.

Table 1. Rank resolution of the 2025 headline items: pairwise economy differences resolvable at the 5 percent level, of fifteen, unadjusted and Holm-adjusted, under three design effects; adjacent-rank pairs resolvable (of five) at a design effect of 2.0; and the mean width of the simulated 95 percent rank interval at 2.0.

Item (2025)	Range (pts)	Pairs resolvable, deff 1.0	1.5	2.0	2.0, Holm	Adjacent pairs resolvable, deff 2.0	Simulated rank-interval width, deff 2.0
Confidence in NATO	78	15	15	15	15	5	1.0
Confidence in armed forces	52	14	14	14	14	4	1.3

Item (2025)	Range (pts)	Pairs resolvable, deff 1.0	1.5	2.0	2.0, Holm	Adjacent pairs resolvable, deff 2.0	Simulated rank-interval width, deff 2.0
Expects EU accession never	29	15	14	14	12	4	1.3
Confidence in the European Union	47	15	14	13	13	3	1.7
EU membership would mean economic prosperity	42	13	13	13	13	3	1.7
Confidence in government	38	13	13	13	13	3	1.7
Confidence in police	54	13	13	13	13	3	1.7
Confidence in courts	37	13	13	13	12	3	1.7
EU membership would be good	45	12	12	12	12	3	2.0
Confidence in parliament	31	12	12	12	12	3	2.0
Confidence in political parties	28	12	12	12	12	3	2.0
Economic situation among most important problems	30	13	13	12	11	2	2.0
What brings citizens together more important (agree)	40	13	13	12	11	2	2.0
Rising prices among most important problems	24	14	13	12	8	2	2.0
Used AI tools in last 12 months	22	11	11	11	11	2	2.3
Corruption among most important problems	41	12	12	11	11	2	2.3
Relations in the region better than 12 months ago (agree)	27	11	11	11	10	2	2.3
Used online shopping in last 12 months	21	13	11	11	8	2	2.2
Confidence in religious organisations (a)	31	12	11	10	9	1	2.5

Item (2025)	Range (pts)	Pairs resolvable, deff 1.0	1.5	2.0	2.0, Holm	Adjacent pairs resolvable, deff 2.0	Simulated rank-interval width, deff 2.0
Regional cooperation contributes to the economy (agree)	25	14	12	10	9	2	2.3
Fake news increased in past 12 months (yes)	18	10	10	10	7	1	2.7
Unemployment among most important problems	18	12	11	10	6	2	2.3
Expects own financial situation to be better	17	9	9	9	9	1	3.0
Lack of trust due to history hampers cooperation	21	12	12	9	6	1	2.7
First feeling about the Western Balkans: hope	12	11	10	8	5	0	3.0
Does not use the Internet	10	11	10	8	4	1	3.0
Internet crime among digital worries	13	8	7	6	3	0	3.8
Economic situation good (top two; see 4.5)	5	10	8	6	3	0	3.5
Air pollution among environmental concerns	13	9	5	5	2	0	4.0
Would consider living and working abroad (yes)	9	8	4	4	2	0	4.0

Note. Valid-response n = 878 to 1,008 per economy, except the economic-situation item (703 to 871; Section 4.5). Range is the difference between the highest and lowest economy. Items marked (a) are adapted per economy (the religious-organisations item names churches, mosques or religious organisations locally), so their cross-economy comparison assumes the adaptation is equivalent, which the files cannot test. The economic-situation item is retained for completeness; its 2025 coding is under query (Section 4.5), and the column totals rise by about one point without it.

3.3 Within-edition tests are interpretable, with one exception

Within each edition, the institutional battery meets the conditions that make a cross-economy test interpretable, with one exception. The 2025 battery passes the single-group gate in every economy (CFI .961 to .993, SRMR .030 to .066, standardised loadings .41 to .97, Appendix B), its thresholds

and loadings are invariant across the six economies by the CFI criterion (configural .987, threshold .985, metric .986), and the two cells that break scalar invariance can be identified and freed (Section 4.4). The 2024 battery passes the gate in every economy once the local dependence between its "military and law enforcement" and "police" items is allowed (single-group CFI .980 to .997; pooled configural CFI .963 with the residual fixed, .991 with it freed; metric .991). The 2023 battery does not pass the gate in Kosovo*: its single-group CFI there is 0.53 (SRMR 0.14, loadings 0.46 to 0.71), and the five-common-item version is little better (CFI .81, police loading .23), so the 2023 six-economy sequence is not interpretable and the 2023 verdicts are taken from the five other economies, where the battery is metric invariant and scalar invariant at the .02 tolerance (Table 2). The scaled chi-square difference tests reject every step in every battery at these sample sizes, as they do for any large multi-group model, and are reported in Appendix B. What this licenses is stated in Section 4.4: comparisons of economies on the anchor items, and on the latent mean once the non-invariant cells are freed.

The four-item regional-cooperation battery, which asks whether cooperation can improve the economic, political and security situation and overall wellbeing, is the cautionary case. Across all six economies its CFI barely moves through the sequence (.998 configural to .997 scalar), which read alone would make it the one fully invariant instrument. It is not. In Kosovo* the four items are nearly uncorrelated (polychoric -.01 to .36 against .58 to .92 in the other five economies) and two of them load at .16 and .17 where the other economies load at .74 to .96. The configural model does not hold there (CFI 0.66), and a CFI at ceiling cannot show it. In the five economies where the items do form a scale, the battery is invariant by the CFI criterion at every step (change at most .0006). Excluding the 43 to 56 percent of respondents in those economies who give the same answer to all four items lowers the configural CFI to .995 and gives step changes of .010, .000 and .011, so the five-economy verdict rests on the .02 tolerance once the straightliners who produce the ceiling are removed. The published percentages on these four items are therefore comparable across five economies at that tolerance, and the Kosovo* figures need a query before they are combined into a scale or compared on the latent mean (Section 4.5). The same Kosovo* pattern, weak inter-item structure with very low straightlining, appears on the 2023 institutional battery and not on the 2024 one, so it spans two editions and two contractors (Section 4.5).

4. What the Barometer cannot rank

4.1 Adjacent positions

The same arithmetic that supports the large contrasts withholds support from the small ones. Of the five differences between adjacent ranks on each item, 51 percent are resolvable under simple random sampling and 38 percent at a design effect of 2.0, or 43 and 28 percent with the Holm adjustment. For the thirty 2025 items, 40 percent of adjacent pairs are resolvable at 2.0 unadjusted and 31 percent adjusted. A complete six-way ordering is supported by the simulated intervals on 3 percent of the 2025 items and 1 percent of all item-editions at a design effect of 2.0, and the mean simulated rank interval is 2.3 positions wide for the 2025 items and 2.4 across all editions. On the EU-membership item in 2025, the three middle economies, North Macedonia, Bosnia and Herzegovina and Montenegro, at 62, 59 and 59 percent, share the positions three to five (Figure 1). A domestic example makes the same point: on the share naming corruption among the most important problems, the six economies span 41 points and eleven of fifteen pairs resolve at a design

effect of 2.0, but North Macedonia (54 percent) and Serbia (48) share the first two positions, Bosnia and Herzegovina, Albania and Montenegro (34, 31 and 31) share the middle three, and only Kosovo* (12) has a position of its own. The median gap between adjacent economies across the 2025 items is 5.0 points, close to the 6.2-point threshold at a design effect of 2.0, so the typical adjacent pair sits just below resolution.

The general statement is simple and the Barometer's reports do not make it: with a thousand interviews per economy and a clustered design, an economy difference of six points or more is a finding on the published dichotomies, and a difference of two or three points is not. The rank intervals in Table 1 are the form in which the ordering would need to be published to make the distinction visible. A comparison of means or of full ordinal distributions on the same items could resolve somewhat more, and the six-point rule is a property of the percentages the RCC prints near one half.

4.2 The regional figure

The Barometer's "WB6" or "Western Balkans" percentage is the arithmetic mean of the six economy percentages. The 2025 key findings say so: the results are "aggregated results from WB6, each contributing equally to the overall dataset. No weighting was applied, ensuring that the regional perspectives are balanced" (RCC, 2025a). The reports for 2023 and 2024 do not say how the figure is formed. Equal weighting is a defensible convention for a council of six equal members, and it is the estimand the RCC has chosen: the average opinion of an average economy. The difficulty is only that it is not the quantity a reader supplies when the figure is reported as what the region's citizens think, which is the opinion of the region's population, in which Serbia's 6.5 million residents count for ten times Montenegro's 0.6 million, and that no report before 2025 said which of the two was published.

