V-Dem INSTITUTE

Conceptual and Measurement Issues in Assessing Democratic Backsliding

Carl Henrik Knutsen, Kyle L. Marquardt,
Brigitte Seim, Michael Coppedge, Amanda Edgell,
Juraj Medzihorsky, Daniel Pemstein, Jan Teorell,
John Gerring, and Staffan I. Lindberg

May 2023

Working Paper

SERIES 2023:140

THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (V–Dem) is a new approach to conceptualization and measurement of democracy. The headquarters—the V–Dem Institute—is based at the University of Gothenburg with 20 staff. The project includes a worldwide team with five Principal Investigators, 22 Project Managers, 33 Regional Managers, 134 Country Coordinators, numerous Research Assistants, and almost 4,000 Country Experts. The V–Dem Project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V–Dem Institute
Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, PO Box 711
SE 40530 Gothenburg
Sweden
E-mail: contact@v-dem.net

V–Dem Working Papers are available in electronic format at www.v-dem.net.

Conceptual and Measurement Issues in Assessing Democratic Backsliding[*]

Carl Henrik Knutsen[†]
Kyle L. Marquardt[‡]
Brigitte Seim[§]
Michael Coppedge[¶]
Amanda Edgell[‖]
Juraj Medzihorsky[**]
Daniel Pemstein[††]
Jan Teorell[‡‡]
John Gerring[§§]
Staffan I. Lindberg[¶¶]

May 26, 2023

[*] This project received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Program (No. 863486). This research was also made possible in part through support by the Knut and Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg (No. 2018.0144) and co-funding from the University of Gothenburg. The V–Dem measurement process was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden and the Swedish National Infrastructure for Computing at the National Supercomputer Center in Sweden, partially funded by the Swedish Research Council through grant agreements No. 2022-06725 and No. 2018-05973.

[†] University of Oslo

[‡] University of Bergen

[§] University of North Carolina at Chapel Hill

[¶] Notre Dame University

[‖] University of Alabama

[**] Durham University

[††] North Dakota State University

[‡‡] Stockholm University

[§§] University of Texas at Austin

[¶¶] University of Gothenburg

Abstract

This paper addresses three interrelated questions. First, how strong is the evidence that democracy has declined globally over the last decade? Second, how should we best measure (change in) democracy? Third, given that much of the recent evidence for global backsliding comes from measurement projects that rely on expert ratings, is there evidence that experts have become harsher judges of democratic quality in recent years? We begin our analysis with a discussion of how to conceptualize democracy and democratic backsliding, stressing that for contested concepts such as democracy, no one operationalization is likely to reign supreme. We then dissect the distinction between "subjective" and "objective" measures, examining how measurement error can affect even seemingly objective indicators, and highlight how subjectivity pervades all measurement enterprises. Next, focusing on V–Dem's methodology, we show—through both theoretical considerations and empirical tests—that it is highly unlikely that time-varying expert biases drive recent declines in estimates of the state of global democracy. Finally we evaluate Little and Meng's (2023) recent attempt to assess the prevailing case for global backsliding using "objective" measures. We demonstrate multiple issues that make their measurement strategy ill-suited to studying trends in global democracy.

Over the past decade, analyses drawing on different democracy measures have documented a global trend of democratic retrenchment. While these democracy measures are taken from measurement projects that use radically different methodologies, they all rely—at least partially—on subjective judgments by experts to produce estimates of the level of democracy within states. But these measurement exercises do not occur in a vacuum. Large-scale measurement projects—notably Freedom House, Polity, and V–Dem—are well-known in policy communities and their findings are routinely covered by global media. In turn, prevailing trends captured by such projects may become self-reinforcing if experts internalize scholarship, policy pronouncements, and news articles about the state of democracy in the world in a manner that biases their ratings.

But eschewing expert judgment while thoroughly measuring conceptually relevant aspects of democracy is difficult. Indeed, this tension—how to balance conceptual coverage with potential for expert bias—has long animated the massive literature on democracy measurement (Przeworski et al., 2000; Munck and Verkuilen, 2002). Little and Meng (2023, henceforth L&M) re-introduce this debate. More specifically, they argue that "objective" democracy measures show little evidence of global democratic backsliding in recent years.[1] In extension, they posit that time-varying expert bias is driving the appearance of democratic retrenchment in measures that incorporate expert judgments.

So how strong is the prevailing evidence for democratic backsliding? How should we measure democracy to effectively answer this question? Can "objective" indicators provide gold-standard reference points from which to assess "subjective" measures of democracy? Is there compelling evidence that experts have raised their standards for democracy in recent years? Does L&M's approach to measuring democracy make better sense than those adopted by established projects?

Addressing the first and second questions, we argue that assessing whether democracy has recently declined hinges crucially on how one conceptualizes and operationalizes democracy, and on how one summarizes country-level measures. Democracy is a contested concept, and indices measuring different notions of democracy will capture different aspects of political systems. For this and other reasons (such as measurement error), democracy measures may disagree on the level of democracy and trends in democratic backsliding, both for individual countries and in the global aggregate.

Nonetheless, certain conceptualizations may be better suited to the broad goal of monitoring the state of global democracy than others. For example, the top cell of Figure 1 plots the relationship between L&M's "objective" democracy index and V–Dem's Electoral Democracy Index (EDI) (Teorell et al., 2019; Coppedge et al., 2023*a*,*b*) for all

[1] Given the way that subjectivity permeates both measurement and the application of measures to research questions, we contend that it is better to think about concepts as more or less observable, and measures as more or less observer-invariant, rather than thinking of specific measures as more or less subjective. We nonetheless use the terms "subjective" and "objective" throughout this paper to engage with L&M's approach.

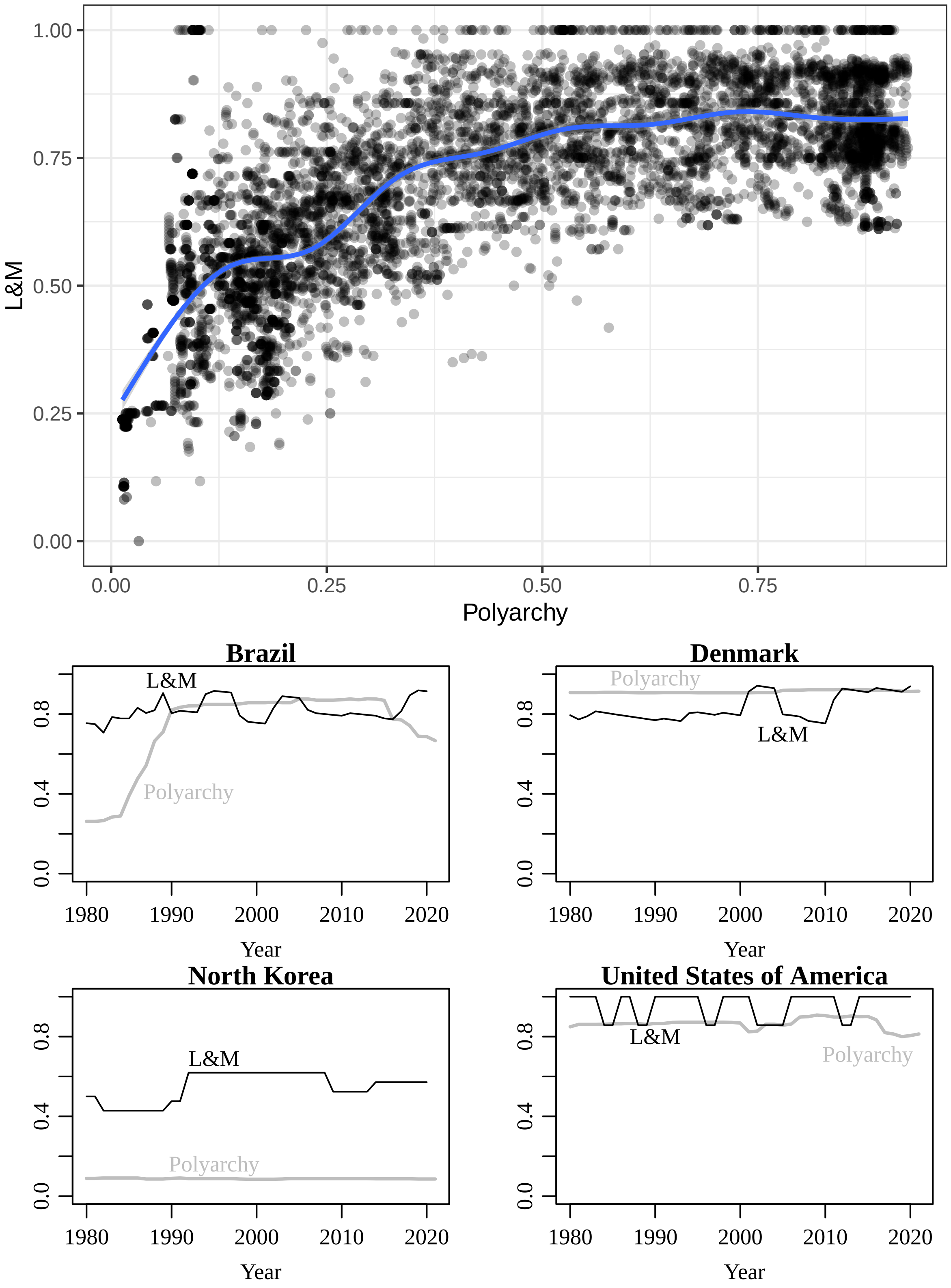

Figure 1: Comparison of L&M index and V–Dem's Electoral Democracy Index, both globally and in specific cases

