Appendices

A Do country experts consistently update their scores in a more "pessimistic direction"? Analyzing changes from V–Dem v.5 and v.9 to v.13

As discussed in Section 3.2 of the paper, if the bias that L&M propose is powerful, some of V–Dem's country experts should reevaluate their perceptions of trends in democracy and therefore revise their earlier ratings to reflect lower levels of democracy in the countries they rate during the period since about 2009. V–Dem's data collection procedures make it easy for the experts to do so. In each annual update, all experts are shown their previous ratings for earlier years. To change these ratings, all they have to do is select a different rating for any of these years and submit it along with their new ratings. We can check the degree to which they do so by comparing the coder-level data from earlier waves of data collection to the most recent wave.

Analyses reveal that V–Dem country experts have not systematically revised their ratings downward for recent years. First, experts have changed less than four percent of their ratings on any of the component variables from the Electoral Democracy Index between v5 (released in 2016) and v13 (released in 2023) and less than 1.4 percent of their ratings between v9 (released in 2019) and v13.[1] The mean change in all ratings, including the unrevised ratings, is statistically indistinguishable from zero for all 20 variables in all years from 1900 to 2012 (the last year of ratings in v5) for the v5 to v13 comparison and for all 45 variables in all years from 1900 to 2019 in the comparison of v9 to v13.

Country experts may not revise scores because it takes extra time, thought, and effort. It may also be psychologically unpleasant to correct a "mistake." With this caveat, among the country experts who did revise their ratings, there has been no consistent trend toward assigning lower ratings in recent years. Appendix Figure A.1–Figure A.3 report patterns for the v5 to v13 analysis (grouped according to whether and in what direction the indicator's revision has led to systematic changes, after year 2000). Figures 4 in the paper as well as Appendix Figures A.4-A.8 do the same for the v9 to v13 comparison. Consdiering the latter comparison, for 18 of the 45 variables, there has been no significant trend in either direction since 2000, although in some cases there were significant trends in some earlier periods (Figure A.1). There was a significant negative trend after 2000 for eleven variables, although with three exceptions (Election Other Voting Irregularities, Opposition Parties Autonomy, and Executive Respects Constitution) the negative revisions began in the 1980s or earlier, which is inconsistent with the DGP proposed by L&M (Appendix Figure A.2). The remaining fifteen variables show positive revisions since 2000 (Appendix Figure A.3).

[1]The v5 to v13 analysis includes 20 expert-coded variables used in the Electoral Democracy Index. It excludes the 18 variables in the Elected Officials Index and Suffrage because they were centrally coded; CSO Repression because its scale was flipped in 2014; and Barriers to Parties, which cannot be analyzed in this way. The v9 to v13 analysis includes 45 variables used in either the Electoral Democracy Index or the Liberal Component Index.

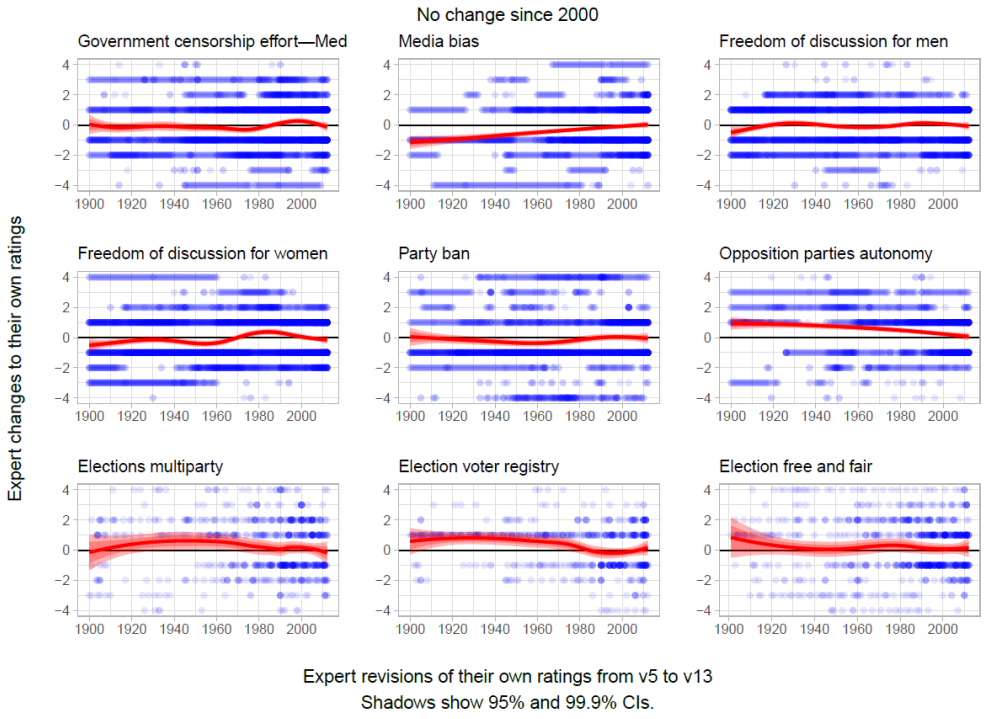

Figure A.1: Expert-coded indicators entering V–Dem's Electoral Democracy index that have not experienced any systematic change after year 2000 due to coder revisions (from V–Dem v.5 to v.13).

Even after confining our attention to the ratings that have been revised, the net revisions have been few and very small on average. They have been positive for some variables, negative in others, and non-significant for more, and usually applied to earlier years rather than just the past decade. None of this is consistent with the DGP L&M's model implies, and it is not enough change in one direction at the right time to produce the observed recent decline in electoral democracy scores, especially if we take into consideration the overwhelming majority of ratings that have not been revised at all. Taking all ratings, revised or not, into consideration, revisions have not been responsible for the backsliding observed in V–Dem's Electoral Democracy Index.

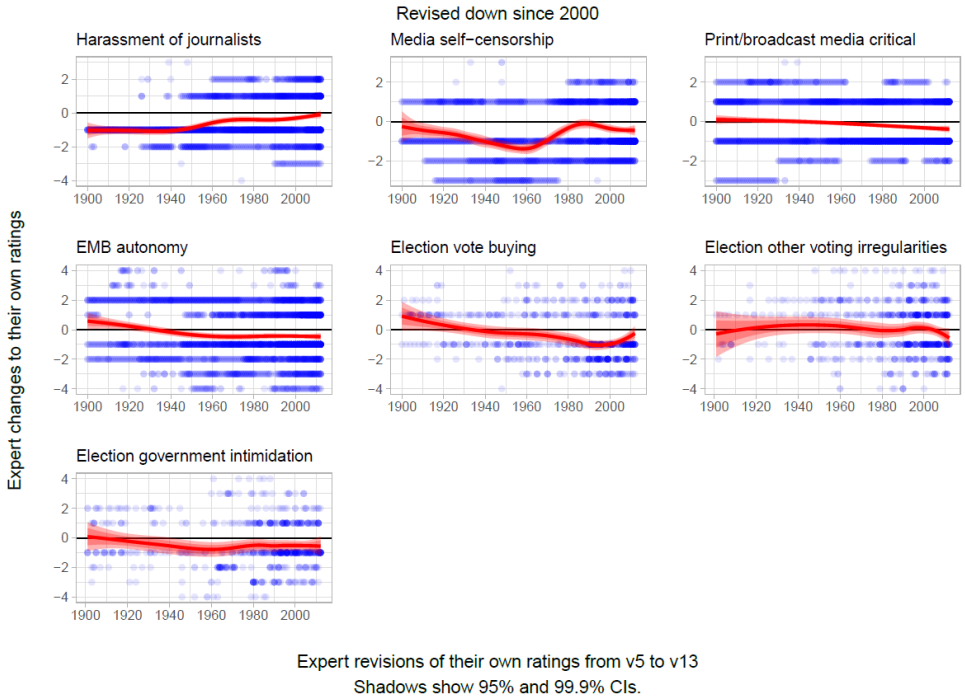

Figure A.2: Expert-coded indicators entering V–Dem's Electoral Democracy index that have have been revised down after year 2000 due to coder revisions (from V–Dem v.5 to v.13).