Under 2024 population weights, the WB6 figure moves by 2.5 points on average across the 77 item-editions (median 1.8), by three points or more on 32 percent of them and by five or more on 10 percent (Figure 2). The direction and size of a shift are properties of the population distribution and of where the item's support lies, not evidence about any economy's opinion. The largest shifts are on the items most often quoted. Support for EU membership in 2025 is 64 percent equal-weighted (published 64) and 58 percent population-weighted, with standard errors of 0.8 and 1.1 points at a design effect of 2.0. In 2024 it was 54 and 49. The record-high headline survives either way, and the rise from 2024 is 10 points on the published basis and 8.5 on the population basis. Confidence in NATO is 57 percent equal-weighted (published 58) and 45 percent population-weighted, a shift of 11.4 points. Confidence in the European Union is 62 (published 63) and 54, agreement that what brings citizens together matters more than what separates them 53 and 46, and confidence in the police 55 (published 56) and 48. Where the published figure differs from our equal-weight recomputation by a point, the cause is the don't-know base of the 2025 tables (Section 4.3). The share expecting never to join the EU moves the other way, from 17 to 22 percent, because it is highest in the most populous economy.

Figure 2. The WB6 figure for the 2025 headline items under equal weights, as published, and under 2024 population weights.

The convention itself has changed without notice. The database's hidden share column records each economy's contribution to the regional figure: population-proportional shares for the editions from

2015 to 2019 (Serbia 33 to 42 percent, Montenegro 3 percent) and equal shares of 17 percent from 2020. The regional series that the database displays from 2015 to 2025 therefore changes estimand in 2020, and a reader comparing a 2019 regional figure with a 2025 one is comparing population opinion with average-economy opinion. In the 2019 row North Macedonia is shown at 42 percent and Serbia at 11, the reverse of their population shares, a labelling error that does not affect the published percentages but shows that the column has not been checked.

4.3 Year-on-year change

The third of the Barometer's claims is change: support "up from 54 percent in 2024", rising prices "up slightly from 64 percent" (RCC, 2025a). Two conditions must hold for such a claim: the same question must have been asked the same way, and the movement must exceed sampling error. On the first, both of the RCC's own 2025 change claims are on items that meet it, which is to the RCC's credit. Across the headline set the picture is different. Of the twenty-three series present in all three editions, four are identical or trivially reworded in wording, scale, base and option set throughout: EU membership, the first feeling about the Western Balkans, whether fake news has increased, and non-use of the Internet. Between 2023 and 2024, 14 of 23 series are comparable, and between 2024 and 2025, 9 of 24 (Appendix A). The changes are not subtle. The economic-situation item moved from a satisfaction scale to a very-good-to-very-bad scale. Household finances moved from five points to eleven, and financial expectations from three categories to five. The most-important-problems question moved from two mentions among sixteen options to three mentions among fourteen, and the option "rising prices" did not exist in 2023, so the 2025 comparison with 2024 is the only one the series allows. The agreement items gained a midpoint in 2025. The institutional battery moved from "trust" (four points) to "confidence" (the World Values Survey wording) and from including don't-know in the base to excluding it, a base change that alone moves the 2025 institutional percentages by 0.8 points on average and 1.8 at most. The abroad item changed its time frame twice and its answer set once, the accession-expectation item gained three options, and the Internet-hours item changed its bands every year. These are the changes Smith (2005) had in mind in stating Duncan's law for trend measurement: to measure change, do not change the measure. Schwarz (1999) documents the size of the effects that scale and wording changes alone produce, and that expectation is the reason to separate the series.

The database is more careful than the reports. Where the RCC treats a new wording as a new series it opens a new table under the same indicator heading, and 21 of the 63 indicators have two or more tables. But the tables are stacked under one heading with no break marker, and within a table the category set is the union across editions, so the Internet-hours table shows eight bands of which no edition used more than five. A reader of the key findings sees neither.

On the second condition, the movements on the comparable series clear sampling error at a design effect of 2.0 in about half of the economy-by-year cases. Movements on the changed series are considerably larger: between 2024 and 2025 the mean absolute movement is 14.4 points on the thirteen changed series that carry a 2025 value (78 economy-movements) against 6.3 on the nine stable ones (54; Mann-Whitney $p < .001$). Between 2023 and 2024 it is 12.0 against 8.9 (54 against 84; $p = .07$). The contrast is descriptive, since content and comparability are confounded, but its size across the 2025 boundary, where most of the scale and base changes fall, is what a change of instrument would be expected to add.

The largest single movements on the stable series are in one economy and one year. In Montenegro, support for EU membership fell from 60 to 39 percent between 2023 and 2024 and rose to 58 in 2025. The share saying fake news had increased fell from 71 to 38 and rose to 53, and the share naming hope as their first feeling about the region fell from 16 to 11 and rose to 18. Two readings are available. One is that opinion in Montenegro moved and moved back, which the political calendar of those years could support. The other is a fieldwork effect in the 2024 edition: the Montenegrin sample that year has the highest don't-know rate of the 2024 edition across the single-response items (8.7 percent) and 10.5 percent don't-know on the EU item itself against 1.4 to 5.2 percent elsewhere, the highest straightlining on the institutional battery of the 2024 edition (39 percent of those who answered all nine items, 33 percent of all respondents; Figure 3), and a fourteen-day field period in the file's interview dates against 17 to 31 days in the other five. The file allows one clean test of the second reading and it does not support it: EU support among straightliners (41 percent, n = 328) and non-straightliners (37.5 percent, n = 672) does not differ (p = .30). Support was higher in the first fieldwork week (47 percent on 137 interviews) than afterwards (37 percent on 863; p = .04), but in a random-route design fieldwork proceeds area by area, so the week contrast is confounded with geography and cannot be read as drift or as its absence. The files contain nothing else, no interviewer, sampling point or contact record, that would decide between the two readings, and that is the point about documentation rather than about either reading.

Figure 3. Straightlining on the institutional battery by economy and edition: the percentage giving an identical answer to every item. The batteries differ in length (eleven items in 2023, nine in 2024, seven in 2025), so the rates are not comparable across editions.

4.4 Economy comparisons on the institutional items

The Barometer publishes, item by item, the share of each economy expressing trust or confidence in each institution, and readers order the economies on them: which economy trusts its police most, which its parliament least. A comparison of translated items across groups is a comparison of the construct only if the items relate to it in the same way in each group (metric invariance) and if a given level of the construct produces the same item responses in each group (scalar invariance for the item concerned). Without the second, a difference in an item's percentage may be a difference in how that item works (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Davidov et al., 2014). The test is standard in cross-national survey research and had not, to our knowledge, been applied to the Barometer.

Table 2 reports it, and the answer is specific rather than general. Metric invariance holds for the institutional battery in every edition. Full scalar invariance fails in 2025 beyond both tolerances (change in CFI .024); in 2023, on the five economies where the model holds, it fails at .01 and holds at .02 (.010); in 2024 the residual-fixed sequence is not interpretable and with the residual freed the change is .009, within both. The failure in 2025 is concentrated in two cells. The alignment locates it in the police item in Kosovo* (aligned intercept 3.44 against 2.41 to 2.69 elsewhere) and the courts item in North Macedonia (1.69 against 2.17 to 2.53), which are also the two largest modification indices under the scalar model (816 and 565, and the next is 138). The DIF analysis confirms it. Police and courts carry the largest economy effects (pseudo-R-squared change .092 and .068), parliament is borderline (.026), civil service is just below the flag (.019), and government, parties and elections show none. At the .035 benchmark only police and courts are flagged. Freeing the police and courts thresholds in every economy gives a five-anchor model that fits as well as the metric model (CFI .983 against .986, a change of .003), so the other five items are comparable across all six economies as printed. Freeing them only in the two cells, police in Kosovo* and courts in North Macedonia, gives a model with six fewer free parameters that still sits within the .01 tolerance of the metric model (CFI 0.980, a change of 0.006, against .024 for the full scalar model), so the police item is comparable across the five economies other than Kosovo* and the courts item across the five other than North Macedonia, at that tolerance. The substantive reading is that, at any given level of confidence in the other institutions, Kosovo* respondents are far more likely to express confidence in the police (their police thresholds sit 1.1 to 1.4 standard units below the common ones; on the aligned intercepts, three-quarters of a point to a point more on the four-point scale) and North Macedonian respondents less likely in the courts (thresholds 0.9 to 1.1 units above; about half a point less). These are institution-specific effects that the RCC could report as findings, not defects of the survey. Whether they owe anything to translation the files cannot say.

What the two cells do to the ordering is the point for readers. On raw scores the economies rank Kosovo* (2.49 on a 1-to-4 scale), Albania (2.24), North Macedonia (2.22), Montenegro (2.08), Serbia (1.90) and Bosnia and Herzegovina (1.75), and the full scalar model, which forces the two cells to fit, reproduces that order. Once the two cells are freed, Kosovo* and North Macedonia cannot be separated at the top (latent means 0.20 and 0.11 with standard errors 0.04 and 0.05, Albania at 0; the five-anchor model gives .20 and .17), and Albania sits below Kosovo* and, marginally, below North Macedonia. On the aligned means (Kosovo* .05, North Macedonia .12, Albania 0, bootstrap standard errors .06 to .08) the three cannot be separated at all, and their bootstrap rank intervals run 1 to 3, 1 to 2 and 1 to 3. Montenegro, Serbia and Bosnia and Herzegovina are separated in that order on the partial-scalar and aligned models; on the full scalar model Serbia and Bosnia and Herzegovina are not separable (-.65 and -.69). Kosovo*'s clear first place on the raw scores is a property of its police item. A two-factor model that separates the five elected or administrative institutions from police and courts fits better (configural CFI .994, SRMR .031) and is scalar invariant at .02 but not at .01 (change .014); on the elected-institutions factor Kosovo* and North Macedonia tie (.20 and .15), and the two-indicator police-and-courts factor restates the two cells (Kosovo* .71 against -.80 to 0).