countries between 1980–2020.[2] There is a moderate correlation between the two indices for country-years with EDI scores below 0.4 ($\rho = .56$), but a much lower correlation for country-years with EDI ≥ 0.4 ($\rho = .16$). In turn, the lower four panels in Figure 1 plot country-level time series for the two measures. In all four cases, there is evidence of divergence in both levels and trends in democracy over time. Clearly, *the two measures tap into very different conceptions of democracy.* As we argue in Section 1, we are now in an age when when anti-democratic leaders have learned how to subtly undermine formal institutions. Capturing democracy therefore requires both a sensitive instrument and a sufficiently thick conceptualization of democracy—covering not only quality of elections but also the depth of political freedoms and the breadth of political opposition. We contend that V–Dem's democracy measures, including EDI, exhibits both of these features.

This focus on conceptualization also animates Section 2's discussion of whether or not we can uncritically rely on "objective" measures as unbiased—or gold-standard—indicators against which to assess expert-coded indicators. We highlight two common misconceptions. First, the distinction between "objective" and "subjective" indicators is often overblown, as coding seemingly objective indicators typically requires multiple hidden judgments by human raters. Measures of contested concepts, such as democracy, are not objective or subjective, but rather more or less judgment-based. Second, indicators that require fewer judgments to code are not, as a rule, any more likely to be unbiased than indicators that rely on substantial human judgment. Whether or not a particular "objective" indicator exhibits bias in capturing a given concept varies dramatically based on the mapping between concept and indicator, and on the data construction process.

Building on these points, Section 3 explores potential for bias in V–Dem's measures, including EDI. We argue that V–Dem's methodology limits the potential for systematic expert bias by 1) asking experts to code specific sub-components of democracy and then aggregating up, 2) rigorous survey design, 3) a systematic expert recruitment process, and 4) statistical methods designed to mitigate the impact that biased and unreliable experts have on aggregate scores. We also interrogate the question of whether or not V–Dem's experts have exhibited systematic pessimism (giving rise to a "bad vibes bias") in recent years, and find no evidence for such claims.

Finally, Section 4 analyzes L&M's measurement strategy. As Figure 1 foreshadows, we find little evidence that L&M operationalize democracy in a way that maps onto any easy-to-recognize version of the concept. Moreover, various subjective coding decisions—such as how they select indicators, aggregate measures, and deal with missing data—crucially undermine the credibility of their measurement exercise, whether the goal is to broadly measure democracy over time or only to evaluate aspects of democratic backsliding on a measure-by-measure basis.

[2] EDI is also referred to as the "Polyarchy index" throughout this paper.

1 Foundational Building Blocks for Assessing Democratic Backsliding

Assessing democratic backsliding requires four components: a conceptualization of democracy; a conceptualization of backsliding; measures that translate the conceptualization of democracy into its constituent (and aggregate) indicators; and an unbiased and reliable approach for collecting data on these measures. Insufficiently developing any of these components will undermine one's ability to obtain an accurate and unbiased assessment of democratic backsliding. In this section, we overview some considerations on conceptualizing democracy and backsliding. These considerations are not novel but delineate a "common ground" upon which we deal with the more technical measurement issues arising in the next sections.

1.1 Conceptualizing Democracy

Conceptualizing (and subsequently measuring) democracy is a vast area of scholarship. Here we highlight a few core points with implications for the measurement of global democratic backsliding. We note first that scholars conceptualize democracy in various ways (Coppedge et al., 2011). A particularly important distinction is that some democracy definitions are minimalist, focusing only on contested elections (e.g., Przeworski et al., 2000); whereas others are maximalist, incorporating several additional attributes (e.g., Beetham, 1999).

Despite the lack of consensus on how to define or best measure democracy, most democracy scholars agree it is desirable to 1) specify the democracy concept that one wants to measure and 2) select indicators and aggregation rules that capture the underlying concept in an unbiased and reliable manner as best as possible (see, e.g., Adcock and Collier, 2001; Munck and Verkuilen, 2002; Goertz, 2006). In other words, conceptualization precedes measurement, and measurement characteristics need to be attentive to the boundaries and logical structure of the concept. (Scholars must also attend to other characteristics such as spatial-temporal coverage and reproducibility).

While a narrow conceptualization of democracy can be helpful in many research contexts, we argue that a more encompassing measure is most useful for monitoring the global state of democracy. Holistically, democracy is about election quality, but it also encompasses political freedoms and the ability of political opposition to effectively compete in both elections and public discourse. Using measures with sufficiently broad conceptualizations of democracy is especially important if we want a measure that is nuanced and sensitive enough to be useful in an age when autocrats have learned how to abuse formal democratic institutions (e.g., Bermeo, 2016; Levisky and Ziblatt, 2018; Lührmann and Lindberg, 2019).

1.2 Conceptualizing Backsliding

The notion that democracy measurement influences reported trends in democracy, including the nature and depth of recent democratic decline globally, is established knowledge among democracy scholars (for a recent analysis, see Treisman, 2023). For example, Waldner and Lust (2018) find that, *inter alia*, V–Dem measures are more conservative in terms of identifying backsliding relative to other prominent measures such as Freedom House. Knutsen and Skaaning (2022) describe how different V–Dem measures have followed different trends when computed as global averages across time. In recent years, aspects of democracy such as freedom of expression and association as well as deliberative democratic aspects have declined, whereas suffrage and the extent to which officials are elected have remained stable. Even the V–Dem Institute's Democracy Report (e.g., Alizada et al., 2022) reports variation in this phenomenon based on measurement technique: the recent decline in democracy is much larger for population-weighted global scores than pure averages where all countries are treated equally.

These examples point to the fact that there are a myriad of ways to conceptualize "democratic backsliding" or the many similar terms that are widely used. Backsliding could refer to country-specific trends or phenomena at a regional or global level. For the global level, it could be measured by taking a simple average across countries or by using population-weighted trends.[3] It could be conceptualized as a short-term phenomenon—something that occurs within a year—or a long-term process—something that takes multiple years—or even an entire "episode."[4] It could apply to changes observed in all countries or only in those that are initially "democratic." Finally, one's conceptualization of democratic backsliding depends, in a large part, on one's conceptualization of democracy, for example whether democracy is conceptualized as unidimensional or multidimensional, as a binary state or a spectrum.[5]

[3] Note that analyses of global democracy trends in recent years tend to show much more backsliding for population-weighted measures than aggregates where all countries are weighted equally (e.g., Alizada et al., 2022). Weighting countries by population is a valid approach to describing trends for the average person on the planet. However, this approach raises a different debate than that of L&M and unweighted average scores remain a more conventional metric for judging global trends. Moreover, population-weighted changes in practice say a lot about developments in large countries — especially India in recent years — and less about the rest of the world, making it less well suited to analyze how systematic cross-country tendencies are.

[4] While we do not pursue such an approach here, we refer those who wish to explore an analysis of custom-defined episodes to, for example, Lührmann and Lindberg (2019), Haggard and Kaufman (2021), Pelke and Croissant (2021), and Maerz et al. (Forthcoming).

[5] If the concept of backsliding pertains primarily to democracies, a prerequisite decision is how to categorize countries as sufficiently democratic. Some choose to use cut-offs on a continuous index of democracy such as V–Dem's Electoral Democracy Index (EDI). Others consider transitions between categorical measures (typologies) that force countries into "democracy" and "autocracy" categories. However, these categorical measures themselves have a judgment call embedded in them, with some using a high threshold for categorizing countries as "democracies" (e.g., the Regimes of the World measure by Lührmann, Tannenberg and Lindberg (2018) used in the V–Dem Institute's Democracy reports) and others using a lower threshold (e.g., Boix, Miller and Rosato (2012). There are many typologies and

Rather than delving into all of these conceptual decisions, we highlight three conceptual points regarding backsliding that have pivotal implications for its measurement. In the list that follows, we delineate the dominant two opposing views on each point and put in parentheses the implication of the view for quantifying backsliding.