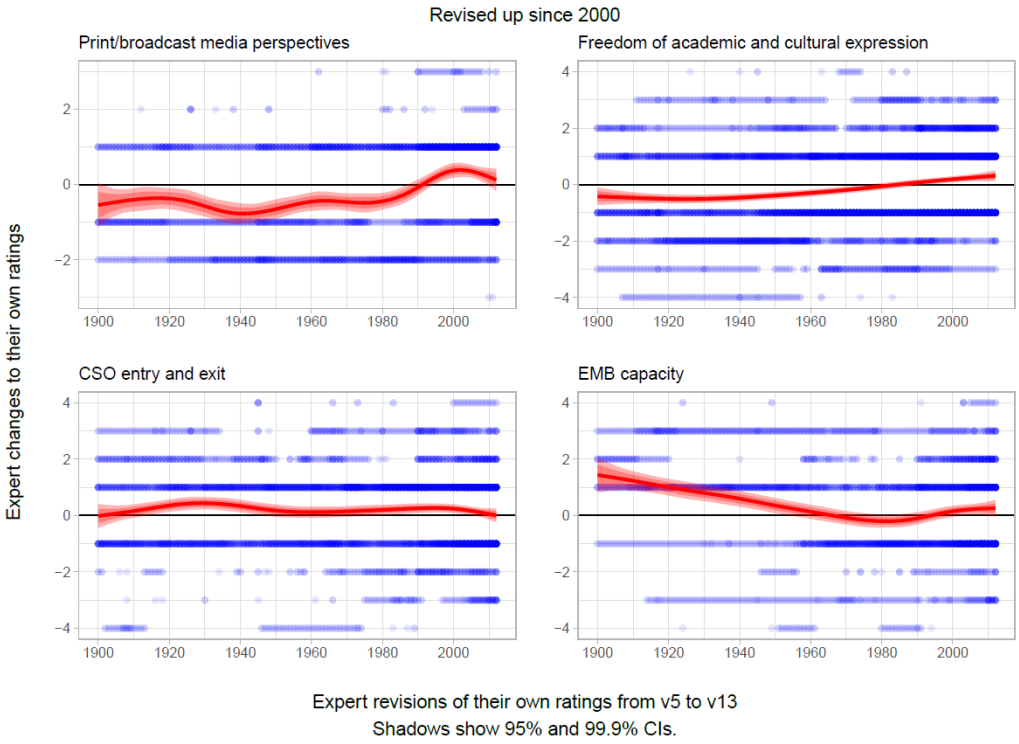

Figure A.3: Expert-coded indicators entering V–Dem's Electoral Democracy index that have have been revised up after year 2000 due to coder revisions (from V–Dem v.5 to v.13).

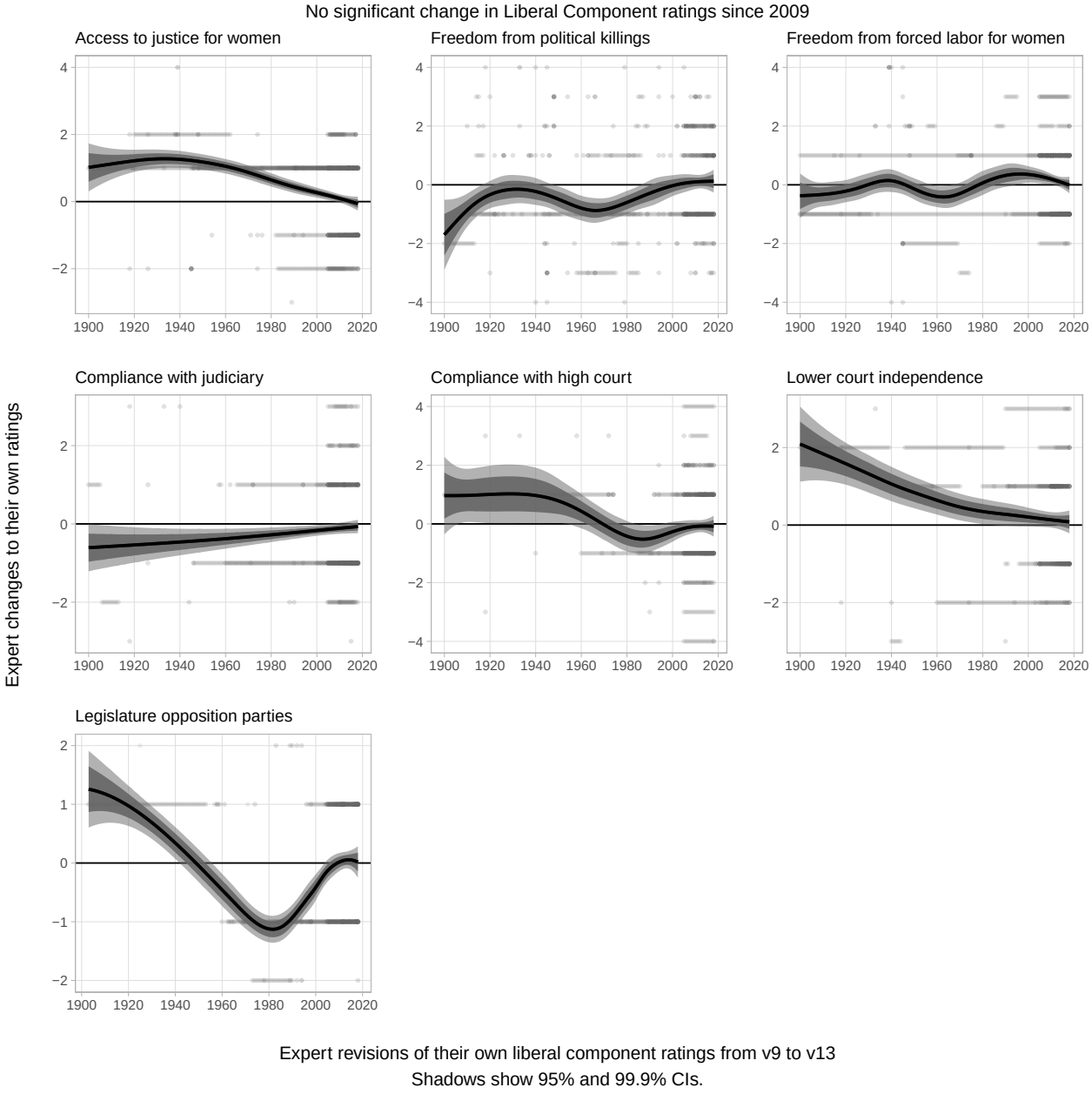

Figure A.4: Expert-coded indicators entering V–Dem's Liberal Component index that have not experienced any systematic change after year 2000 due to coder revisions (from V–Dem v.9 to v.13).