Two further batteries are not comparable across economies as published. The 2025 trust-in-information battery, modelled as two factors (traditional and online sources), fails at the metric step by the .01 criterion (change in CFI .017) and passes it at .02; on the strict criterion the six sources do not relate to their factors the same way across economies and no economy mean on it is

comparable, and on the lenient one the scalar step fails (.026) so economy means remain incomparable either way. The regional-cooperation battery does not form a scale in Kosovo* (Section 3.3).

Table 2. Measurement invariance of the multi-item batteries across the six economies: scaled CFI at each step, the change from the previous step, and the verdict under the .01 (Chen, 2007) and .02 (Rutkowski & Svetina, 2014) tolerances.

Battery	n	Configural	Thresholds (change)	Metric (change)	Scalar (change)	Partial (change vs metric)	Verdict at .01 / .02
Institutional 2025, 7 items, six economies	5,407	0.987	0.985 (-0.002)	0.986 (+0.001)	0.962 (-0.024)	cells: 0.980 (-0.006); anchors: 0.983 (-0.003)	metric holds; scalar fails / fails; partial (cells) holds / holds
Institutional 2024, 7 items, military-police residual freed	5,358	0.991	0.991 (-0.001)	0.991 (-0.000)	0.981 (-0.009)		metric holds; scalar holds (.009) / holds
Institutional 2024, residual fixed (baseline fails the gate; not interpretable)	5,358	0.963	0.962 (-0.001)	0.959 (-0.002)	0.953 (-0.006)		not interpretable
Institutional 2023, 7 items, five economies without Kosovo* (Kosovo* fails the gate, CFI .53)	4,503	0.970	0.970 (+0.000)	0.970 (-0.000)	0.960 (-0.010)		metric holds; scalar fails (.010) / holds
Institutional 2023, 7 items, six economies (not interpretable)	5,450	0.956	0.953 (-0.003)	0.954 (+0.001)	0.935 (-0.019)		not interpretable
Five common items 2025, six economies	5,588	0.989	0.987 (-0.001)	0.988 (+0.001)	0.960 (-0.028)		metric holds; scalar fails / fails

Battery	n	Configural	Thresholds (change)	Metric (change)	Scalar (change)	Partial (change vs metric)	Verdict at .01 / .02
Five common items 2024, six economies	5,541	0.994	0.993 (-0.001)	0.993 (-0.000)	0.983 (-0.010)		metric holds; scalar holds (.010) / holds
Five common items 2023, five economies without Kosovo*	4,662	0.995	0.994 (-0.001)	0.992 (-0.002)	0.983 (-0.009)		metric holds; scalar holds (.009) / holds
Regional cooperation 2025, six economies (Kosovo* fails the gate, CFI .66; at ceiling)	5,717	0.998	0.998 (-0.000)	0.997 (-0.000)	0.997 (-0.000)		not interpretable
Regional cooperation 2025, five economies without Kosovo*	4,777	0.998	0.998 (+0.000)	0.998 (-0.000)	0.998 (-0.001)		scalar holds / holds
Regional cooperation 2025, five economies, straightliners excluded	2,434	0.995	0.985 (-0.010)	0.985 (-0.000)	0.974 (-0.011)		thresholds holds (.010); scalar fails (.011) / holds
Trust in information 2025 (6 items, ten-point, two factors, robust ML)	4,962	0.959		0.943 (-0.017)	0.917 (-0.026)		metric fails / holds; scalar fails / fails

Note. WLSMV with the Wu and Estabrook (2016) identification for ordered items; robust ML for the ten-point items. Thresholds = thresholds constrained; metric = thresholds and loadings; scalar = thresholds, loadings and intercepts; partial (cells) = scalar with the police thresholds freed in Kosovo and the courts thresholds in North Macedonia; partial (anchors) = scalar with both items' thresholds freed in every economy. Single-group gate: CFI at least .95 and SRMR at most .08. Scaled chi-square difference tests reject every step at $p < .001$ in every battery. Full fit tables with RMSEA intervals, per-economy fit and per-economy n in Appendix B.*

Response style differs by economy in ways a published percentage does not show, and its relationship to the non-invariance is not simple. Straightlining on the institutional battery is lowest in Kosovo* in 2023 and 2025 (5 and 7 percent) but not in 2024 (24 percent, fifth of six), and highest in Serbia in 2025 (27 percent) and Montenegro in 2024 (39 percent). The extreme-response rate on the 2025 battery runs from about 32 percent in Albania and Kosovo* to 47 percent in Serbia. Response styles of this size are a recognised source of non-invariance (Baumgartner & Steenkamp, 2001), but here the economy with the largest intercept deviation, Kosovo*, has one of the two lowest extreme-response rates, and the sensitivity that excludes straightliners makes the scalar failure larger (change in CFI .049), not smaller, so the two cells are not a straightlining artefact.

Language and context can be probed in the 2023 file, whose nationality variable gives economy-by-nationality groups: Albanians in Albania (857) and North Macedonia (204), Macedonians in North Macedonia (590), Bosniaks (455) and Serbs (337) in Bosnia and Herzegovina, Montenegrins (474) and Serbs (314) in Montenegro, and Serbs in Serbia (836); Albanians in Kosovo* (855) are excluded because the battery does not fit there (Section 3.3). Across these eight groups the 2023 battery is metric invariant (change in CFI .001) and scalar invariant at both tolerances (.009), so the groups' item functioning differs less than the six economies' did with Kosovo* included. Where it differs, the aligned intercept profiles do not cluster by language. The mean absolute difference between two groups' aligned intercepts is 0.08 for Serbs and Bosniaks in Bosnia and Herzegovina, 0.08 for Serbs and Montenegrins in Montenegro and 0.12 for Macedonians and Albanians in North Macedonia (the largest single item difference within a context is .27, on the judiciary item in North Macedonia); the same-language pairs across contexts are 0.12 for Albanians in Albania and North Macedonia, 0.20 for Serbs in Montenegro and Serbia and 0.11 for Serbs in Bosnia and Herzegovina and Serbia, and the closest pairs of all share neither (0.05, 0.05). Levels differ a great deal (the two North Macedonian groups' aligned means differ by .40), but item functioning between two language groups in one context is no further apart than between two contexts in one language. Within the limits of a nationality proxy, and on the presumption, undocumented, that one Serbian version served all three economies where Serbs were interviewed, this is evidence that the non-invariance does not follow language; it cannot attribute it to context either, and the language-of-interview variable the RCC's contractors hold is the proper test.

What could not be tested is as important as what could. Language of interview is unrecorded, so within-economy version effects can only be proxied as above, and in one edition. Interviewer effects, a documented source of variable error in face-to-face surveys (West & Blom, 2017), cannot be estimated because no file records the interviewer. Mode effects do not arise, since every edition is CAPI. Each of these is a variable that the contractor holds and that a documented file could carry.

4.5 Errors found

We report three findings for correction rather than as measurement properties. All three are put to the RCC in the courtesy copy that precedes posting (see Declarations). First, in the database table for the economic-situation item, the 2024 rows for Kosovo* and North Macedonia are exchanged: the row published as Kosovo* (very good 1, rather good 6, neither 24, rather bad 38, very bad 31) is the microdata distribution for North Macedonia, and the row published as North Macedonia (7, 18, 48, 23, 4) is the distribution for Kosovo*. The 2024 report describes North Macedonia as the most negative economy and Kosovo* as the most optimistic on the companion item (RCC, 2024, p. 27), which agrees with the microdata and not with the database. Second, the 2019 share column

exchanges North Macedonia and Serbia, as noted in Section 4.2. Third, in the 2025 file the economic-situation and political-situation items contain no "very good" responses in any economy, although the value labels list the category, and carry 13 to 34 percent system-missing values (Bosnia and Herzegovina 30 percent on the economic item, 34 on the political), where the 2024 file has 1 to 7 percent "very good" and 0 to 2 percent don't-know. The published 2025 percentages for these items are computed on the non-missing base, so Bosnia and Herzegovina is published as 62 percent "very bad" where the all-respondent figure is 44. A category lost in processing, a routing filter we cannot see, and a third of respondents declining to answer are all consistent with the file as released. Only the RCC or its contractor can say which, and the economic-situation series is treated here as unreliable for 2025. A fourth pattern is reported as a query rather than an error. In Kosovo* the multi-item batteries of two editions show weak inter-item structure: on the four 2025 regional-cooperation items the responses are nearly uncorrelated (polychoric -.01 to .36 against .58 to .92 elsewhere) and only 11 percent of respondents give the same answer to all four against 43 to 56 percent elsewhere, and on the 2023 institutional battery the one-factor model does not fit (CFI .53) where it fits in the other five economies (Section 3.3); the 2024 Kosovo* battery shows neither feature. The pattern spans two editions and two contractors, the published Kosovo* marginals on these items reproduce from the file to within a point, and it is reported without attribution of cause: it may be how the batteries were answered, how they were administered, or how the files were assembled, and only the RCC and its contractors can say.