1. Does our conceptualization of democracy—and therefore backsliding—focus on competitive elections (less backsliding) or consider checks on executive power, protections of civil liberties, a critical media, and active civil society (more backsliding)?
2. Does our conceptualization of backsliding treat it as a short-term phenomenon (less backsliding) or longer-term process, occurring over, say, five or ten years (more backsliding)?
3. Does our conceptualization of backsliding pertain to democratic, or even very democratic countries (less backsliding), or can backsliding theoretically occur in any country, regardless of its level of democracy (more backsliding)?

Before turning to more of a technical discussion regarding measurement in Section 2, we discuss and illustrate the implications of these three conceptual points. Concerning point 1), many scholars studying backsliding (Bermeo, 2016; Levisky and Ziblatt, 2018; Lührmann and Lindberg, 2019) agree that backsliding in democracies predominantly occurs through processes driven by elected incumbents who gradually concentrate power in their own hands – aka "executive aggrandizement." This leads to a more subtle degrading of (less formalized and harder to observe) supporting pillars of democracy, such as civil liberties, civil society mobilization, press freedom, and judicial independence. These features of political regimes are not conceptually included in a minimalist definition or operationalization of democracy. By contrast, more directly observable characteristics of elections do not feature as prominently in recent backsliding episodes. This could partly explain why "objective" measures of democracy, typically centered on electoral outcomes, show less backsliding.

Regarding point 2), a conceptualization of backsliding that focuses on short-term shocks to democracy is going to result in much lower rates of backsliding than a conceptualization of backsliding that treats it as a long-term process. To illustrate this point, consider Figure 2, which shows the relative frequencies of clear negative or positive changes in V–Dem's EDI over one-, five-, and ten-year periods. Over a one-year period, the modal outcome is for EDI to remain constant (top panel, lighter shading) in the sense there is no registered change or the change fails to achieve conventional thresholds for "statistical significance". In contrast, over the five-year or ten-year periods (middle and bottom panels), most countries experience either positive change (blue shading) or negative change

many differences across them, and each difference implies a different operationalization of backsliding.

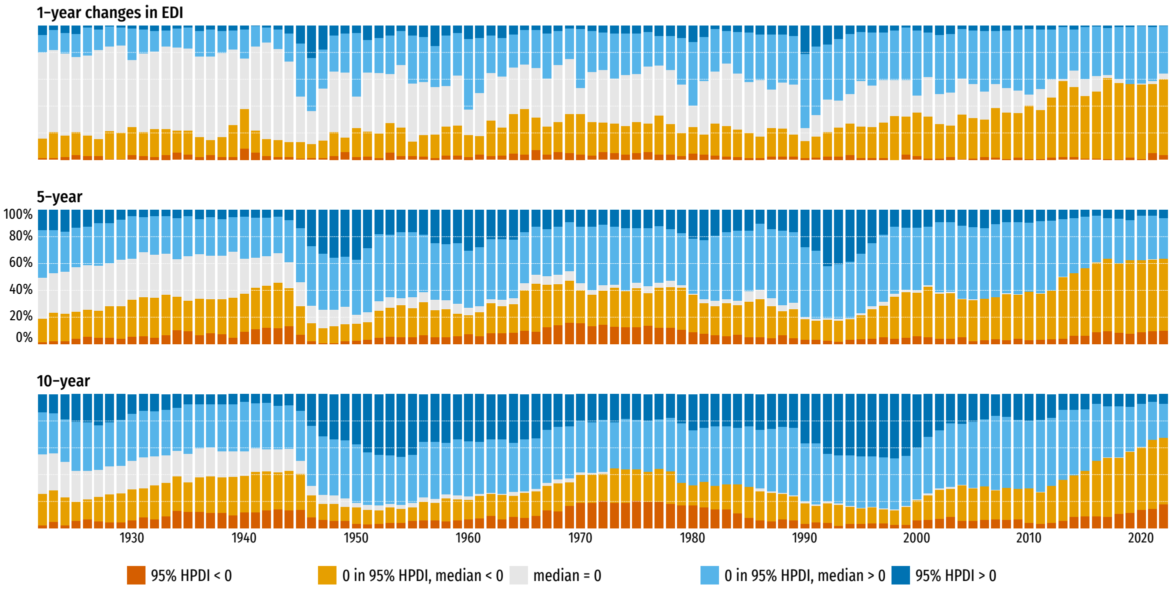

Figure 2: Relative frequencies of one-, five-, and ten-year changes in EDI by year. Orange-shaded areas indicate negative changes and blue-shaded positive changes without 0 in 95% highest posterior density intervals.

(orange shading). A more long-term conceptualization of backsliding would view these negative changes as democratic backsliding, but a more shock-focused conceptualization would not.

Point 3) pertains to the scope of countries eligible for backsliding. A narrower conceptualization of backsliding will only include the most democratic countries as potential backsliders. A broader conceptualization of backsliding will include countries across the democracy spectrum (or most of it). Figure 3 separates countries by EDI quartile. Within each quartile, 10-year trends of countries resulting in positive change are shown in blue and those resulting in negative change are shown in orange. Countries with negative trendlines occur across the EDI quartiles. Most notably, the bottom two EDI quartiles include countries widely regarded as paradigmatic examples of backsliding, such as Nicaragua and Thailand.

2 “Objective” vs. “Subjective” Democracy Measurement: An Exaggerated Distinction

As stated in the previous section, conceptualization should precede measurement. In their foundational article, Munck and Verkuilen (2002, henceforth M&V) specify the tasks faced by those who want to measure democracy in a valid manner. These include identifying all the relevant attributes of the concept as well as what indicators tap into each specific attribute, avoiding the problems of redundancy and conflation. All relevant attributes

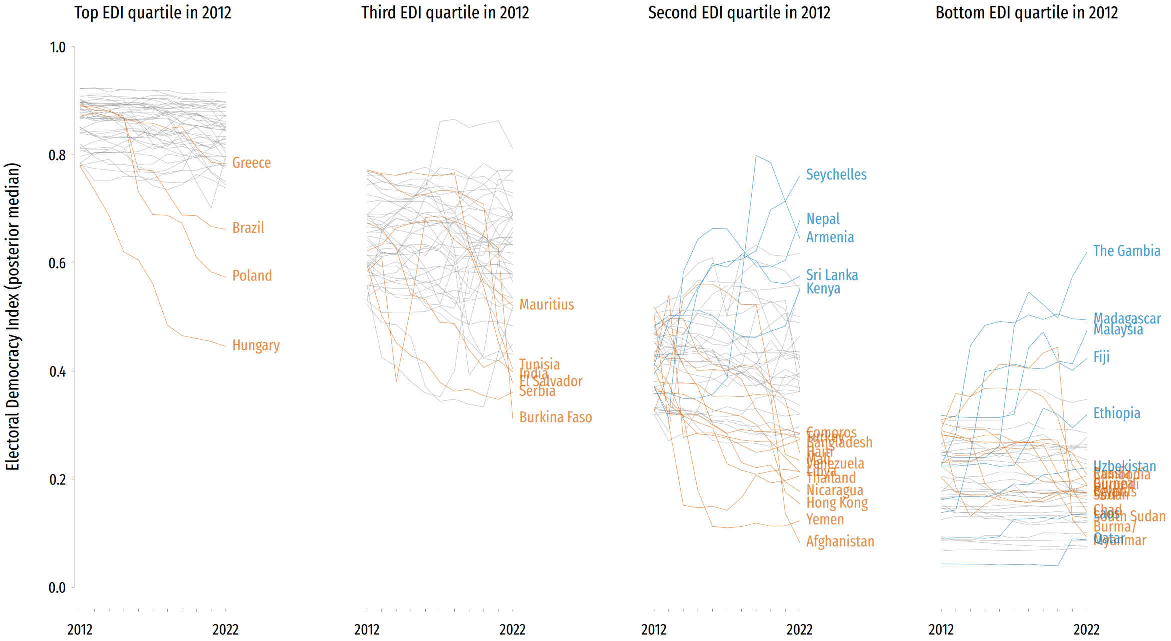

Figure 3: 10-year trajectories of countries with clearly visible (without 0 in 95% highest posterior density intervals) negative (32) and positive (13) changes.

and indicators should be included, and irrelevant indicators unrelated to the democracy concept excluded. The authors also highlight the importance of selecting aggregation rules that reflect the concept's logical structure, avoiding simple default rules (e.g., just taking the unweighted average across all indicators by default). In cases where some indicators reflect more important aspects of the concept, they should be weighted more heavily. In cases where indicators reflect attributes that are necessary for high democracy scores, the aggregation rule should be multiplication instead of averaging (see also Goertz, 2006).