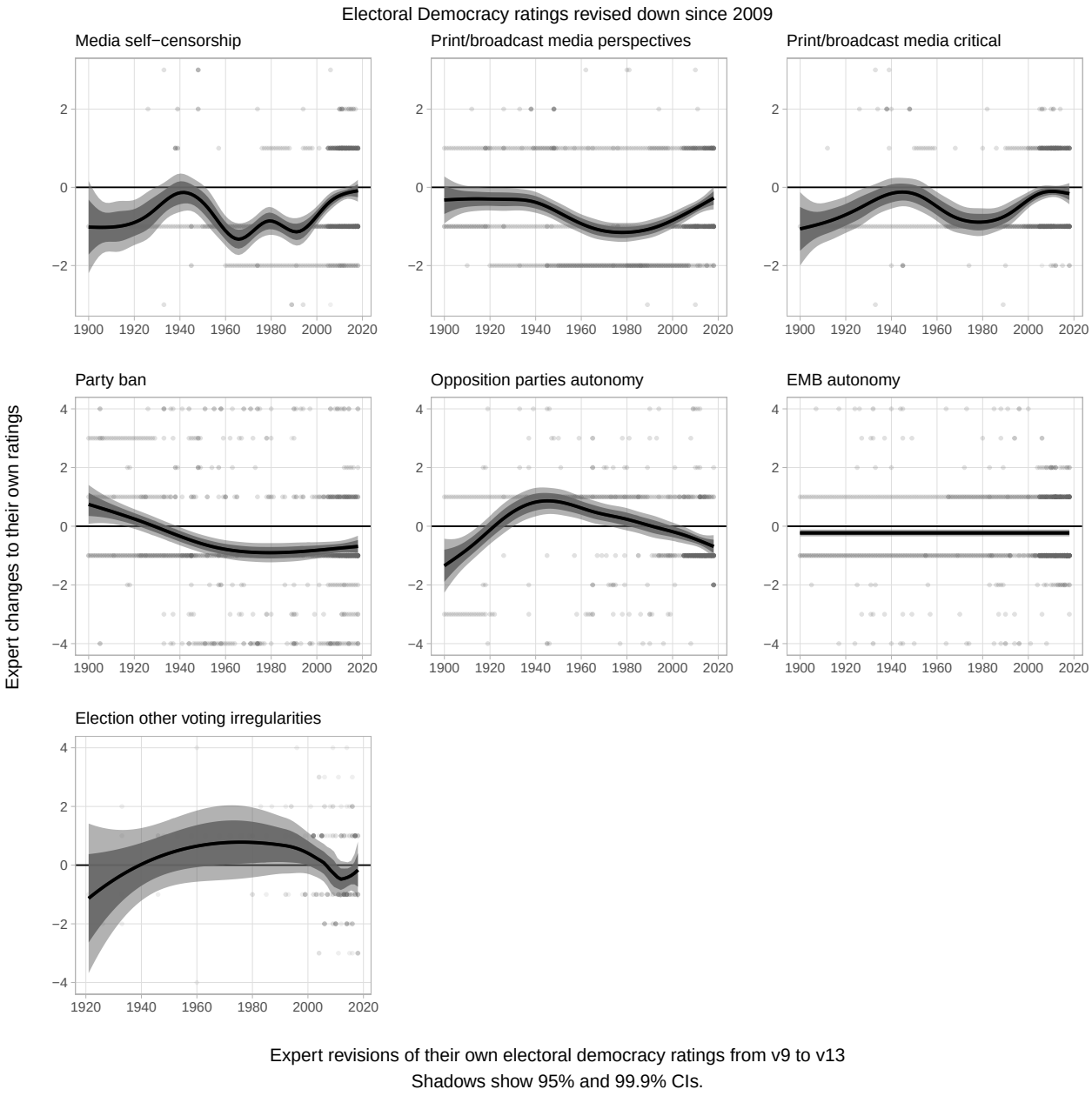

Figure A.5: Expert-coded indicators entering V–Dem's Electoral Democracy index that have have been revised down after year 2000 due to coder revisions (from V–Dem v.9 to v.13).

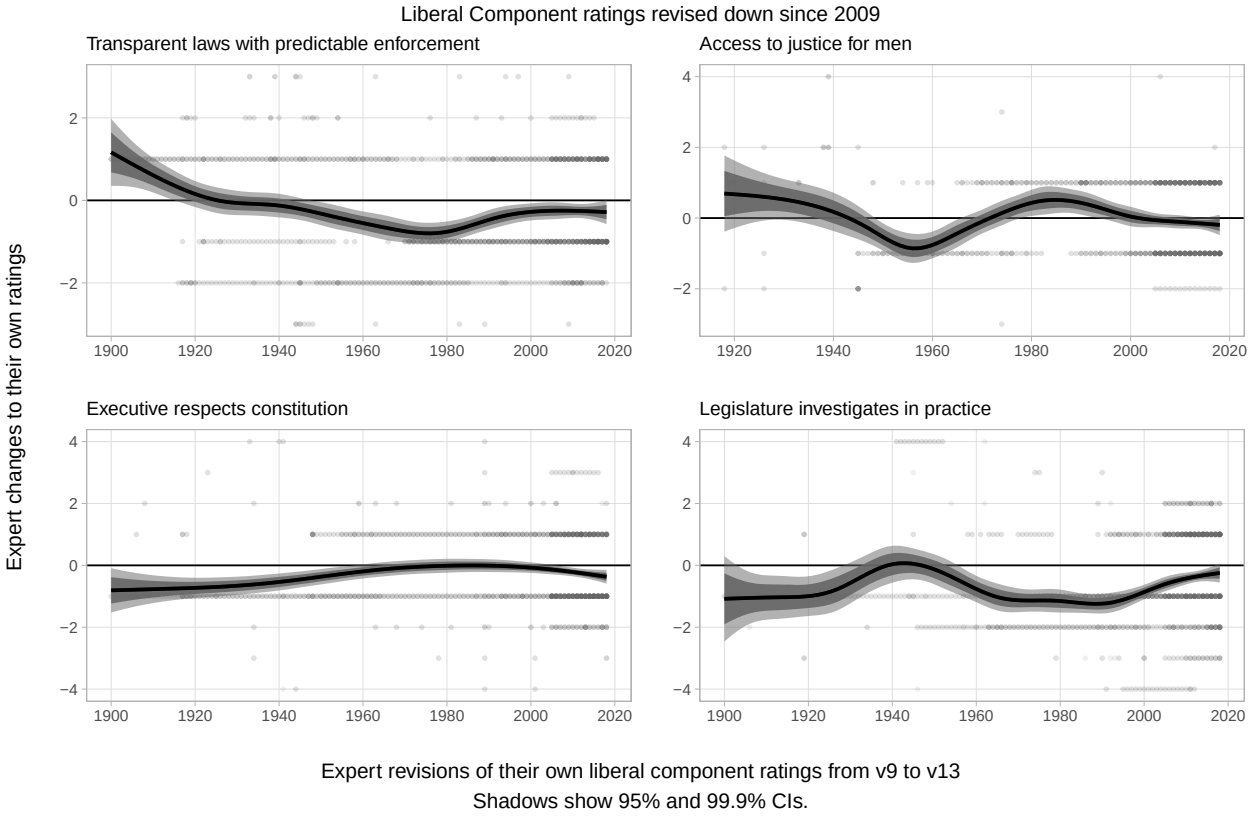

Figure A.6: Expert-coded indicators entering V–Dem's Liberal Component index that have have been revised down after year 2000 due to coder revisions (from V–Dem v.9 to v.13).