5. What a technical report would need to contain

The findings of Sections 3 and 4 are not findings about the region. They are findings about what a reader can and cannot learn from the Barometer as it is published. We score in Table 3 the public documentation of the three editions against sixteen elements drawn from what the AAPOR Transparency Initiative, the Standard Eurobarometer and the European Social Survey publish routinely. The three programmes are the source of the elements, not scored subjects, and the ESS in particular has a documentation budget no regional survey matches, so the score describes a gap, not a rebuke. The two published editions score 16 of 32, and the 2025 edition, scored provisionally on the site's pages, 15. Three elements score zero in every edition: precision (no margin of error, confidence interval, standard error or design effect is published in the three editions examined), comparability (no edition lists what changed in the questionnaire) and limitations. Weighting scores zero in two editions and one in the third, because the only statement about it in the three editions is the 2025 sentence that none was applied to the regional figure, while the 2023 file carries a weight that the 2023 report appears to have used and that no published 2023 document describes. The full questionnaire, the translation protocol and the response rate score one at best: the wording can be recovered item by item from the figures and the SPSS labels, the 2024 note names two languages and one undefined rate, and nothing describes how the translations were produced, assessed or pretested. Microdata access scores one in every edition: the files exist, which is more than most regional surveys offer, but without a codebook, without design variables and without terms of use.

Table 3. Transparency of the public documentation, scored 0 (not found), 1 (partial) or 2 (published and specific) against the reference programmes; single rater, written rubric, evidence for each cell in Appendix C. The 2025 score is provisional (report not yet published).

Element	Reference programmes	2023	2024	2025
Sponsor and data-collection organisation named	AAPOR; EB; ESS	2	2	2
Population under study defined	AAPOR; EB; ESS	2	2	1
Sampling frame and design described	AAPOR; EB; ESS	2	1	1
Completed interviews per economy	AAPOR; EB; ESS	2	2	2
Fieldwork dates	AAPOR; EB; ESS	2	2	1
Mode of data collection	AAPOR; EB; ESS	2	2	2
Full questionnaire published	AAPOR; EB; ESS	1	1	1
Translation protocol and language versions	ESS; EB	1	1	1
Response rate or contact outcomes	AAPOR; ESS	0	1	0
Weighting described	AAPOR; EB; ESS	0	0	1
Precision: margin of error or design effect	AAPOR; EB; ESS	0	0	0
Data processing and quality control	AAPOR; ESS	1	1	0
Comparability of the series documented	ESS; EB	0	0	0
Limitations statement	AAPOR	0	0	0
Microdata access with documentation	ESS; EB	1	1	1
Regional aggregate: composition and weighting stated	EB	0	0	2
Total (of 32)		16	16	15

Four additions would close two of the three elements that score zero (precision and comparability), raise weighting, the questionnaire, translation and microdata access from one to two, and leave a limitations statement, a defined response rate and quality-control documentation as the remaining gaps; none requires a change to the questionnaire or to the fieldwork design.

First, a precision statement. One table, of the kind the Standard Eurobarometer prints in every technical annex, giving the confidence interval for a percentage at the economy sample size, together with the design effect estimated from the sampling-point structure the contractor already holds, would allow every reader to apply the six-point rule of Section 4.1 themselves. The rank intervals of Table 1 are a compact way of publishing the same information for an ordering.

Second, a statement of what the regional figure estimates. Either the equal-weight convention is stated in every report, with the population-weighted figure alongside it for the headline items, or the aggregate is population-weighted, as the Eurobarometer's EU averages are, with the convention

stated. Either is defensible. Silence is not, and the 2020 change of convention needs a note in the database.

Third, a change log. A one-page annex per edition listing each item whose wording, scale, base or option set differs from the previous edition, with the RCC's own judgement of which series it regards as continuous, would let the key findings distinguish a change in opinion from a change in question. The database's table structure already encodes half of this and needs only to be made visible.

Fourth, a documented file. The released microdata would be a research resource of the first rank for the region if the public file carried the language of interview, the weights actually used, a codebook with the questionnaire and its translations, and terms of use, and if a documented restricted file carried the sampling-point and interviewer identifiers, which have disclosure risk in small economies. With those variables, the tests of Section 4.4 could be run within economies and across interviewers rather than only across economies, and the RCC could report measurement invariance for its batteries as the European Social Survey does for its own.

For a reader of the Barometer today, four rules follow from Sections 3 and 4. Treat an economy difference of six points or more as a finding and a smaller one as a tie. Treat the WB6 figure as the average of six economies, not the region's population, until a report says otherwise. Check the change log in Appendix A before reading a year-on-year movement. And read economy comparisons on the institutional items with the police and courts items set aside.

6. Limitations and conclusion

The audit is bounded by what the files contain. The design effect is proxied from municipalities and regions, not estimated from sampling points. The proxies sit between 1.2 and 2.2 and the headline results are given at 2.0, with the full range in the package. The rank inference uses pairwise tests and simulated intervals on the RCC's published dichotomies, and comparisons on means or full distributions would resolve somewhat more. The invariance tests are unweighted, listwise on don't-know, and treat economy as the group. The nationality analysis separates language from context for one edition only, and the alignment and DIF criteria are conventional. The population weights use total population rather than the adult population, a difference of at most 0.61 points in any weight. The 2025 report was not yet published when the audit was run, so the 2025 documentation was scored on the site's About and key-findings pages, and a full report may raise its score. The crosswalk covers the headline items and not the whole questionnaire, and the 2017 to 2022 files, which are also public, were not examined. We examined the public-opinion survey only. The business survey, fielded by a different contractor, was not examined.

Within those bounds our conclusion is plain. The Balkan Barometer can rank what it is precise enough to rank: economy differences of six points or more, the broad contrasts that make up most of its headlines, and economy comparisons on items that work the same way everywhere, which the institutional battery does once two items are set aside. It cannot rank adjacent positions, it cannot report its regional figure as the region's opinion without saying which quantity it publishes, it cannot report year-on-year change on the series that changed, and it cannot report the police and courts items as comparable in the one economy each where they behave differently. None of this is a criticism of the fieldwork, which reproduces its published numbers to the point, or of the RCC's decision to release the files, which is what made the audit possible and is more than most regional programmes do. It is a description of the distance between what the instrument measures and what

is said in its name, and of the four documents that would close it. The method transfers: any survey programme that releases microdata without design variables can be audited the same way, with the geography the file carries standing in for the sampling design it does not.

Data and code availability

All data are public: the microdata zips and the interactive database at rcc.int/balkanbarometer, the 2023 and 2024 reports at rcc.int, and the World Bank World Development Indicators. The analysis package (extraction, crosswalk, resolution, reweighting, invariance, DIF, checklist and figures, in Python and R, with session information and seeds) and the results register are deposited at Zenodo, https://doi.org/10.5281/zenodo.22919169. The package contains code and derived results only. The microdata carry no published terms of use and are not redistributed.

Declarations

The author received no funding for this work and reports no competing interests. The author is a resident of Montenegro and has no relationship with the Regional Cooperation Council or with any of the fieldwork contractors named. The three items reported for correction in Section 4.5 and the query that follows them are addressed to the Regional Cooperation Council, which had not been notified of them when this version was posted; any correction the RCC supplies will be reflected in a later version. The designation Kosovo* is used as the RCC uses it, without prejudice to positions on status, and in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence. Generative AI tools (Anthropic Claude) were used to assist with data extraction, code and drafting under the author's direction. Every analysis was specified, checked and interpreted by the author, who is responsible for the content.

References

AAPOR. (2026). *Transparency Initiative: items for immediate disclosure*. American Association for Public Opinion Research. https://aapor.org/standards-and-ethics/transparency-initiative/

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1-30. https://doi.org/10.18637/jss.v039.i08

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1), 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

ESS. (2026). *ESS methodology*. European Social Survey. https://www.europeansocialsurvey.org/methodology

European Commission. (2017). *Standard Eurobarometer 88, autumn 2017: Technical specifications* [Annex to *Public opinion in the European Union: First results*]. Directorate-General for Communication.

European Commission. (2025). *Public opinion in the European Union: Eurobarometer report, March-April 2025* (Standard Eurobarometer 103.3). Directorate-General for Communication, Publications Office of the European Union. https://doi.org/10.2775/1134214

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A, 159*(3), 385-443. https://doi.org/10.2307/2983325

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849-879. https://doi.org/10.1093/poq/nfq065

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115-140). Wiley.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (R package version 0.5-8).