M&V find that many existing measures of democracy fail to meet many of these challenges. For example, they argue that Freedom House does not build on a clearly specified democracy concept and (as a likely consequence) includes several redundant indicators; they also use simple aggregation rules that lack justification. M&V also criticize measures Polity for problems including problematic aggregation procedures and the omission of core democracy aspects such as participation. In contrast, one measure that M&V assess positively is the dichotomous ACLP measure (e.g., Przeworski et al., 2000), which they laud for having a carefully specified underlying concept and for being aligned with this conceptualization. However, this measure only captures a minimalist democracy concept focusing on contested elections, and codes regimes as either democratic or dictatorial. The ACLP is thus not fit for capturing a more multidimensional concept of democracy.

Despite the method's limitations, it remains common to rely on "objective" measures to operationalize democracy (see, e.g., Cheibub, Gandhi and Vreeland, 2010). Therefore, before turning to our analysis of bias in the V–Dem data in Section 3, we discuss the blurred line between objective and subjective measures and the potential for bias in supposedly objective measures.

2.1 "Objective" Measures Require Subjective Decisions

Although the more general principle underlying the distinction is relevant for understanding differences between democracy measures, a crisp objective vs. subjective categorization constitutes a false dichotomy: seemingly "objective" measures often have considerable elements of subjectivity baked into them.[6] Although there may be relative differences in bias between different types of human raters—country experts, citizens answering surveys, or research assistants—there is no type of rater, and thus no democracy measure, that escapes the need for human judgment (see Marquardt et al. (2017)). In this subsection, we illustrate the human subjectivity inherent in "objective" measures with several examples.

First, "fact-based" measures from V–Dem illustrate this point. These variables, which complement V–Dem's expert-coded variables, cover concepts such as the seat or vote share of the largest party in parliament and are coded by research assistant raters.[7] Despite being based on observable facts, these variables require several subjective decisions (e.g., how to code independents, which round of an election to consider, and how to deal with diverging sources). These V–Dem variables are therefore accompanied by careful protocols and routines for deliberation within the broader team for dealing with ambiguous cases. Even so, these "objective" indicators involve human judgment and, therefore, are not perfectly reproducible.

As a second illustration of the difficulty of coding even seemingly straightforward "objective" measures, consider the most widely used and highly regarded "objective" democracy measure: the binary, minimalist ACLP (also known as DD; Przeworski et al., 2000; Cheibub, Gandhi and Vreeland, 2010). ACLP relies on four coding rules pertaining to observable factors (elections to legislature, to cabinet, elections are multi-party, alternation of government after loss in elections). Yet, Knutsen and Wig (2015) note that this measure is not always easily reproducible due to the judgment call needed to both apply the alternation rule and then to determine exactly how long the current regime institutions have existed.[8]

[6] As the analysis in Section 3.2 indicates, there are arguably also differences in "degree of subjectivity" for different measures with clear evaluative components, such as V–Dem's expert-coded measures.

[7] The individuals who compile and code "objective" indicators are referred to differently across projects. We refer to them as "raters" in this article.

[8] Knutsen and Wig (2015) discuss the case of Mexico at the turn of the 20th century. The ACLP coding for this case assumes that the adoption of an autonomous election monitoring board in 1996 marked the relevant change in regime institutions. As a result, Mexico changed from an autocracy to a democracy before government alternation, which happened four years later, in 2000. While a plausible marker for changed regime institutions, this coding decision is not perfectly reproducible. Hence, the ACLP measure is not perfectly reproducible either.

Another commonly employed “objective” indicator, turnover in the executive, provides another illustration of the subjective elements related to coding even “objective” indicators. Consider Kenya’s 2022 general elections. The winning candidate, William Ruto, was at the time the Deputy President. However, he ran in opposition to Raila Odinga, the candidate endorsed by outgoing President Uhuru Kenyatta. Does this result constitute electoral turnover? From a purely partisan perspective, it does. From a “changing of the guard” perspective, it does not. Kenya is far from the only example of an ambiguous case, indicating that even this seemingly objective indicator involves considerable subjective evaluation.

The degrading of competitiveness in elections — even if multiparty elections formally remain in place — results in other situations that often require judgment calls. For example, if Turkey as of April 2023 is an autocracy, exactly when did institutions shift so significantly that the country’s regime turned from a democracy to an Erdoğan-centered autocracy? ACLP’s coding rules for raters are not clear in such cases, hindering reproducibility even for this gold standard for objective measures.

Another case in point is electoral interruptions, for example in the form of a coup followed by the dissolution of parliament. One way to deal with them is to not account for them at all. That is, an objective democracy measure capturing features of the last election could simply ignore making any “subjective judgments” on whether or not a coup has happened. Another version is to apply a stopping rule, for example that an election remains relevant for six years unless there is a new one. This approach amounts to continuing to count the results from the last elections as relevant even when they are, in practice, not. Using data from NELDA on whether the incumbent party lost, L&M do so with their “objective” democracy index, which we discuss in more detail in section 5. They thus do not account for the coups in, e.g., Yemen in 2015 or Sudan in 2019. Using these rules, one would thus still count the 2019 presidential election in Afghanistan as relevant when coding the degree of democracy under today’s Taliban regime, which seems unreasonable.

Most generally, the lack of perfect inter-rater reliability in “objective” indicators provides a final illustration of our point. Consider the NELDA project, which codes objective phenomena related to elections. The project is transparent about rater disagreement and subjectivity, providing analyses of inter-rater reliability for approximately 20% of their cases (Hyde and Marinov, 2012, 2019).[9] Across their 58 indicators, the proportion of cases with agreement between two raters ranges from 58% to 98%, with an average of

[9] Other projects elaborate on only their decisions for presumed difficult cases. For example, the DPI project provides information about “ambiguous” cases to justify coding decisions where other raters would plausibly have come to other conclusions (Cruz, Keefer and Scartascini, 2021); for some indices, they assign intermediate values to cases in which they are unsure about the specific ordinal category to which it belongs. This case again illustrates that subjectivity is pervasive in gold-standard “objective” data gathering enterprises.

83%. The fact that disagreement exists for all their indicators — and is often pervasive — is evidence that judgment could substantially affect "objective" data.

Given these concerns, it is worth noting that "subjective" approaches that use expert evaluations—such as that of the V–Dem Projec —*can* be more reproducible than their "objective" counterparts. Although V–Dem uses multiple experts to provide openly subjective assessments of numerous political phenomena, and is thus superficially less reproducible than "objective" measures, it is transparent about the use of judgment and the uncertainty this entails. All V–Dem expert scores are publicly available, as are the aggregation methods and criteria, from the aggregation of expert-level to country-level scores to the creation of high-level indices from indicators. Moreover, uncertainty based on judgment is explicitly estimated when aggregating expert codings and propagated throughout meso- and high-level indices, meaning that users of the V–Dem data can easily identify contexts where there is more disagreement between experts' evaluations. The V–Dem data production process is thus *replicable*, in that anyone could, in principle, apply the same process with a new set of experts. It also includes estimates of measurement error, which provides a systematic tool for predicting replication variability. While no two replications of the V–Dem process are likely to produce identical data, a replicable process that thoroughly operationalizes a given concept may often be far preferable to a reproducible approach that only partially captures that concept.

2.2 "Objective" Is Not Unbiased

Scholars regularly use the terms "objective" and "subjective" to describe measures of concepts like democracy, typically without defining the terms. In everyday language, "objective" tends to mean either "impartial" or "unbiased." However, this meaning does not transfer to the world of measurement. In measurement vernacular, "objective" instead corresponds to what we call "observer-invariance:" all observers get the same result when using the measure to code the same case.[10] In the previous subsection, we demonstrated that "objective" measures may not always be observer-invariant. However, even if they were, objective measures can nonetheless be more biased than subjective measures that explicitly include an evaluative component. Critically for the measure of backsliding trends over time, time-varying biases can also afflict objective measures.