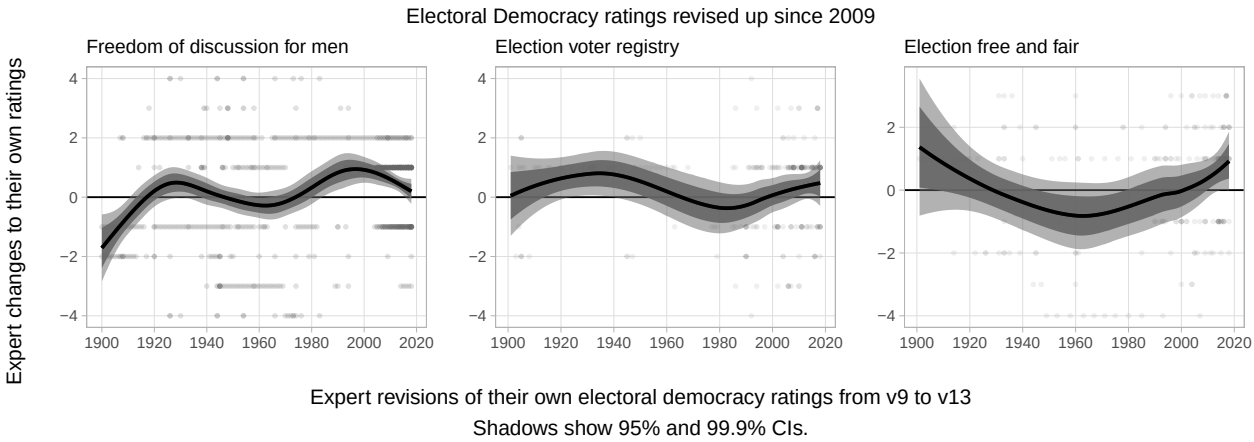

Figure A.7: Expert-coded indicators entering V–Dem's Electoral Democracy index that have have been revised up after year 2000 due to coder revisions (from V–Dem v.9 to v.13).

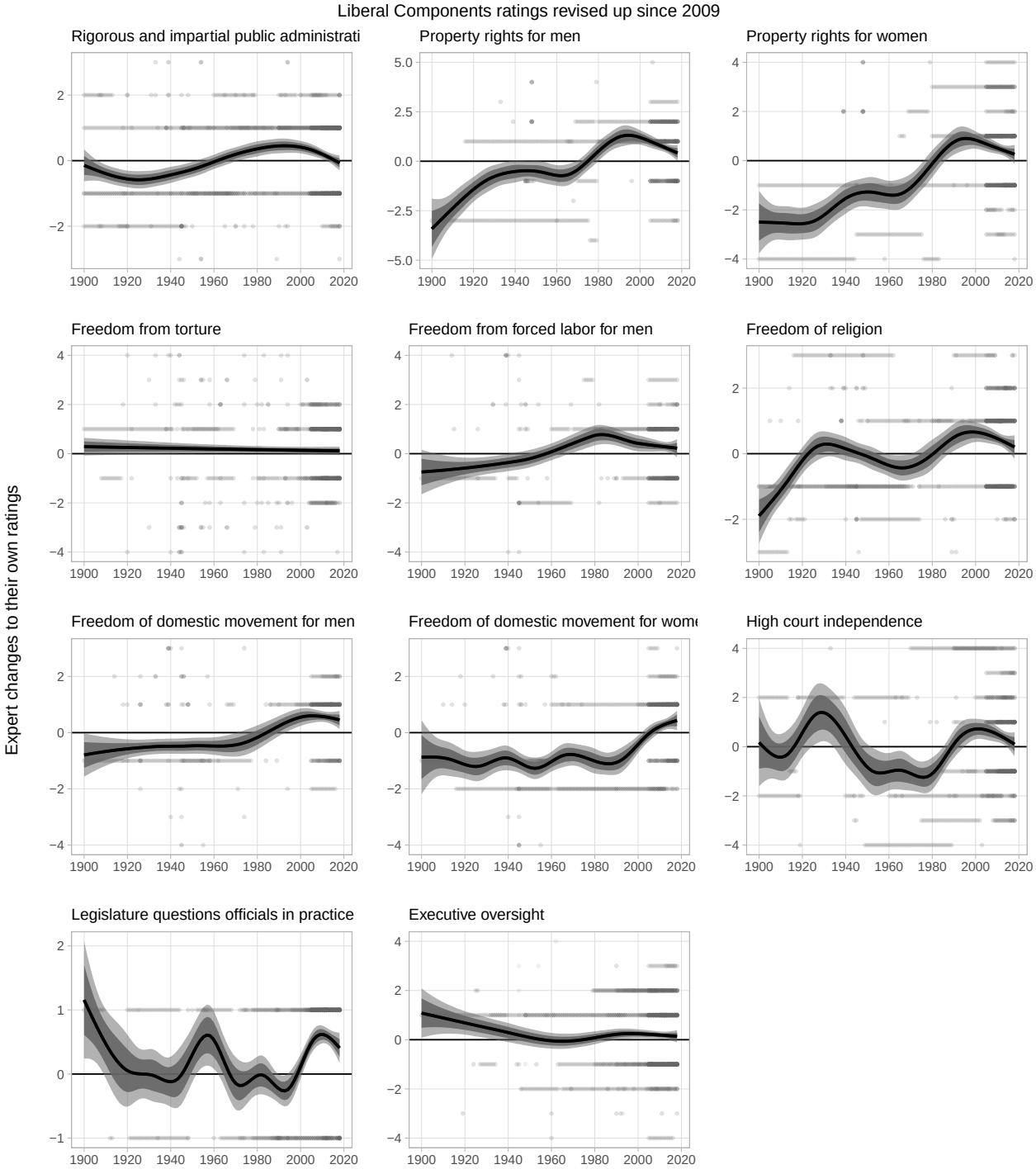

Figure A.8: Expert-coded indicators entering V–Dem's Liberal Component index that have have been revised up after year 2000 due to coder revisions (from V–Dem v.9 to v.13).

B Differences in coder disagreement between more and less subjective V–Dem indicators

As noted in Section 3.2, we assess patterns in coder disagreement for two expert-coded ("C") V–Dem indicators with different levels of subjectivity, namely election free and fair (*v2elfrefair*) and election boycotts (*v2elboycot*).

The former summary indicator, which is far less specific than the typical V–Dem indicator, asks coders to evaluate: "Taking all aspects of the pre-election period, election day, and the post-election process into account, would you consider this national election to be free and fair?" Between the multiple periods of evaluation, the broad and multidimensional nature of assessing how "free and fair" elections are, and the inherently subjective language "would you consider," this is a highly subjective question. The latter asks coders to evaluate: "In this national election, did any registered opposition candidates or parties boycott?" This question subsequently provides a clarification about the definition of a "boycott," so there is some subjectivity involved in applying the definition. Otherwise the question involves coding a categorical, observable trait pertaining to a limited set of actors. This expert-coded indicator is therefore associated with relatively low subjectivity.