Kankaraš, M. (2026). *The illusion of precision: Interpreting international rankings within their measurement limits*. SocArXiv. https://doi.org/10.31235/osf.io/mbgxs_v1

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. *Social Science Computer Review, 37*(2), 214-233. https://doi.org/10.1177/0894439317752406

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*(2), 183-200.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods and Research, 47*(4), 637-664. https://doi.org/10.1177/0049124117701488

RCC. (2023). *Balkan Barometer 2023: Public opinion*. Regional Cooperation Council. https://www.rcc.int/download/docs/BB%202023-PO.pdf/5495f3e223e456e99fc3bdce76054b7e.pdf

RCC. (2024). *Balkan Barometer 2024: Public opinion*. Regional Cooperation Council. https://www.rcc.int/download/docs/BB2024-PO.pdf/c29cfed20c3776d280077cdfc2617abc.pdf

RCC. (2025a). *Balkan Barometer 2025: Public opinion key findings*. Regional Cooperation Council. https://www.rcc.int/balkanbarometer/key_findings_2025/2/public

RCC. (2025b). *Balkan Barometer: About*. Regional Cooperation Council. https://www.rcc.int/balkanbarometer/about

RCC. (2025c). *Balkan Barometer 2025 public presentation, 11 December 2025, Brussels*. Regional Cooperation Council. https://www.rcc.int/events/1642/balkan-barometer-2025-public-presentation

Robitzsch, A. (2024). *sirt: Supplementary item response theory models* (R package version 4.2-133).

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis* (pp. 233-247). Kluwer. https://doi.org/10.1007/978-1-4615-4603-0_17

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. https://doi.org/10.1037/0003-066x.54.2.93

Smith, T. W. (2005). *The laws of studying societal change* (GSS Social Change Report No. 50). NORC at the University of Chicago. https://gss.norc.org/content/dam/gss/get-documentation/pdf/reports/social-change-reports/SC50.pdf

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-107. https://doi.org/10.1086/209528

Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys* (4th ed.). Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175-211. https://doi.org/10.1093/jssam/smw024

World Bank. (2026). *World Development Indicators: Population, total (SP.POP.TOTL); Population ages 0-14 (SP.POP.0014.TO.ZS)*. Retrieved 18 September 2026.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014-1045. https://doi.org/10.1007/s11336-016-9506-0

Appendix A. Headline crosswalk

The table lists each headline series, the variable that carries it in each edition, the grade of each adjacent pair (identical; reworded, meaning a trivial wording change with the same scale, base and options; scale or base changed; option set changed; construct changed), and what changed. A pair is comparable if identical or reworded. Twenty-three of the twenty-four series are present in all three editions (rising prices was not offered in 2023); four are comparable across all three. Full rules, including the response codes counted, are in the analysis package (bb_common.py).

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
EU membership would be good	B16	Q20	Q15	identical	identical	Identical wording and 3-point plus DK response set in all three editions.
Relations in the Western Balkans better than 12 months ago (agree)	B12	Q15	Q3	reworded	scale or base	2023 and 2024: 4-point agree scale without midpoint (order reversed in 2024). 2025: 5-point scale with 'Neither agree nor disagree' and wording changed to 'political relations'.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
What brings citizens together is more important than what separates (agree)	B14	Q16_A1	Q13	identical	scale or base	4-point scale without midpoint in 2023 and 2024; 5-point with midpoint in 2025.
Regional cooperation can contribute to the situation of the economy (agree)	B15	Q16_A2	Q13.1	reworded	construct	2023: one item, 'political, economic or security situation', 4-point. 2024: one item, 4-point. 2025: four separate items (economic, political, security, wellbeing), 5-point with midpoint; the economic item is used here.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Economic situation of the economy good / satisfied (top two)	A02	Q2	Q4	construct	scale or base	2023: satisfaction wording ('How satisfied are you with the economic situation', completely dissatisfied to completely satisfied). 2024 and 2025: evaluative wording ('very good ... very bad'), with the coding order reversed between 2024 and 2025.
Household financial situation good / satisfied (top two)	A03	Q3	Q1.3	construct	scale or base	2023: 5-point satisfaction. 2024: 5-point good/bad. 2025: 11-point (1 to 10 plus DK) satisfaction scale; no top-two dichotomy is defined, so 2025 is NA here.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Expects own financial situation to be better in 12 months	A05	Q6	Q5	identical	scale or base	2023 and 2024: 3-point (better / worse / the same). 2025: 5-point (much worse to much better); 'better' plus 'much better' used.
Unemployment among the most important problems	A07	Q7_5	Q6.5	option set	identical	2023 asked for the TWO most important problems from 16 options; 2024 and 2025 asked for THREE from 14 and 13 options. Shares are not comparable across the 2023 to 2024 boundary.
Corruption among the most important problems	A07	Q7_6	Q6.6	option set	identical	As problem_unemployment.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Economic situation among the most important problems	A07	Q7_1	Q6.1	option set	identical	As problem_unemployment; 2023 option read 'Economic situation', later 'Economic situation in general'.
Rising prices / inflation among the most important problems	A07	Q7_2	Q6.2	not asked	identical	The option did not exist in the 2023 list; 2024 to 2025 is comparable in wording and mention count.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Trust / confidence in courts (top two)	T_H60_1	Q35_A1	Q33.8	reworded	scale or base	2023: 'trust', totally distrust to totally trust, 'Judicial institutions (e.g. courts)'. 2024: 'trust', do not trust to trust, 'Judicial institutions'. 2025: 'confidence', none at all to a great deal, 'the courts' (World Values Survey wording).
Trust / confidence in parliament (top two)	T_H60_2	Q35_A2	Q33.11	reworded	scale or base	As trust_courts.
Trust / confidence in government (top two)	T_H60_3	Q35_A3	Q33.9	reworded	scale or base	As trust_courts.
Trust / confidence in political parties (top two)	T_H60_5	Q35_A5	Q33.10	not asked	scale or base	As trust_courts.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Trust / confidence in police (top two)	T_H60_11	Q35_A9	Q33.7	reworded	scale or base	As trust_courts.
Trust in public TV and radio (top two)	T_H61_1	Q39_A1	Q35.1	identical	scale or base	4-point trust scale in 2023 and 2024; 11-point (1 to 10 plus DK) in 2025, no top-two dichotomy defined.
Fake news and disinformation increased in past 12 months (yes)	H62	Q40	Q36	reworded	identical	Same yes/no/DK item; 2023 wording 'has there been increase ... of various fake news', later 'a raise of fake news'.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Would consider living and working abroad (yes)	E39	Q28_A1..A4	Q22.1	construct	construct	2023: 'Would you consider living and working abroad?' yes/no. 2024: 'in the next five years', asked per destination (WB, EU, US, other); yes = any destination. 2025: 'if you have a chance', yes permanently / yes temporary / no.
First feeling about the Western Balkans: hope	B13	Q13	Q12	identical	identical	Same item and option set.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Expects EU accession never	B18	Q22	Q17	option set	identical	2023 offered by 2030 / by 2035 / never / DK; 2024 and 2025 offered 2030 / 2035 / 2040 / after 2040 / never / DK. The 'never' share depends on the option set.
Does not use the Internet	D35	Q24	Q18	identical	identical	Hour bands differ every year (2023: at least 1 hour / 2-3 / 4 and more; 2024: up to 1 / 2-3 / 4 and more; 2025: up to 1 / 2-3 / 4-5 / 6 and more); the non-user category, which this indicator uses, is stable in wording and base.

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Air pollution among greatest environmental concerns	G51	Q31_1	Q25.1	option set	option set	Multi-mention; option list changed (2023 had 'extreme temperature', 'biodiversity'; 2024 added 'heavy traffic', 'noise', 'landscape damage').

Series	2023	2024	2025	2023 to 2024	2024 to 2025	What changed
Internet crime among worries about digital tools	D37	Q26_1	Q20.1	scale or base	option set	2023 option read 'Cyber-attacks and cybercrime such as theft or abuse of personal data, ransomware or phishing'; later 'Internet crimes (theft of personal data, theft of credit card information, etc.)'. Base changed: 2023 percentages are of respondents routed to the item (Internet users, D35 answered; report Figure 47, n=5,374), 2024 and 2025 of all respondents.

2025-only items used in Table 1: confidence in the European Union (Q33.17), in religious organisations (Q33.1), in the armed forces (Q33.2) and in NATO (Q33.18); AI-tool use (Q19.5); online

shopping (Q19.4); lack of trust due to history hampers cooperation (Q14.1); EU membership would mean economic prosperity (Q16.3).

Appendix B. Invariance fit tables

Fit of every step, every battery (scaled indices; RMSEA with 90 percent interval; scaled chi-square difference test against the previous step).