We illustrate these points with several examples. First, consider a measure that said: "A country is democratic if and only if it has 'democratic' in the country name." If observer-invariance were the most important criterion on which to judge democracy measures, this measure would be preferable to most other democracy measures: it is

[10] Thinking of democracy measurement as estimation helps to clarify this issue. In estimation, two kinds of error can be present, bias and variance: systematic deviation from the estimand and noise, respectively. Observer-invariance eliminates noise but not bias. This further highlights how "objectivity" in democracy measurement does not correspond to the everyday notion of objectivity.

perfectly reproducible. Yet the scores of such a measure have very low face validity. Indeed, it will be a strongly biased measure of any plausible definition of democracy, since some authoritarian regimes are eager to use their country's name to signal democratic credentials (e.g., the People's Democratic Republic of Korea, the German Democratic Republic), whereas democratic regimes do not have the same need to signal their credentials.

The ACLP/DD measure provides a less extreme example of potential bias in objective measures (Przeworski et al., 2000; Cheibub, Gandhi and Vreeland, 2010). Knutsen and Wig (2015) argue that this measure systematically miscodes young democracies with a track record of high economic growth as autocracies. As a result, the measure produces bias when estimating the relationship between regime type and growth. Indeed, the bias is mitigated for otherwise similar electoral democracy measures (notably Boix, Miller and Rosato, 2012) that invoke subjective evaluations on the freeness and fairness of elections (instead of only observing government alternation after election losses). In this case, including subjective evaluations mitigates a bias.

When attempting to document a trend over time (for example, backsliding), time-varying measurement bias is particularly problematic. "Objective" measures are not immune to this affliction. For example, one method for detecting election fraud is to consider the distribution of second digits in officially reported results from electoral precincts (for a review and general critique of such "election forensics" methods, see Medzihorsky, 2017). This method relies on the idea that regime officials who try to cheat will be prone to select numbers that often end on the same digit (e.g., 0). However, once researchers have publicly discussed this pattern and methods for detecting fraud, authoritarian regime officials may adapt and randomize second digits for fraudulent election results so they are not detected. Consequentially, such digit-based tests will become an increasingly poor proxy for detecting electoral fraud over time.

This methodological point has important implications for L&M's finding that some objective measures show less backsliding than subjective measures. In an age where democracy is a high-legitimacy system that even autocrats try to emulate (for example by holding multi-party elections, see Miller (2015)), and where the main threat to democracy is self-coup by elected leaders (Lührmann and Lindberg, 2019), "objective" measures that typically pertain to elections and the electoral process might increasingly be a poor proxy of wider concepts of democracy. In a context where would-be autocrats are elected but then concentrate power in their own hands, successful autocratizers follow subtle and less easily discernible acts to (gradually) undermine democracy (Levisky and Ziblatt, 2018). Conversely, they avoid blatant and easily documentable actions that can serve as focal points for opposition mobilization and push-back. For example, winning an election with 55% is still winning, and it is a less suspicious-looking win than achieving 95% of the votes. In brief, "objective" indicators that are sensitive to blatant and documentable actions such as rigging elections may be increasingly biased as a measure of the broader state of democracy over time.

2.3 Conceptualization Revisited

In principle, conceptual thickness is orthogonal to whether a measure relies heavily on human judgment. Yet "objective" measures of democracy tend to tap narrow conceptualizations out of necessity. Indeed, we know of no such objective measure that captures all of the attributes that would be included in a more multidimensional and continuous concept of democracy. In contrast, the V–Dem democracy measures tap into clearly specified, more encompassing, and continuous democracy concepts and have aggregation rules that reflect these concepts (for a measure capturing Robert Dahl's polyarchy concept, see Teorell et al., 2019). However, to have this multidimensional characteristic, the V–Dem measures include not only "objective" indicators such as the share of adult population with suffrage and the election of legislators but also several "subjective" indicators that rely on expert coding, such as freedom of discussion for women and the autonomy of election management bodies. These subjective indicators are essential because they capture dimensions of democracy where relevant and comprehensive "objective" indicators are lacking. This fact highlights a concluding point about the difference between "subjective" and "objective" measures of democracy: *they generally operationalize different concepts.*

3 Biases in V–Dem Measures? Clarifications and Tests

We have argued that the best way to monitor global democracy trends is to rely on sensitive indicators that broadly cover the concept. And, such measures will have to rely on subjectively coded data, at least to some extent. Expert-coded data gathering enterprises such as V–Dem thus face the fundamental question of whether or not experts are systematically biased when rating institutions related to democracy.[11]

Here we focus on the particular form of bias L&M put forth as a possible explanation for reported global democratic backsliding: Experts could be (increasingly) negatively biased in recent years due to pessimism about the state of democracy. While we cannot entirely discount the possibility of "bad vibes bias," we are skeptical that it strongly influences reported global trends in democracy. In this section, we explain this skepticism, first by clarifying important aspects of the V–Dem methodology and next by presenting empirical tests for expert bias.

[11] For more detailed discussions on the V–Dem methodology and measures, see, Coppedge et al. (2020, 2023*c*).

3.1 V–Dem's Methodology: Limiting Expert Expert Biases

V–Dem's approach to gathering and aggregating expert-coded data militates against bad vibes bias in multiple ways. First, V–Dem experts do not code broad concepts like "democracy" directly.[12] Instead, V–Dem experts predominantly code more specific concepts such as election violence or the degree to which media self-censors in a given country-year. These concepts are presumably less prone to general mood affecting coding, due to their specificity. To create meso-level (e.g., freedom of expression) and high-level indices (e.g., the Electoral Democracy Index), V–Dem algorithmically aggregates these scores upwards. In order for pessimism to affect ratings of high-level concepts like democracy, it would have to similarly affect not just most V–Dem experts, but the ratings of these experts across multiple indicators related to different, specific, low-level concepts.

Second, V–Dem experts use an ordinal scale with set definitions for each item that serve as meaningful benchmarks to guide their coding. Again, such specificity ameliorates concerns that a general unease about democracy could greatly affect estimates. They also code anchoring vignettes designed to help standardize how they use these ordinal scales and many experts rate multiple time periods in a given coding session, which should encourage further strictness norming.

Third, the specificity of V–Dem experts' *expertise* should lessen the risk of bad vibes bias. V–Dem experts are individuals — typically scholars with advanced degrees — with deep knowledge of both the countries and sectors (e.g., judiciary, civil rights, political parties) they rate. Most of them are citizens of, or residents in, the countries they rate. Although experts might contextualize their ratings based on their overall sense of political trends worldwide and in the country they are coding, we expect that their detailed expertise allows them to also understand whether or not their cases are exceptions to these general trends.[13]

Fourth, the V–Dem method for aggregating expert data accounts for the possibility that individual experts have idiosyncratic biases that could affect the accuracy of V–Dem data overall. Specifically, V–Dem uses a Bayesian Ordinal Item-Response Theory Measurement Model (MM) to aggregate expert-coded data (Pemstein et al., 2023). It corrects for two forms of error that might emerge from the argument that L&M propose: (1) variation in expert reliability and (2) scale perception. Concerning expert reliability, the MM assumes that experts who diverge from other experts in terms of directionality (i.e., coding higher scores higher) are less reliable.[14] The MM accordingly downward adjusts the contribution

[12] The indicator L&M use to exemplify V–Dem's strategy is an exception to this general rule: V–Dem's composite "free and fair elections" summary measure.

[13] More specifically, L&M note that many V–Dem experts could be influenced by living in the US and the particular pessimistic mood in that country, regardless of the country they code. However, only 12 percent of V–Dem country experts who participated in the 2023 coding actually live in the United States, which makes this argument implausible as an explanation for the broader trends in global democratic backsliding.

[14] The model initially assumes that all experts are equally reliable and correctly perceive directionality

of less reliable experts to the measurement process. As a result, if some experts for a given country shift their scores downward due to bad vibes about democracy—while most other experts for that country do not—the bad-vibing experts will likely be considered less reliable and thereby contribute less to the estimation process. A country's score on a variable is therefore unlikely to drop much unless the majority of experts, within each country, experience similar bad vibes.

The MM also assumes that experts vary in their strictness, thus accounting for some experts tending to score higher or lower than others on an ordinal scale or having different thresholds for changing scores.[15] The experts most likely to change their ratings due to bad vibes about democracy presumably have more compressed thresholds than more vibe-resistant experts, which means that the effect of these experts changing their scores should be less drastic. That is, if (only) some experts systematically lower their scores in response to pessimism about global democracy, such actions will probably not result in a major change in any country's overall score.