For each variable, we regress coder disagreement (the standard deviation of coder ratings at the country-year level) on country and coder traits: the century of coding, level of freedom of expression, level of democracy, and number of coders. Table B.1 reveals that none of these variables predict coder disagreement for the low-subjectivity variable, *v2elboycot*. In contrast, coder disagreement is higher for the high-subjectivity variable, *v2elfrefair*, for countries and years that are more recent, have higher freedom of expression, lower levels of democracy, and more coders. Collectively, these findings suggest that coder disagreement does systematically vary with the level of democracy and level of freedom of expression, but only for the highly subjective V–Dem indicator.

While these two variables display similar average levels of disagreement (0.602 for *v2elboycot* and 0.733 for *v2elfrefair*, both on a 5-level scale), these aggregate figures belie nuanced patterns in how disagreement varies with country- and coder-level characteristics. Both (relatively) low-subjectivity and high-subjectivity questions are common also among the expert-coded indicators in the V–Dem dataset. Hence, it is critical to think through and analyze the drivers of disagreement likely to affect a particular set of questions rather than simply assuming common patterns affecting all questions similarly. These results indicate that the plausibility of L&M's hypothesized form of bias varies across indicators, and is minimal in many.

Table B.1: Predicting Respondent Disagreement

	v2elboycot	v2elfrefair
Century	0.307	0.375**
	(0.236)	(0.171)
Freedom of Expression	0.387	1.829***
	(0.537)	(0.359)
Level of Democracy	-2.688	-7.176***
	(2.256)	(2.241)
Level of Democracy × Level of Democracy	0.889	4.460
	(2.913)	(3.029)
Number of Coders	0.006	0.057*
	(0.032)	(0.031)
R^2	0.092	0.407
N. Countries	43	45
N. Observations	156	187

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. This analysis was produced with v13 of the V–Dem dataset.

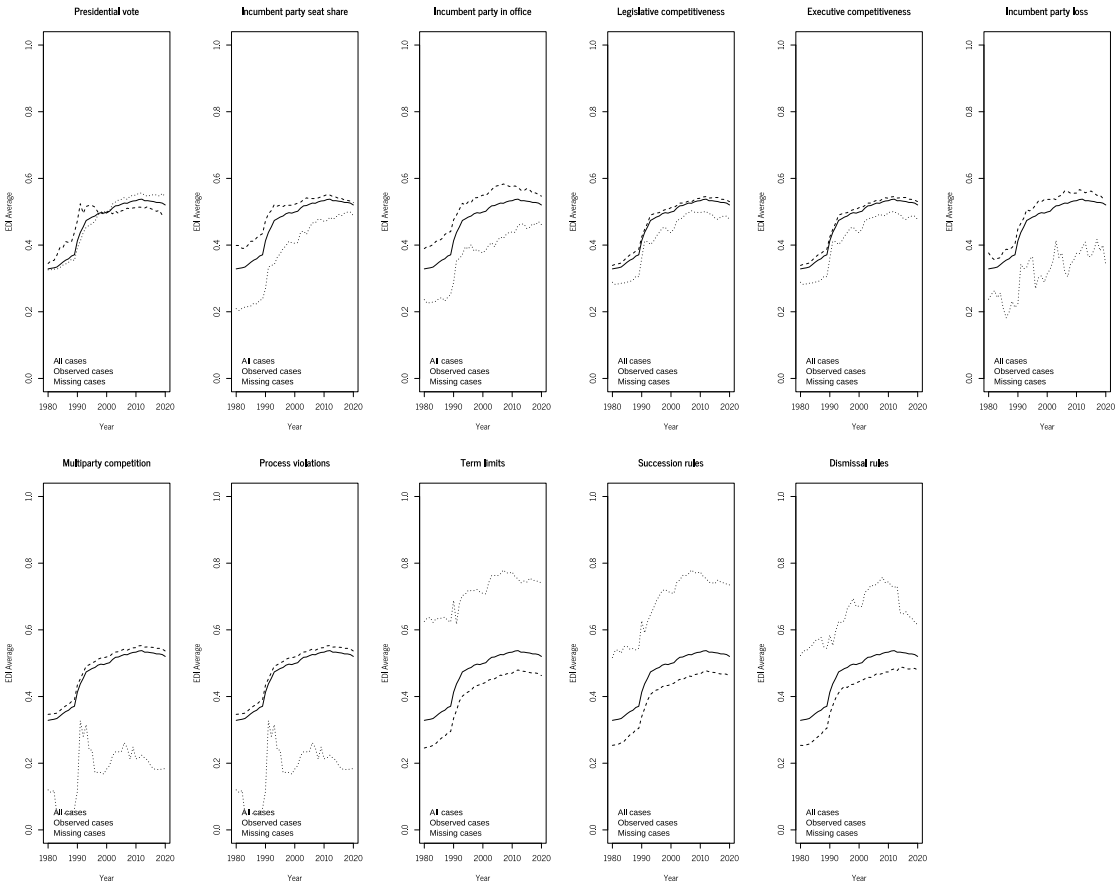

Figure C.1: V–Dem EDI world averages across L&M indicator case coverage

C Missingness in L&M data

Figure C.1 displays missingness patterns in L&M's democracy indicators. Each panel represents an L&M indicator, and plots three world average EDI scores over time: for all cases (solid line), for cases where the L&M indicator has data (dashed), for cases where the indicator is missing (dotted). For most indicators, all three lines show some backsliding over the 2010–2020 period. For seven of the 11 measures, the dotted lines show substantially more backsliding than the other two lines. Much of the recent downward action in the EDI world average is therefore driven by cases that simply do not factor into L&M's analysis. These patterns make it difficult to use trends in these individual indicators as points of reference against which to judge subjectively coded democracy measures.