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Institutional 2025, 7 items	configural	1503.8	84	0.987	0.981	0.137 (0.131, 0.143)	0.046	
Institutional 2025, 7 items	thresholds	1778.3	119	0.985	0.984	0.124 (0.119, 0.130)	0.046	chi-square diff 382.6, df 35, p < .001
Institutional 2025, 7 items	metric	1697.0	149	0.986	0.988	0.107 (0.103, 0.112)	0.047	chi-square diff 153.3, df 30, p < .001
Institutional 2025, 7 items	scalar	4433.1	179	0.962	0.973	0.163 (0.158, 0.167)	0.048	chi-square diff 2359.0, df 30, p < .001
Institutional 2024, 7 items, residual fixed	configural	5335.4	84	0.963	0.945	0.265 (0.259, 0.271)	0.092	
Institutional 2024, 7 items, residual fixed	thresholds	5576.5	119	0.962	0.959	0.227 (0.222, 0.232)	0.092	chi-square diff 277.2, df 35, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Institutional 2024, 7 items, residual fixed	metric	5918.5	149	0.959	0.966	0.208 (0.204, 0.213)	0.092	chi-square diff 515.6, df 30, p < .001
Institutional 2024, 7 items, residual fixed	scalar	6823.2	179	0.953	0.967	0.204 (0.200, 0.208)	0.092	chi-square diff 1039.5, df 30, p < .001
Institutional 2024, 7 items, residual freed	configural	1302.6	78	0.991	0.986	0.133 (0.126, 0.139)	0.037	
Institutional 2024, 7 items, residual freed	thresholds	1430.8	113	0.991	0.990	0.114 (0.109, 0.120)	0.037	chi-square diff 277.2, df 35, p < .001
Institutional 2024, 7 items, residual freed	metric	1492.2	143	0.991	0.992	0.103 (0.098, 0.108)	0.038	chi-square diff 201.6, df 30, p < .001
Institutional 2024, 7 items, residual freed	scalar	2811.1	173	0.981	0.987	0.131 (0.127, 0.135)	0.038	chi-square diff 1141.1, df 30, p < .001
Institutional 2023, 7 items	configural	3291.0	84	0.956	0.934	0.205 (0.199, 0.211)	0.073	

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Institutional 2023, 7 items	thresholds	3538.9	119	0.953	0.951	0.178 (0.173, 0.183)	0.073	chi-square diff 365.1, df 35, p < .001
Institutional 2023, 7 items	metric	3505.3	149	0.954	0.961	0.158 (0.153, 0.162)	0.074	chi-square diff 328.6, df 30, p < .001
Institutional 2023, 7 items	scalar	4933.3	179	0.935	0.954	0.171 (0.167, 0.175)	0.073	chi-square diff 1385.0, df 30, p < .001
Five common items 2025	configural	999.9	30	0.989	0.977	0.186 (0.177, 0.196)	0.053	
Five common items 2025	thresholds	1142.1	55	0.987	0.986	0.146 (0.139, 0.153)	0.053	chi-square diff 234.4, df 25, p < .001
Five common items 2025	metric	1077.5	75	0.988	0.991	0.120 (0.114, 0.126)	0.054	chi-square diff 90.2, df 20, p < .001
Five common items 2025	scalar	3470.2	95	0.960	0.975	0.195 (0.190, 0.201)	0.055	chi-square diff 2114.0, df 20, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Five common items 2024	configural	636.6	30	0.994	0.988	0.148 (0.138, 0.158)	0.034	
Five common items 2024	thresholds	732.7	55	0.993	0.992	0.116 (0.108, 0.123)	0.034	chi-square diff 209.6, df 25, p < .001
Five common items 2024	metric	796.7	75	0.993	0.994	0.102 (0.096, 0.109)	0.035	chi-square diff 125.6, df 20, p < .001
Five common items 2024	scalar	1754.0	95	0.983	0.989	0.138 (0.132, 0.143)	0.035	chi-square diff 858.0, df 20, p < .001
Five common items 2023	configural	575.7	30	0.989	0.978	0.140 (0.130, 0.150)	0.037	
Five common items 2023	thresholds	785.3	55	0.985	0.984	0.119 (0.112, 0.127)	0.037	chi-square diff 272.4, df 25, p < .001
Five common items 2023	metric	1217.6	75	0.977	0.982	0.128 (0.121, 0.134)	0.043	chi-square diff 429.0, df 20, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Five common items 2023	scalar	1999.3	95	0.962	0.976	0.146 (0.141, 0.152)	0.042	chi-square diff 735.7, df 20, p < .001
Cooperation 2025, six economies	configural	278.3	12	0.998	0.993	0.153 (0.137, 0.169)	0.023	
Cooperation 2025, six economies	thresholds	336.7	52	0.998	0.998	0.076 (0.068, 0.084)	0.023	chi-square diff 123.4, df 40, p < .001
Cooperation 2025, six economies	metric	403.9	67	0.997	0.999	0.073 (0.066, 0.080)	0.027	chi-square diff 71.9, df 15, p < .001
Cooperation 2025, six economies	scalar	475.9	82	0.997	0.999	0.071 (0.065, 0.077)	0.027	chi-square diff 80.1, df 15, p < .001
Cooperation 2025, without Kosovo*	configural	189.6	10	0.998	0.995	0.137 (0.120, 0.155)	0.012	
Cooperation 2025, without Kosovo*	thresholds	223.7	42	0.998	0.999	0.067 (0.059, 0.076)	0.012	chi-square diff 110.1, df 32, p < .001
Cooperation 2025, without Kosovo*	metric	256.3	54	0.998	0.999	0.063 (0.055, 0.070)	0.013	chi-square diff 29.7, df 12, p = 0.003

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Cooperation 2025, without Kosovo*	scalar	342.6	66	0.998	0.999	0.066 (0.059, 0.073)	0.013	chi-square diff 90.8, df 12, p < .001
Cooperation 2025, straightliners excluded	configural	114.4	12	0.985	0.955	0.125 (0.105, 0.147)	0.037	
Cooperation 2025, straightliners excluded	thresholds	228.2	52	0.974	0.982	0.079 (0.069, 0.090)	0.037	chi-square diff 106.0, df 40, p < .001
Cooperation 2025, straightliners excluded	metric	260.7	67	0.972	0.985	0.073 (0.064, 0.082)	0.049	chi-square diff 50.9, df 15, p < .001
Cooperation 2025, straightliners excluded	scalar	328.9	82	0.964	0.984	0.074 (0.066, 0.083)	0.050	chi-square diff 70.5, df 15, p < .001
Institutional 2025, straightliners excluded	configural	915.5	84	0.984	0.975	0.115 (0.109, 0.122)	0.051	
Institutional 2025, straightliners excluded	thresholds	1231.8	119	0.978	0.977	0.112 (0.106, 0.118)	0.051	chi-square diff 331.4, df 35, p < .001
Institutional 2025, straightliners excluded	metric	1142.2	149	0.980	0.983	0.095 (0.090, 0.100)	0.055	chi-square diff 147.3, df 30, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Institutional 2025, straightliners excluded	scalar	3632.0	179	0.932	0.952	0.161 (0.157, 0.166)	0.056	chi-square diff 2174.7, df 30, p < .001
Institutional 2025, two factors	configural	789.1	78	0.994	0.990	0.101 (0.094, 0.107)	0.031	
Institutional 2025, two factors	thresholds	1143.6	113	0.991	0.990	0.101 (0.095, 0.106)	0.031	chi-square diff 382.6, df 35, p < .001
Institutional 2025, two factors	metric	1173.3	138	0.991	0.992	0.091 (0.086, 0.096)	0.032	chi-square diff 127.5, df 25, p < .001
Institutional 2025, two factors	scalar	2721.4	163	0.977	0.982	0.132 (0.128, 0.136)	0.034	chi-square diff 1389.8, df 25, p < .001
2023, nine economy-by-nationality groups	configural	3269.6	126	0.965	0.947	0.214 (0.207, 0.220)	0.076	
2023, nine economy-by-nationality groups	thresholds	3548.5	182	0.962	0.961	0.184 (0.179, 0.189)	0.076	chi-square diff 375.7, df 56, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
2023, nine economy-by-nationality groups	metric	3667.5	230	0.962	0.968	0.166 (0.161, 0.170)	0.079	chi-square diff 440.7, df 48, p < .001
2023, nine economy-by-nationality groups	scalar	5099.8	278	0.946	0.963	0.178 (0.174, 0.182)	0.078	chi-square diff 1478.0, df 48, p < .001
Institutional 2025, partial scalar, five anchors (police and courts thresholds freed in every economy)	partial (anchors)	2064.4	159	0.983	0.986	0.115 (0.111, 0.120)	0.054	
Institutional 2023, five economies without Kosovo*	configural	2162.0	70	0.970	0.955	0.182 (0.176, 0.189)	0.059	
Institutional 2023, five economies without Kosovo*	thresholds	2181.1	98	0.970	0.968	0.154 (0.148, 0.159)	0.059	chi-square diff 177.0, df 28, p < .001
Institutional 2023, five economies without Kosovo*	metric	2214.7	122	0.970	0.974	0.138 (0.133, 0.143)	0.059	chi-square diff 229.0, df 24, p < .001
Institutional 2023, five economies without Kosovo*	scalar	2950.7	146	0.960	0.971	0.146 (0.142, 0.151)	0.059	chi-square diff 744.6, df 24, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Five common items 2023, five economies	configural	261.1	25	0.995	0.990	0.101 (0.090, 0.112)	0.025	
Five common items 2023, five economies	thresholds	354.5	45	0.994	0.993	0.086 (0.078, 0.094)	0.025	chi-square diff 129.3, df 20, p < .001
Five common items 2023, five economies	metric	452.0	61	0.992	0.993	0.083 (0.076, 0.090)	0.026	chi-square diff 113.6, df 16, p < .001
Five common items 2023, five economies	scalar	899.4	77	0.983	0.989	0.107 (0.101, 0.113)	0.026	chi-square diff 423.4, df 16, p < .001
Cooperation 2025, five economies, straightliners excluded	configural	42.8	10	0.995	0.985	0.082 (0.058, 0.108)	0.025	
Cooperation 2025, five economies, straightliners excluded	thresholds	140.4	42	0.985	0.989	0.070 (0.057, 0.082)	0.025	chi-square diff 97.6, df 32, p < .001
Cooperation 2025, five economies, straightliners excluded	metric	152.9	54	0.985	0.992	0.061 (0.050, 0.073)	0.033	chi-square diff 25.6, df 12, p = 0.012
Cooperation 2025, five economies, straightliners excluded	scalar	235.2	66	0.974	0.988	0.073 (0.063, 0.083)	0.034	chi-square diff 74.5, df 12, p < .001