Despite the demonstrated ability of the MM to recover latent scores across various scenarios (Marquardt and Pemstein, 2018, 2023; Marquardt, 2020), the MM does not directly correct for time-varying changes in strictness (e.g., time-varying bias induced by bad vibes): if many of the experts who rate a given country become more strict in unison, for a subset of years that they code, the MM would likely be unable to adjust for this time-sensitive change in rating behavior. However, V–Dem has long been aware of this general issue (see Knutsen et al., 2019; Pemstein et al., 2023). For example, historical V–Dem experts, who typically code 1789-1920, tend to be somewhat less strict than contemporary experts, likely because their frame of reference is different. Similarly, there is some evidence that experts who only code recent years (2005-present) have different levels of strictness than experts who coded the entire time series for a country, e.g., from 1900 and onward. Accordingly, V–Dem has long deployed technical fixes to account for these differences.[16] Another particularly important innovation started in 2020, when V–Dem began asking newly-recruited experts coding only recent years to also code a sequence of dates covering a country's entire time series from 1900 onward. By expanding these experts' frame of reference, this strategy should facilitate contextualization of short-term recent trends and thereby reduce the likelihood that general pessimism about democracy affects contemporary codings.

before it adjusts idiosyncratic expert reliability based on their coding patterns.

[15] The MM assumes that experts have idiosyncratic thresholds for converting latent country-year values to ordinal ratings, which are clustered first by country of expertise and then universally. In addition to data from codings, the model incorporates data from anchoring vignettes and other bridging patterns (Pemstein, Tzelgov and Wang, 2015).

[16] The MM uses empirical priors that are offset by the period for which an expert provided ratings. The specific offset is the difference in ratings between newly-recruited experts for a given country-variable and those experts who rated the entire time period. In principle, this strategy should downweight the ratings for country-years with newly-recruited experts who provide relatively high ratings and upweight the ratings for country-years in which these experts provided relatively low ratings.

3.2 Evaluating Expert Expert Biases in V–Dem

Despite all of these strategies, temporal proximity to events that are being coded could still affect expert strictness: the issue might not be an expert's frame of reference, but uncertainty due to the act of coding political changes that are still unfolding. Unfortunately, it is difficult to assess the scope of this concern in a direct manner, as multi-expert bias manifests behaviorally in the same way as actual change in the concepts of interest: experts adjust their ratings in similar ways. Since V–Dem experts rate political institutions for which there are no easy-to-observe and directly-measurable indicators, we cannot use "objective" indicators for these phenomena to assess potential bias in contemporary expert codings in the manner that L&M attempt. Instead, in this section, we employ a series of indirect tests that speak to the extent of bias in V–Dem's expert-coded measures.[17]

First, the "bad vibes" theory implies a fairly uniform erosion of democracy across countries *and* across components and indicators of democracy.[18] Contrary to this implication, V–Dem data show that only a minority of countries have registered significant erosion in the EDI. For example, if we define "significant" as a 95% highest posterior density interval that does *not* contain zero change, the great majority of countries do not meet this criterion. If we consider the time span from 2012 to 2022, 32 countries experienced negative change and 13 had positive change; more than 100 countries experienced no significant change in EDI in this period (see Figure 2 and Figure 3). Moreover, for countries that *do* meet the above criteria for a "significant" democracy decline, we observe that some components and indicators have declined, others have remained unchanged, and in some cases some have even improved.[19] In fact, the indicators most vulnerable to the "bad vibes" phenomenon—those on elections such as vote buying and EMB capacity—are least likely to show significant declines.

Second, given the prevailing zeitgeist, "bad vibes bias" should affect the coding of a country's scores for several years but it should *also* be stronger for coding decisions taking place in later years. Insofar as recent coding by V–Dem experts mainly centers on their country's situation in recent years, any such increase in bias is hard to distinguish from a change in the democracy situation on the ground. However, one feature of V–Dem's coding

[17] Various scholars within and external to V–Dem have extensively grappled with assessing and addressing expert biases, and our analysis in this section builds on this body of work. See, for example, McMann et al. (2022); Coppedge et al. (2020); Marquardt et al. (2019); Schedler (2012); Weidmann (2023).

[18] L&M's formal model (their Appendix C1) explicitly assumes that expert bias is linear and additive, affecting all units equally.

[19] From 2010 to 2020, more than 60 countries, globally, dropped significantly on five specific indicators from the EDI and V–Dem's Liberal Component index (CSO repression, government censorship effort, freedom of academic and cultural expression, freedom of expression for men, and media self-censorship), while fewer than 25 countries lost ground on property rights, voter registry, or party bans. These statistics are based on lenient 68% confidence bounds; if we use more conventional 95% bounds, the number of countries experiencing backsliding on the different variables shrinks to a range of two to 42 countries. For all expert-coded variables that make up the EDI and Liberal component index, more countries did not change significantly than declined in this period.

allows us to separate the two processes: V–Dem country experts can update/change their previous scores when coding annual updates for V–Dem. If "bad vibes bias" is more prevalent in recent years, we should observe that: a) many experts update their previous coding; and b) that they consistently do so in a "more pessimistic" direction. To assess this, we combine expert-level information from the 2019 version of the V–Dem dataset (v.9) and the most recent version (v.13; 2023) to assess whether we see country experts changing their scores in a manner that makes the countries they code systematically less democratic. For this analysis, we focus on the expert-coded indicators (45) that enter V–Dem's Electoral Democracy Index (EDI) or the Liberal Component Index (LCI). In brief, we find that less than two percent (1.4%) of experts change scores for any variable, and the ones who do change do not systematically alter their scores in a more pessimistic direction. The variables with averages that are unaffected by revisions are plotted in Figure 4 for EDI and Figure 5 for LCI.[20]

Third, insofar as a common bias is affecting experts across countries, one would arguably expect a greater synchronization of changes in V–Dem's indices in recent years.[21] To assess this possibility, we have taken all 1-year changes for each year on V–Dem's EDI and computed their mean, median, standard deviation (SD), median absolute deviation from the median (MAD), and estimated entropy (EE). Figure 6 reports these change statistics for the last 100 years. The posterior intervals for mean and median change show that, in recent years, these quantities are minuscule, and their posterior intervals generally contain zero. This finding does not fit well with the claim that there has been increasing systematic downward bias over the last decade. Moreover, the trends in the SD, MAD, and EE show that, if anything, one-year changes are now *less* homogeneous than they have historically been. These various measures of dispersion all systematically, but modestly, increase over time, as one might expect given that experts will generally have access to a wider range of information for recent cases than for historical ones. Overall, these findings reinforce the information in Figure 2 and Figure 3: in most years, including the past ones, only a small fraction of countries shows clear changes up or down, e.g., in V–Dem's EDI. Experts' perceptions of backsliding are narrowly focused on specific cases, in addition to specific indicators.

Finally, it is possible that ideological bias (including, but not limited to, "bad vibes bias") manifests itself in systematic expert disagreement.[22] We interrogate this concern in

[20] See Appendix A for the full analysis. Appendix Figures A.4– A.7 report v.9 to v.13 patterns for variables with averages revised either up or down. Appendix Figures A.1-A.3 report the results of the same analysis comparing v5 to v13, grouped according to whether and in what direction the indicator's revision has led to systematic changes.

[21] Such greater synchronization of changes can occur in a variety of ways that do not necessarily involve any bias (e.g., increased diffusion of democratic institutions). Conversely, no increase in synchronization may be visible if there are universal changes equal in magnitude but opposite in direction to the hypothesized bias.

[22] L&M make fairly strong and general claims about expert disagreement in the V–Dem data. However, their Appendix A.2 only shows results for the government media censorship effort, which they claim is a

several ways. First, in their assessment of V–Dem's corruption indicators, McMann et al. (2022) examine the correlates of V–Dem expert disagreement at the question-country-year observation level and find no evidence of "situational closeness," or the idea that experts will be biased in favor of countries that align with them ideologically. Specifically, experts with stronger allegiance to the liberal principle of democracy are *not* more likely to rate liberal countries as less corrupt, and experts who support the free market are *not* more likely to rate free market economies as less corrupt. Second, in Appendix B, we examine patterns in expert disagreement for two V–Dem expert-coded indicators: a low-subjectivity indicator and a (particularly) high-subjectivity indicator. We find no evidence of systematic expert disagreement for the low-subjectivity indicator, whereas expert disagreement on the high-subjectivity indicator is greatest for countries and years that are more recent, have higher freedom of expression, lower levels of democracy, and more experts coding them. Though we hasten to note that expert disagreement is, at best, a weak signal of ideological bias, our analysis demonstrates that only highly subjective V–Dem indicators are vulnerable to systematic expert disagreement.