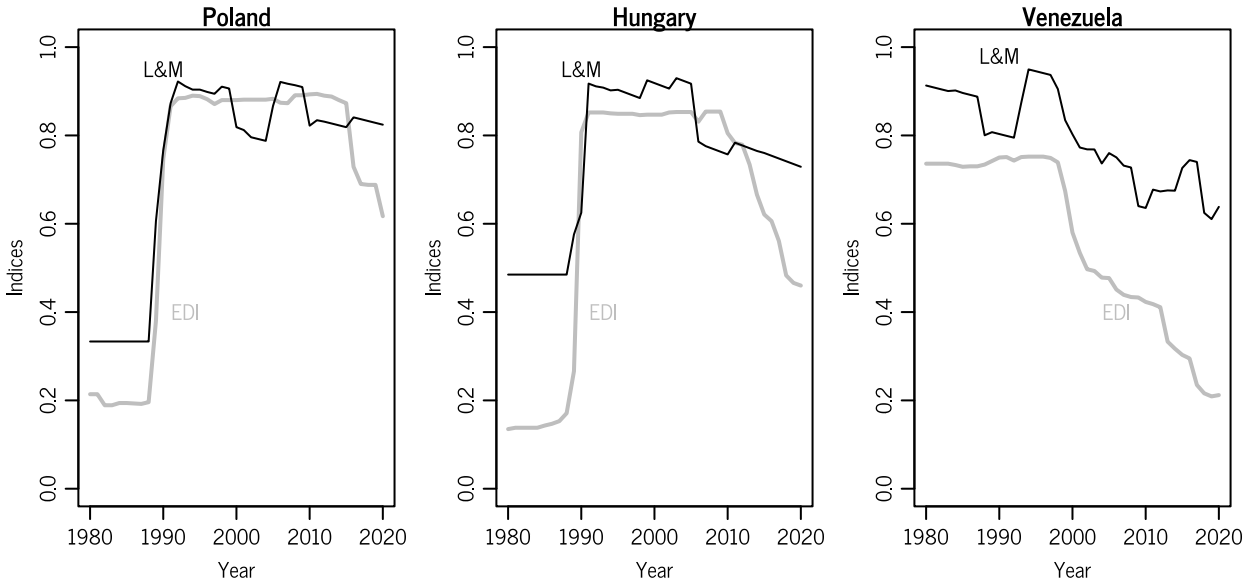

Figure D.1: L&M index validation using country trends

D Problems with using L&M's index for country-level analyses

As discussed in Section 5.3 of our paper, careful analysis of trends within countries is an important way of validating measures of democracy. Indeed, in previous versions of their manuscript, L&M validated their index using analyses of trends in levels of democracy in three widely acknowledged cases of recent democratic backsliding: Hungary, Poland and Venuezela (reproduced in Figure D.1).

Though L&M caution against using country-level analyses in the final version of their manuscript, country-level trends remain implicit in their population-weighted analyses of levels of democracy (Figure 8); even more so when they remove China and India from their analyses because these countries "happen to have noisy trends in recent years" (p. 18).

As such, we believe discussion of problems with analyzing the L&M data at the country-level provides both important insights and clear illustrations of the severity of the measurement issues discussed in our main paper, for instance pertaining to missingness and democracy thresholds. These issues have implications for the validity their analyses.

D.1 Data issues

The L&M index scales all their variables to 0-1 and takes the unweighted average of the indicators to create country-level values; missing data are treated as missing "as if" random. As discussed in Section 5, given that eight of their 12 indicators have an unclear relationship with democracy — in terms of either scaling or conceptualization — this aggregation technique is already problematic.

Although the remaining four indicators that L&M use in their analyses are conceptually closely

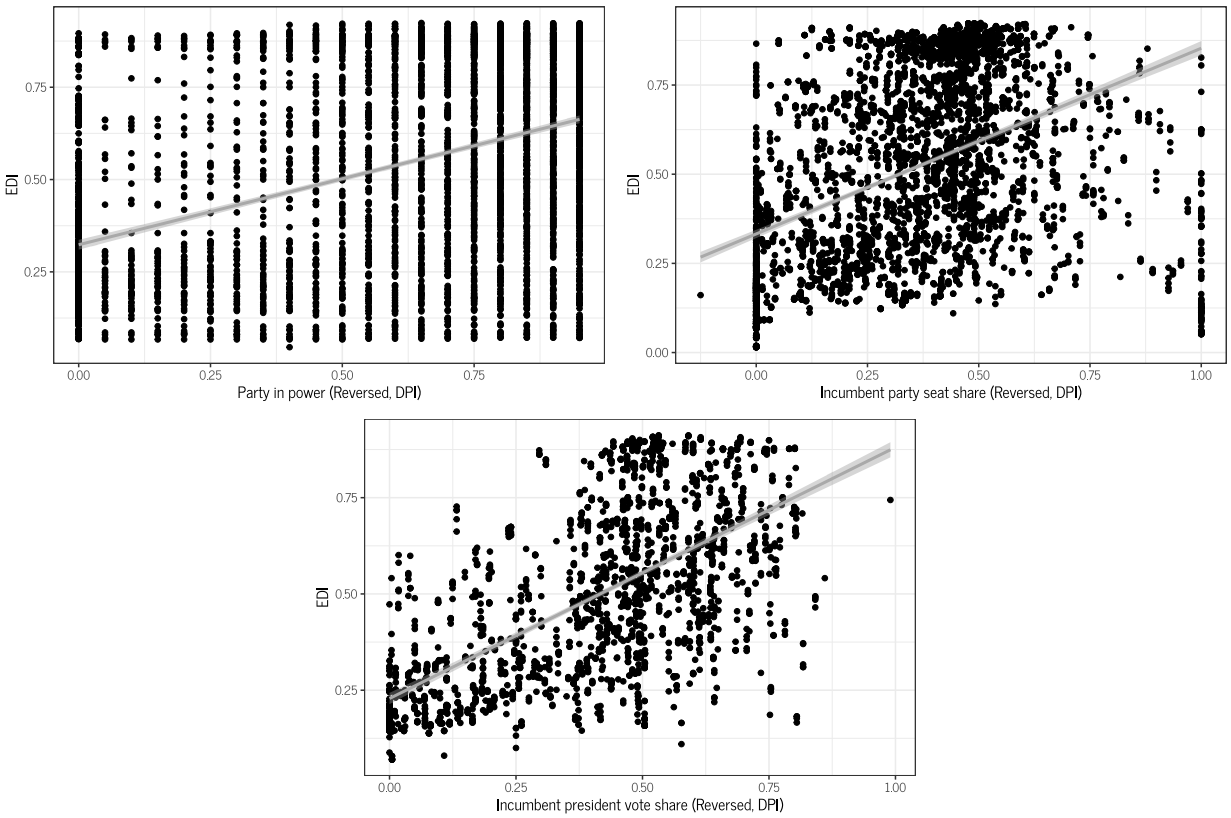

Figure D.2: Relationship between L&M's continuous indicators and EDI

related to democracy and have fewer scaling issues since they are continuous or dichotomous, they are nevertheless untenable for use in analyses of country-level data. Figure D.2 presents scatterplots of the relationship between L&M's three remaining continuous indicators of democracy and EDI. These indicators — party years in power and incumbent vote and seat share— have both conceptual and empirical issues. Conceptually, political candidates and parties in democracies — including incumbents — are supposed to pursue policies that make them popular. As a result, there is no theoretical relationship to expect a linear relationship between winning votes or seats — or staying in power — and levels of democracy. Indeed, there is clear and consistently high levels of variation in levels of democracy for all values of all three variables, indicating the the correlation is incredibly noisy.[2]

The final variable in L&M's democracy index — a dichotomous indicator of whether or not the incumbent party won the last election — epitomizes the problems with using electoral outcomes to measure democracy. Since incumbents in democratic countries are supposed to at least sometimes win elections, this variable makes no sense at the country-level as an indicator of democracy.[3] If an

[2]Also note that there is clear overlap between the vote and seat share indicators and the DPI indices; indices created using both thus violate the local independence principle for index construction.

[3]This variable make more sense in L&M's analysis of global trends since, on average, a greater proportion of incumbent

incumbent wins an election, this automatically results in a *minimum* .08 drop in a country's score on the L&M democracy index.

D.2 Additional discussions of problematic cases for L&M's democracy indicators

In Section 5 of the article, we discussed how different measurement choices and issues with L&M's democracy measure result in several cases displaying scores and trends with low face validity. We consider an additional few such cases here in this appendix.