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
2023, eight nationality groups (without Kosovo* Albanians)	configural	2212.3	112	0.976	0.963	0.192 (0.185, 0.199)	0.062	
2023, eight nationality groups (without Kosovo* Albanians)	thresholds	2316.5	161	0.975	0.974	0.162 (0.157, 0.168)	0.062	chi-square diff 204.4, df 49, p < .001
2023, eight nationality groups (without Kosovo* Albanians)	metric	2462.2	203	0.974	0.978	0.148 (0.143, 0.153)	0.065	chi-square diff 331.2, df 42, p < .001
2023, eight nationality groups (without Kosovo* Albanians)	scalar	3229.9	245	0.965	0.976	0.155 (0.150, 0.160)	0.065	chi-square diff 830.9, df 42, p < .001
Institutional 2025, partial scalar, two cells freed	partial (cells)	2416.8	173	0.980	0.985	0.120 (0.116, 0.124)	0.047	vs metric: chi-square diff 650.7, df 24, p < .001
Institutional 2025, pairwise-available missing	configural	1577.1	84	0.987	0.980	0.134 (0.128, 0.139)	0.046	
Institutional 2025, pairwise-available missing	thresholds	1883.5	119	0.985	0.984	0.122 (0.117, 0.127)	0.046	
Institutional 2025, pairwise-available missing	metric	1778.2	149	0.986	0.988	0.105 (0.100, 0.109)	0.048	

Battery	Step	Scaled chi-square	df	CFI	TLI	RMSEA (90 percent CI)	SRMR	Difference test
Institutional 2025, pairwise-available missing	scalar	4678.7	179	0.960	0.972	0.159 (0.155, 0.163)	0.048	

Per-economy configural fit and standardised loadings (single-group WLSMV models). Gate: CFI at least .95 and SRMR at most .08; Kosovo* fails it on the 2023 institutional battery and on the cooperation battery.

Battery	Economy	n	CFI	RMSEA	SRMR	Standardised loadings
Institutional 2025	ALB	902	0.987	0.141	0.040	0.72, 0.83, 0.87, 0.91, 0.93, 0.79, 0.85
Institutional 2025	BIH	935	0.985	0.149	0.055	0.51, 0.73, 0.93, 0.88, 0.96, 0.86, 0.83
Institutional 2025	XKX	936	0.961	0.129	0.066	0.41, 0.62, 0.76, 0.84, 0.85, 0.57, 0.70
Institutional 2025	MNE	824	0.980	0.156	0.041	0.73, 0.77, 0.90, 0.85, 0.94, 0.81, 0.79
Institutional 2025	MKD	930	0.990	0.113	0.041	0.66, 0.66, 0.89, 0.86, 0.95, 0.73, 0.80
Institutional 2025	SRB	880	0.993	0.140	0.030	0.84, 0.87, 0.95, 0.88, 0.97, 0.88, 0.78

Battery	Economy	n	CFI	RMSEA	SRMR	Standardised loadings
Institutional 2024, residual fixed	ALB	900	0.934	0.318	0.103	0.74, 0.86, 0.93, 0.89, 0.77, 0.82, 0.83
Institutional 2024, residual fixed	BIH	944	0.926	0.253	0.097	0.82, 0.87, 0.88, 0.79, 0.76, 0.80, 0.83
Institutional 2024, residual fixed	XKX	922	0.945	0.325	0.168	0.80, 0.94, 0.84, 0.83, 0.71, 0.93, 0.92
Institutional 2024, residual fixed	MNE	871	0.989	0.204	0.030	0.91, 0.96, 0.96, 0.90, 0.87, 0.89, 0.89
Institutional 2024, residual fixed	MKD	853	0.963	0.207	0.080	0.79, 0.95, 0.91, 0.72, 0.83, 0.80, 0.83
Institutional 2024, residual fixed	SRB	868	0.977	0.203	0.066	0.79, 0.92, 0.93, 0.86, 0.81, 0.83, 0.85
Institutional 2024, residual freed	ALB	900	0.980	0.183	0.044	0.77, 0.88, 0.95, 0.91, 0.78, 0.63, 0.66
Institutional 2024, residual freed	BIH	944	0.985	0.118	0.038	0.85, 0.89, 0.90, 0.82, 0.78, 0.63, 0.66
Institutional 2024, residual freed	XKX	922	0.991	0.138	0.056	0.83, 0.95, 0.87, 0.86, 0.74, 0.62, 0.59

Battery	Economy	n	CFI	RMSEA	SRMR	Standardised loadings
Institutional 2024, residual freed	MNE	871	0.997	0.117	0.015	0.91, 0.96, 0.96, 0.91, 0.87, 0.84, 0.85
Institutional 2024, residual freed	MKD	853	0.993	0.092	0.033	0.81, 0.96, 0.92, 0.74, 0.84, 0.64, 0.71
Institutional 2024, residual freed	SRB	868	0.990	0.137	0.038	0.81, 0.93, 0.94, 0.87, 0.82, 0.73, 0.77
Institutional 2023	ALB	863	0.984	0.137	0.040	0.80, 0.88, 0.94, 0.88, 0.69, 0.67, 0.80
Institutional 2023	BIH	951	0.975	0.134	0.049	0.81, 0.86, 0.90, 0.83, 0.76, 0.54, 0.72
Institutional 2023	XKX	947	0.534	0.268	0.143	0.51, 0.71, 0.59, 0.58, 0.56, 0.46, 0.51
Institutional 2023	MNE	912	0.960	0.195	0.065	0.80, 0.88, 0.89, 0.73, 0.62, 0.79, 0.84
Institutional 2023	MKD	892	0.961	0.200	0.070	0.75, 0.92, 0.91, 0.61, 0.78, 0.82, 0.85
Institutional 2023	SRB	885	0.971	0.216	0.070	0.74, 0.92, 0.92, 0.89, 0.84, 0.83, 0.86
Cooperation 2025	ALB	974	0.996	0.183	0.023	0.85, 0.93, 0.92, 0.88

Battery	Economy	n	CFI	RMSEA	SRMR	Standardised loadings
Cooperation 2025	BIH	973	0.999	0.080	0.010	0.80, 0.79, 0.89, 0.85
Cooperation 2025	XKX	940	0.658	0.199	0.077	0.17, 0.16, 0.77, 0.45
Cooperation 2025	MNE	905	1.000	0.033	0.006	0.74, 0.79, 0.84, 0.86
Cooperation 2025	MKD	985	0.999	0.180	0.011	0.93, 0.94, 0.96, 0.96
Cooperation 2025	SRB	940	0.999	0.135	0.009	0.87, 0.93, 0.94, 0.90

Latent means of the 2025 institutional battery under five models (Albania = 0; standard errors in parentheses; alignment standard errors from 200 within-economy bootstrap resamples; bootstrap 95 percent rank interval for the aligned means). Partial (cells) frees the police thresholds in Kosovo* and the courts thresholds in North Macedonia; partial (anchors) frees both items' thresholds in every economy.

Economy	Raw mean (1 to 4)	Full scalar	Partial (cells)	Partial (anchors)	Aligned (tuned)	Aligned (defaults)	Aligned rank interval	Two-factor: elected	Two-factor: police and courts
ALB	2.24	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00	1 to 3	0.00 (0.00)	0.00 (0.00)
BIH	1.75	-0.69 (0.05)	-0.70 (0.05)	-0.64 (0.05)	-0.77 (0.07)	-0.74	6 to 6	-0.72 (0.06)	-0.60 (0.06)
XKX	2.49	0.28 (0.04)	0.20 (0.04)	0.20 (0.04)	0.05 (0.06)	0.05	1 to 3	0.20 (0.04)	0.71 (0.06)
MNE	2.08	-0.25 (0.05)	-0.24 (0.05)	-0.19 (0.05)	-0.17 (0.08)	-0.20	4 to 4	-0.23 (0.06)	-0.29 (0.06)
MKD	2.22	-0.02 (0.05)	0.11 (0.05)	0.17 (0.05)	0.12 (0.08)	0.12	1 to 2	0.15 (0.05)	-0.62 (0.06)
SRB	1.90	-0.65 (0.07)	-0.64 (0.07)	-0.45 (0.06)	-0.43 (0.08)	-0.49	5 to 5	-0.55 (0.07)	-0.80 (0.07)

Alignment of the 2025 institutional battery: R-squared .992 for loadings and .992 for intercepts; alignment settings scale (.2, .4) and power (.25, .25), with the same ordering under the package defaults. Mean absolute deviation of aligned intercepts from the item mean, by item: police .25, courts .20, government .03, parties .04, parliament .05, civil service .11, elections .10; by economy: Albania .06, Bosnia and Herzegovina .06, Kosovo* .24, Montenegro .09, North Macedonia .13, Serbia .09 (descriptive; the two cells driving them are the Kosovo* police intercept, 3.44 against 2.41 to 2.69, and the North Macedonia courts intercept, 1.69 against 2.17 to 2.53).