Summing up the various analyses in this section, we fail to find any empirical evidence of L&M's proposed general (increase in) pessimism bias in V–Dem's expert-coded data. We must therefore look elsewhere to account for differences in estimated global backsliding trends between work that uses V–Dem indicators and L&M's analysis. In the next section, therefore, we discuss possible issues with the L&M indicators and index.

"representative variable."

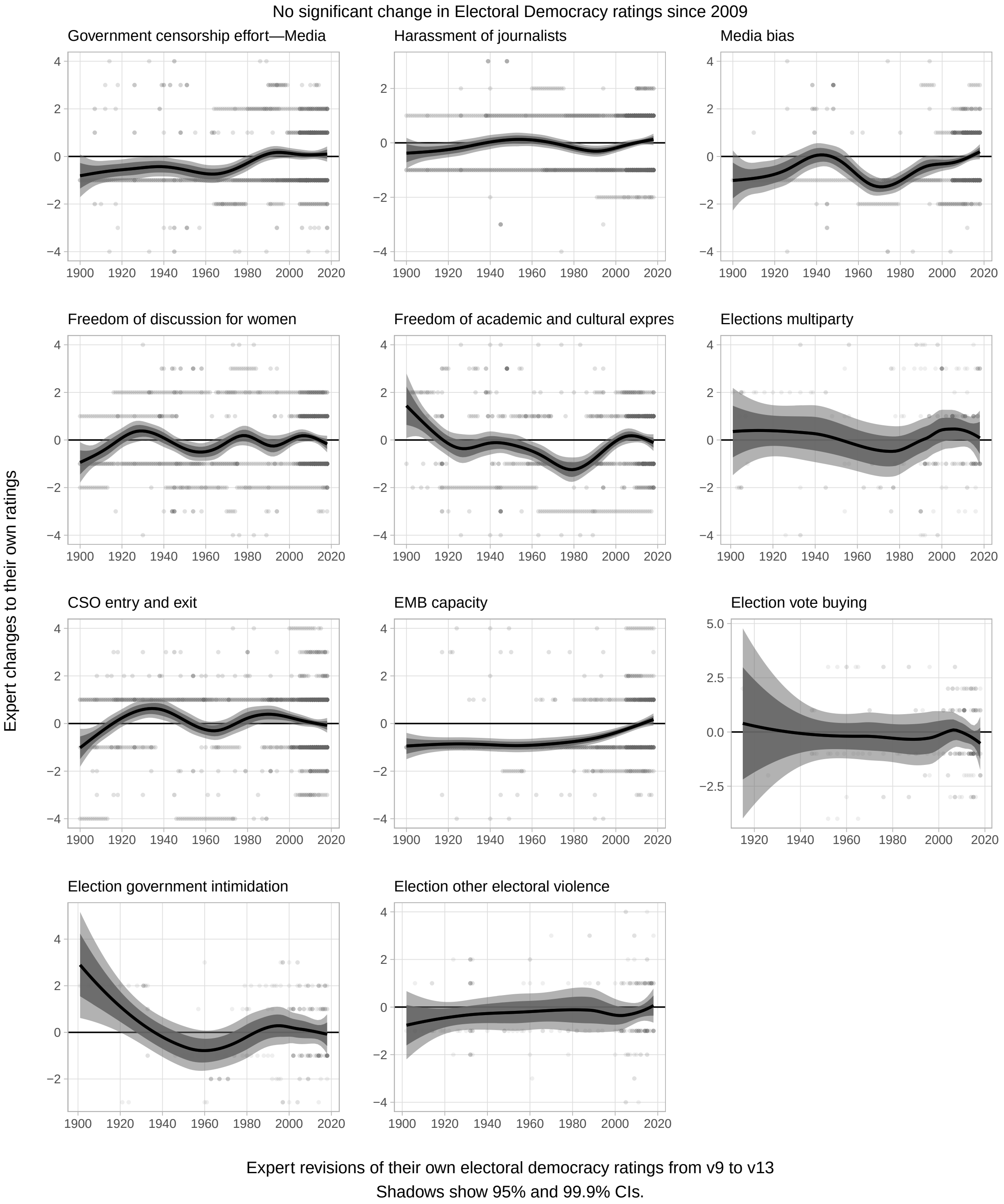

Figure 4: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have not experienced any systematic change after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

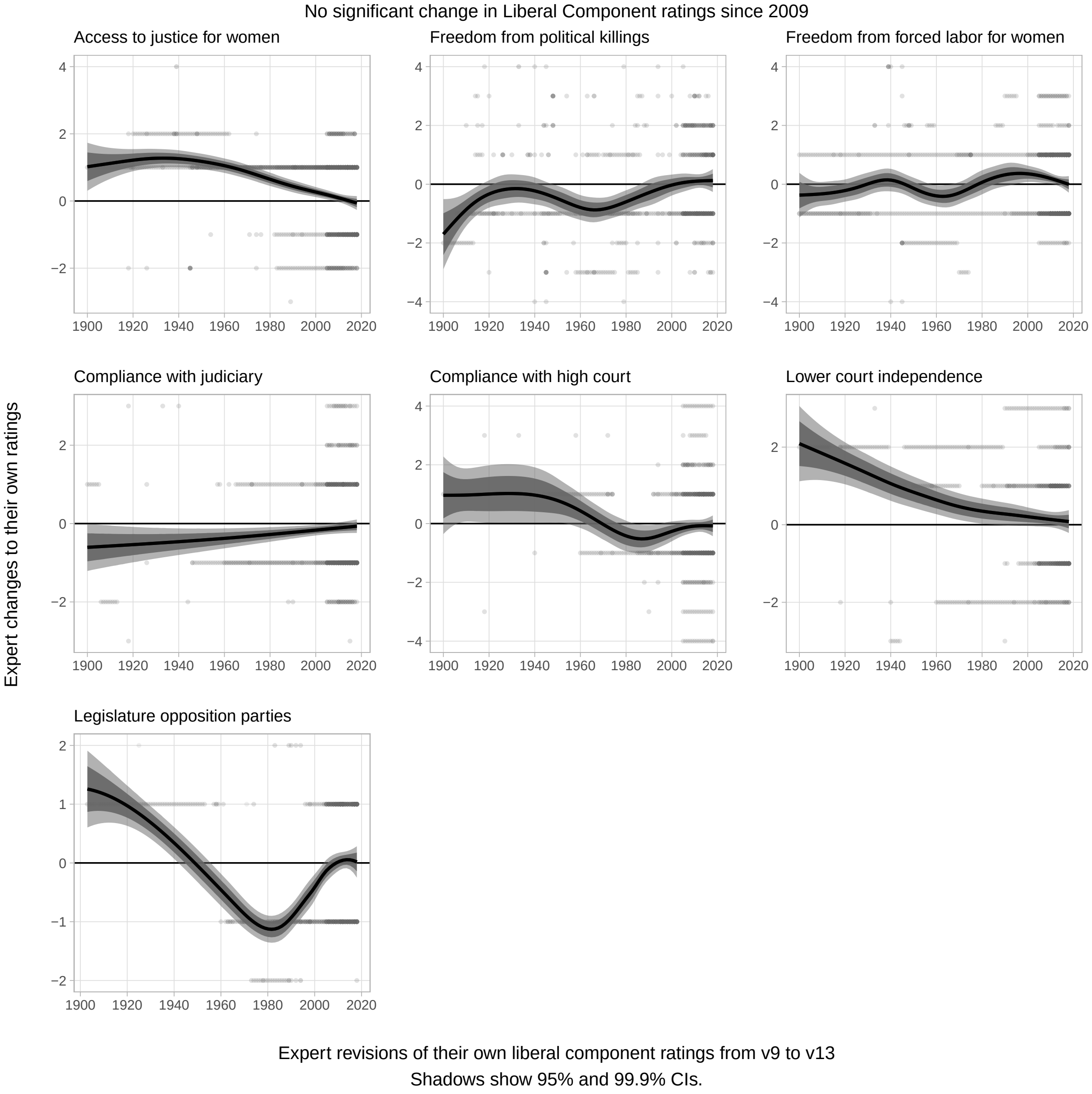

Figure 5: Expert-coded indicators entering V–Dem's Liberal Component Index that have not experienced any systematic change after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