First, consider China (Figure D.3), which was also discussed in the article text. The coding of China illustrates how L&M's weighting and missing-data imputation criteria can result in problematic results. Specifically, since L&M treat missing data as missing at random, a country's index value is the average across whatever indicators are not missing. Across all country-years (1980-2020), all indicators save four (suffrage, term limits, succession rules and dismissal rules) are missing for China. Across 1982-2017, L&M's index therefore reports China as a perfect democracy because it scores 1 on all four indicators with data. Similarly, highly autocratic Turkmenistan – displayed in Figure D.3 – has very high scores on L&M's measure, much as North Korea. It has an average value of .56, and a minimum value of .44 across all observations. Given patterns of missingness, the fact that Turkmenistan has both succession and dismissal rules — as well as *de jure* universal suffrage, consistently high DPI legislative competition scores, and L&M process violation scores — essentially establishes this relatively high score as the minimum value for the country regardless of other democratic violations. Finally, Equatorial Guinea has a high score of .61 due to a combination of missing data for 2017-2020 (six indicators) and high scores in suffrage, succession rules and the NELDA indices included by L&M. Such high scores for autocracies illustrates that the L&M index weighting criteria often result in high scores for countries that achieve only basic democratic thresholds. In this context, substantial backsliding would likely require either fundamental constitutional changes or a high degree of missingness in key indicators.

Second, Norway, which is also included in Figure D.3, illustrates important issues with L&M's codings of relatively stable democracies. Unlike the United States, Turkmenistan and China (and 28 other countries, see Figure D.4) Norway never becomes a perfect democracy in the L&M index. This is because the data for Norway include more continuous indicators than the US (the government share of seats and party years in power). Since these values never reach 1, Norway always scores below 1 on the overall index. Similar to the US, Denmark and Brazil; the largest source of variation in Norway's scores is the indicator of whether or not the incumbent party lost the last election: whenever the incumbent wins, Norway loses 12 points on the index until they lose an election, at which point the 12 points return. This latter indicator actually explains all the variation in the United States L&M scores. The

losses across countries is evidence of greater electoral competition.

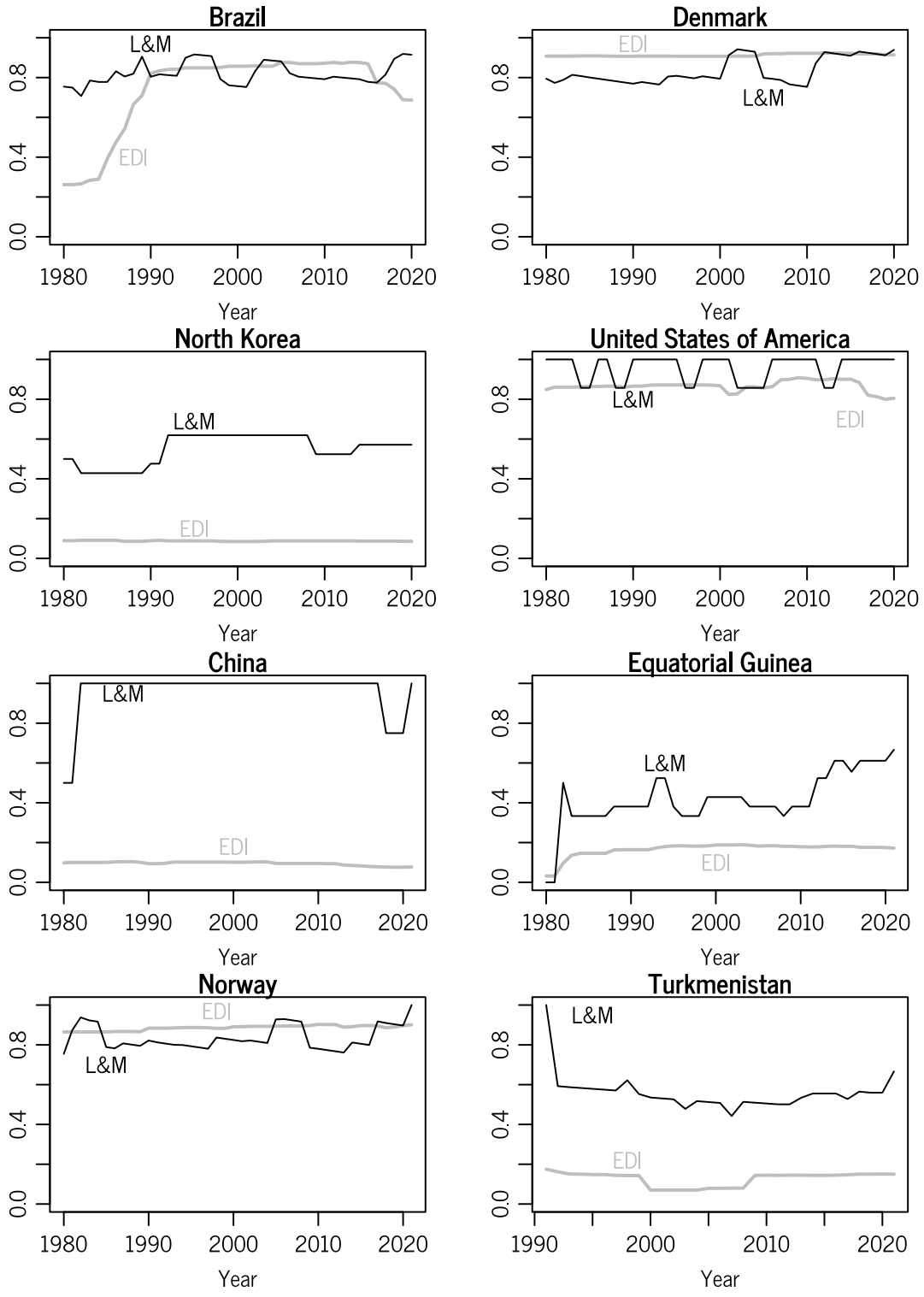

Figure D.3: A selection of cases with substantively different assessments under EDI and L&M.

US only includes 7 indicators, all of which universally have a perfect score of 1, except for the indicator of incumbent party losing an election. Because of this variable, the US becomes less democratic every year the incumbent party wins (a score of .86) before returning to perfect democracy when it loses.

D.3 Assessing recent backsliding across the world: 10-year differences in V–Dem's EDI and L&M index, 2010-2020

These issues with country-level data also make it difficult to assess the source of divergences between the L&M and EDI indices. Figure D.5 in the paper plots the 10-year difference from 2010 to 2020 on L&M's democracy index on the vertical and changes during the same interval for V–Dem's EDI on the horizontal axis, for each country in the world with data. (2020 is chosen because it is the last year in which L&M is estimated). The marked countries are the ones where their estimated degree of change is +.18 higher on L&M than EDI. +.18 is arbitrarily chosen, with the aim being to identify notably diverging countries. One interpretation of the figure is as follows: a primary reason why L&M arrive at the conclusion that there was no global democratic erosion between 2010 and 2020 is that their measure (a) shows small or no change in countries such as Hungary, India, Poland, Turkey and Venezuela, which are clearly "backsliding" countries according to EDI and other widely used measures of democracy. The diverging conclusions from using different measures also partly stems from the fact that L&M measure notable, positive democratic change in countries such as Bolivia, Yemen, Brazil, Egypt, Sudan, Equatorial Guinea and Denmark. These are countries where V–Dem's EDI either identifies backsliding or (as in the case of, e.g., stable Denmark) no substantial change at all. The combination of these two divergences that makes the global aggregate scores on L&M's index paint a more optimistic picture than EDI.