DIF by ordinal logistic regression with economy as the grouping variable (lordif; six-group test; iterative anchor purification, seven iterations; anchors government, parties, civil service, elections). McFadden pseudo-R-squared change, uniform (model 1 vs 2) and total (1 vs 3): Q33.7 0.084 / 0.092; Q33.8 0.067 / 0.068; Q33.9 0.007 / 0.011; Q33.10 0.003 / 0.004; Q33.11 0.022 / 0.026; Q33.12 0.017 / 0.019; Q33.14 0.008 / 0.009. Flagged at .02: police, courts, parliament; at .035: police, courts. Uniform DIF dominates in every flagged item (non-uniform change at most .008).

Eight economy-by-nationality groups, 2023 (Kosovo* Albanians excluded): aligned means BIH-Serbian 0.00, BIH-Bosniak -0.36, MNE-Serbian -0.02, MNE-Montenegrin -0.10, SRB-Serbian 0.29, ALB-Albanian -0.02, MKD-Macedonian -0.53, MKD-Albanian -0.12; alignment R-squared loadings 0.986, intercepts 0.998; sequence CFI configural 0.976, metric change -0.0012, scalar change -0.0085. Nine groups including Kosovo* Albanians: scalar change -0.0155.

Pairwise mean absolute difference between groups' aligned intercepts, 2023 institutional battery, eight nationality groups (Kosovo* Albanians excluded; the nine-group matrix including them is in the package).

	BIH-Serbian	**BIH-Bosniak**	**MNE-Serbian**	**MNE-Montenegrin**	**SRB-Serbian**	**ALB-Albanian**	**MKD-Macedonian**	**MKD-Albanian**
BIH-Serbian		0.08	0.09	0.05	0.11	0.17	0.14	0.10
BIH-Bosniak	0.08		0.14	0.07	0.07	0.20	0.12	0.13
MNE-Serbian	0.09	0.14		0.08	0.20	0.14	0.13	0.05
MNE-Montenegrin	0.05	0.07	0.08		0.12	0.14	0.09	0.08
SRB-Serbian	0.11	0.07	0.20	0.12		0.20	0.16	0.20
ALB-Albanian	0.17	0.20	0.14	0.14	0.20		0.09	0.12
MKD-Macedonian	0.14	0.12	0.13	0.09	0.16	0.09		0.12
MKD-Albanian	0.10	0.13	0.05	0.08	0.20	0.12	0.12	

Appendix C. Transparency checklist evidence

Element	Reference programmes	2023	2024	2025	Evidence
Sponsor and data-collection organisation named	AAPOR 2; EB; ESS	2	2	2	2023: 'Indago-led consortium of research companies' (p. 102). 2024: 'Data Centrum Research Institute' (p. 111). 2025: 'CeSID Ltd, Leader of Consortium with IDRA Research and Consulting' (about page). Sponsor RCC throughout.
Population under study defined	AAPOR 4; EB; ESS	2	2	1	2023: residents 18+, private households, able to speak local language (p. 102). 2024: 18+ residents (p. 112). 2025: 'respondents aged 18+' only.
Sampling frame and design described	AAPOR 5; EB; ESS (per-country design summaries)	2	1	1	2023: stratified two-stage cluster, PSU = enumeration districts, PPS, random route, last birthday, about 10 per sampling point, regional allocation table (pp. 102-104). 2024: multistage stratified cluster, random route every third household, nearest birthday, 5-10 per point (p. 112); no allocation table. 2025: one sentence ('stratified according to the region or economy first', maximum 10 per sampling point).
Number of completed interviews per economy	AAPOR 8; EB; ESS	2	2	2	Printed in each edition and shown in the database (2023: 1,000 to 1,010; 2024: 1,000 each; 2025: 1,000 to 1,008).
Fieldwork dates	AAPOR 7; EB (per country); ESS (per country)	2	2	1	2023: 3 March to 4 April 2023. 2024: 5 March to 4 April 2024. 2025: 21 May to 30 June 2025 on the about page; the released file's interview dates run 3 to 30 June 2025. No per-economy dates in any edition.
Mode of data collection	AAPOR 6; EB; ESS	2	2	2	CAPI face-to-face household interviews in all three editions.
Full questionnaire published	AAPOR 3; EB (questionnaire published with the wave); ESS (source and national questionnaires)	1	1	1	No questionnaire document is published; wording is recoverable item by item from the report figures, the database headers and the SPSS variable labels. Show-cards, filters, interviewer instructions and the translated versions are not published.

Element	Reference programmes	2023	2024	2025	Evidence
Translation protocol and language versions	ESS (TRAPD, Translation Assessment); EB (national questionnaires)	1	1	1	2023: 'translated into related local languages by Indago. RCC reviewed and approved all the translations' (p. 102); languages not named. 2024: languages named for Kosovo* (Albanian and Serbian) and North Macedonia (Macedonian and Albanian) (p. 111); no method. 2025: 'translated into related local languages'. No edition describes the translation method, back-translation, adjudication or pretesting of translations, and no file records the language of interview.
Response rate or contact outcomes	AAPOR 8 (precision), AAPOR outcome rates; EB (not routinely); ESS (per country, AAPOR-defined)	0	1	0	2024: 'Response Rate: 22% on average' (p. 111), without definition, per-economy value or contact outcomes. Not reported in 2023 or 2025.
Weighting described	AAPOR 9; EB (post-stratification and population weights); ESS (design, post-stratification, population weights)	0	0	1	2023: the released file carries a 'Weight' variable (range 0.73 to 1.46) that no published document mentions. 2024 and 2025: no weight variable in the file; 2025 key findings: 'No weighting was applied' for the WB6 aggregate. No edition states how, or whether, economy-level estimates are weighted to population benchmarks.
Precision: margin of error or design effect	AAPOR 8; EB (confidence-interval table by sample size); ESS (design effects per country)	0	0	0	No margin of error, confidence interval, design effect or standard error is published in any edition; economy differences and year-on-year changes of one or two points are reported as findings.
Data processing and quality control	AAPOR 10; ESS (fieldwork monitoring, back-checks)	1	1	0	2023: interviewer training and instructions (p. 102). 2024: training, pilot of up to 50 interviews, interview length 20-25 minutes (p. 111). No back-checks, validation rates or data-cleaning rules in any edition.

Element	Reference programmes	2023	2024	2025	Evidence
Comparability of the series documented	ESS (source questionnaire changes documented per round); EB (trend questions kept verbatim, breaks flagged)	0	0	0	The questionnaire had 77, 43 and 39 questions in 2023, 2024 and 2025, fieldwork moved from March-April to June, and the contractor changed each year; no edition lists which items changed wording, scale or base, and the database presents multi-edition series without a break marker (it does separate tables where the RCC treats a wording as a new series).
Limitations statement	AAPOR 11	0	0	0	None found.
Microdata access with documentation	ESS (open microdata, codebook, documentation report); EB (via GESIS with codebook)	1	1	1	SPSS files for 2017 to 2025 are downloadable from the results page (a modal labelled 'Balkan Barometer Micro Data'). No codebook, no weight or design variables (except the undocumented 2023 weight), no language-of-interview, PSU or interviewer identifiers, and no terms of use. The 2025 file stores don't-know as system-missing on many items and the two 'situation' items carry 13 to 34 percent missing values with no 'very good' responses.
Regional aggregate: composition and weighting stated	EB (EU averages weighted by population, stated in every report)	0	0	2	2025 key findings state the equal-weight, unweighted aggregation. The 2023 and 2024 reports do not say how the 'Western Balkans' figure is formed. The database's hidden 'share' column shows population-based shares up to 2019 and equal shares from 2020, an undocumented change of estimand.

Sources scored: Balkan Barometer 2023 Public Opinion (PDF, 107 pp.), 'Note on methodology' pp. 102-105; interactive database; microdata file 1_BB2023_PO_database_v2.0_19042023.sav; Balkan Barometer 2024 Public Opinion (PDF, 121 pp.), 'Note on methodology' pp. 111-116; interactive database; microdata file 3. RCC BB24_PO - SPSS Update 04.07.24.sav; rcc.int/balkanbarometer/about (methodology paragraph), key_findings_2025/2/public, interactive database, infographics; microdata file RCC_2025_PO_Data_AllCountries_F.sav; the full 2025 report was 'Coming soon' on the publications page on 2026-09-18.