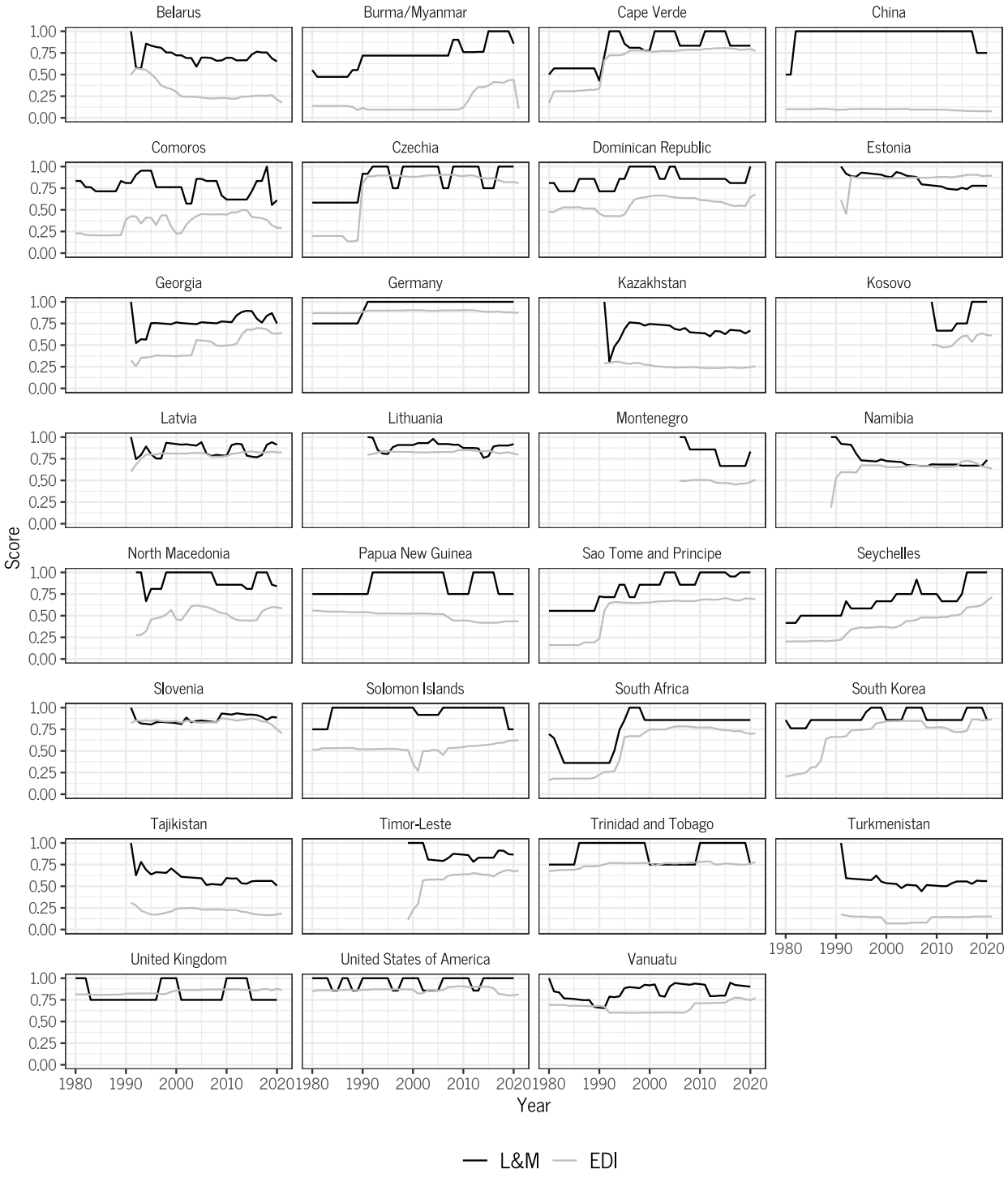

Figure D.4: Countries that earn a perfect score for at least one year on the L&M index

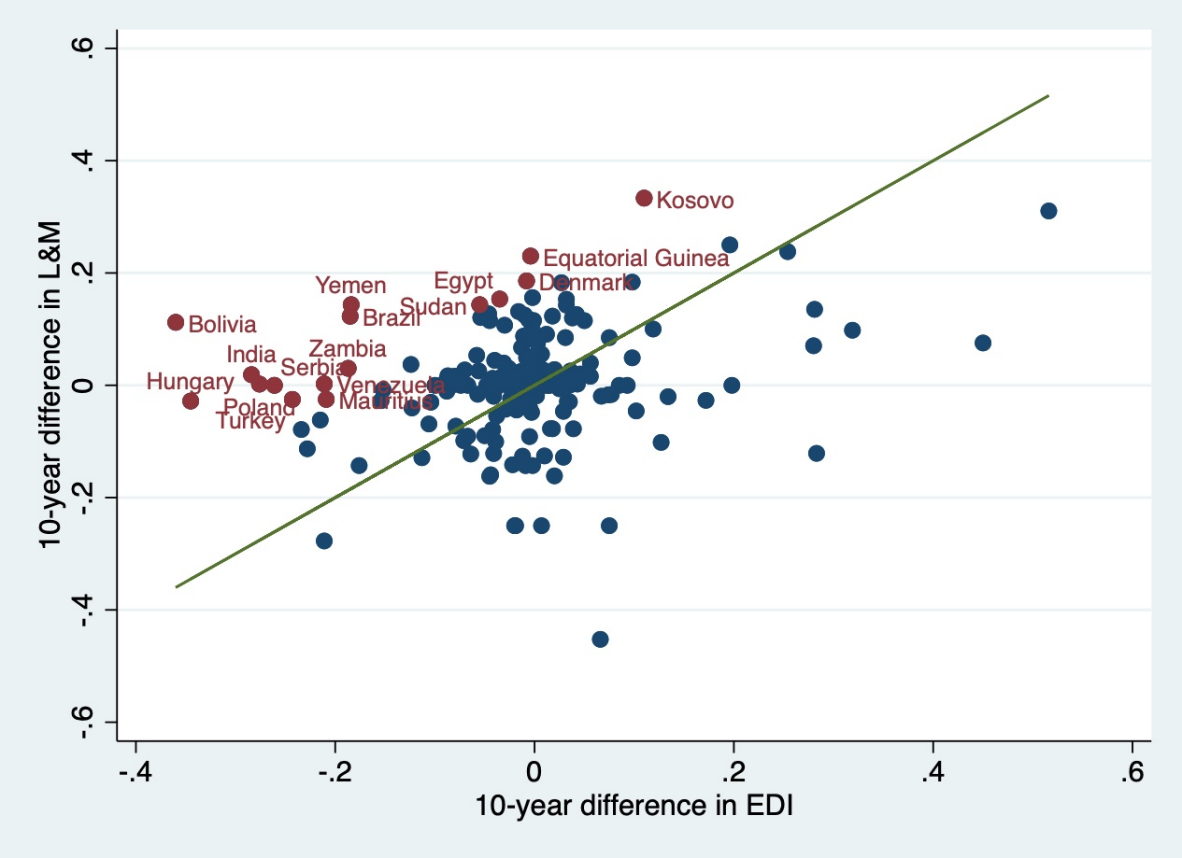

Figure D.5: 10-year differences in V–Dem's EDI and L&M's index from 2010-2020