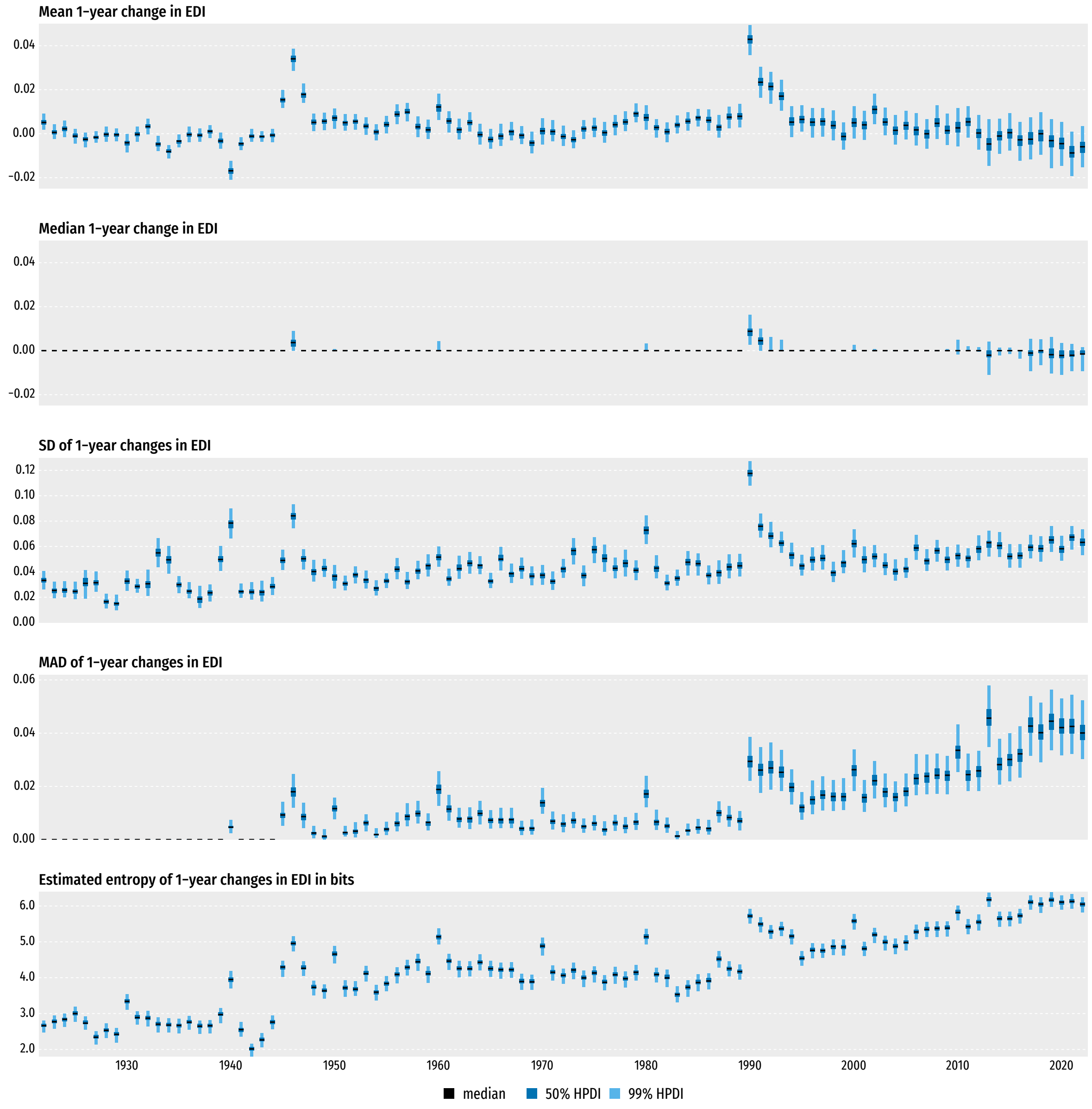

Figure 6: 1-year changes in EDI: mean, median, standard deviation from the mean, median absolute deviation from the median, and estimated entropy in bits, with highest posterior density intervals.

4 Evaluating L&M's Democracy Indicators and Index

The specific index that L&M create to measure global trends in democracy exemplifies many of the concerns with uncritical use of "objective" democracy indicators that we highlight in sections 1 and 2. This section discusses how their selection, measurement, and aggregation of indicators reflect theoretically untenable assumptions about how democracy works, leading to serious validity issues.[23]

As Table 1 illustrates, L&M's Objective index uses 12 indicators from several different sources: proportion suffrage (V–Dem); the proportion of presidential vote shares, the proportion of incumbent party seat shares, incumbent party time in office (truncated at 20 years), and seven-point indices of legislative and executive competitiveness (DPI); a dichotomous indicator of incumbent party loss in the previous election and two L&M-indices on multiparty competition and process violations (NELDA); and, finally, dichotomous indicators of term limits, succession rules, and dismissal rules (Meng, 2020). L&M scale these variables so they are bounded by 0 and 1 and have the same directionality (higher scores are more democratic). They then take the (unweighted) average of all 12 indicators. Missing data are treated as missing at random; i.e., the index value is determined by the average across whatever indicators are not missing for a given country-year.

We will focus on the implications of three key decisions embedded in L&M's measurement strategy: 1) the conceptualization of democracy reflected by the chosen variables, 2) the way that indicators are aggregated, and 3) decisions about how to deal with missing data. All three decisions are theoretically hard to defend and cumulatively result in problematic measurement. Equally importantly, issues 1 and 3 point to problems with interpreting even simple descriptive summaries of the individual indicators. Consequently, neither the aggregate results nor the analyses of individual indicators provide reasonable benchmarks for assessing democratic backsliding.

4.1 Indicators In the L&M Index

The main inclusion criteria for indicators in the L&M index seem to be that they are available, "objective," and measure some aspect of democracy. Before discussing their aggregation schema, we detail issues with the chosen indicators that affect analyses when using them individually (as reflections of democracy) and as part of a wider democracy index.

First, a quarter of L&M's indicators are dichotomous measures of executive constraints from Meng (2020). Meng collected these data to study how autocracies move from personalized to institutionalized rule and, as such, these indicators have an unclear theoretical relationship with many common democracy concepts and weak or otherwise

[23] Our analysis relies on a replication package provided by L&M (Little, 2023); we replicated their results before proceeding with our analyses.

Table 1: Indicators in L&M index

	Source	Concept	Type	% Missing
Proportion suffrage	V–Dem	Suffrage	Proportion	0%
Presidential vote	DPI	Exec. comp.	Proportion	65%
Incumbent party seat	DPI	Leg. comp.	Proportion	25%
Incumbent party in office	DPI	Exec. comp.	Years (max 20)	35%
Legislative competitiveness	DPI	Leg. comp.	7pt scale	20%
Executive competitiveness	DPI	Exec. comp.	7pt scale	20%
Incumbent party loss	NELDA	Elec. comp.	Dichotomous	16%
Multiparty competition	L&M/NELDA	Elec. comp.	4pt scale	6%
Process violations	L&M/NELDA	Elec. comp.	4pt scale	6%
Term limits	Meng	Exec. constraints	Dichotomous	24%
Succession rules	Meng	Exec. constraints	Dichotomous	26%
Dismissal rules	Meng	Exec. constraints	Dichotomous	26%

Note: All indicators scaled to 0-1; index is unweighted average.

problematic empirical relationships with common democracy measures. Figure 7 illustrates this concern using a common measure of electoral democracy, namely V–Dem's EDI (Polyarchy). While the presence of both term limits and dismissal rules correlate with higher EDI scores, two aspects of these boxplots should give pause: 1) there is a very large range in EDI scores for countries with these constraints in place, indicating that they are mainly suited for differentiating highly autocratic countries from all other countries; 2) there are numerous very democratic outliers that do not have such constraints.[24] Furthermore, Figure 7 indicates that succession rules are ill-suited for differentiating countries according to levels of democracy, given the substantially overlapping EDI distributions of countries with and without such rules.

The ordered categorical variables that L&M use to measure democracy are more clearly related to common concepts of democracy. All four variables measure aspects of electoral competition, a fundamental part of democracy (according to various definitions). However, the specific operationalizations of these concepts create various measurement issues. In particular, the categories within each variable scale onto the concept of democracy differently. For example, in the seven-point executive and legislative competitiveness indices from DPI, the difference between having a score of "1" vs. "2" is not identical to the difference between having a score of "6" vs. "7" with regard to the concept of democracy.[25]

Figure 8 illustrates this problem using the DPI codebook entry for the legislative index of electoral competitiveness (Cruz, Keefer and Scartascini, 2021). The first five categories all relate to highly uncompetitive situations in which opposition parties won no seats. And, even the top category (7) represents a very low bar for legislative competition (the

[24] Many high-quality parliamentary democracies (with strong political parties) do not have term limits, suggesting that term limits are not universally relevant to democratic-ness.

[25] Although the DPI indices have seven categories, there are additional categories between levels (e.g., 5.5) when DPI could not determine the correct value (e.g., 5 vs. 6).

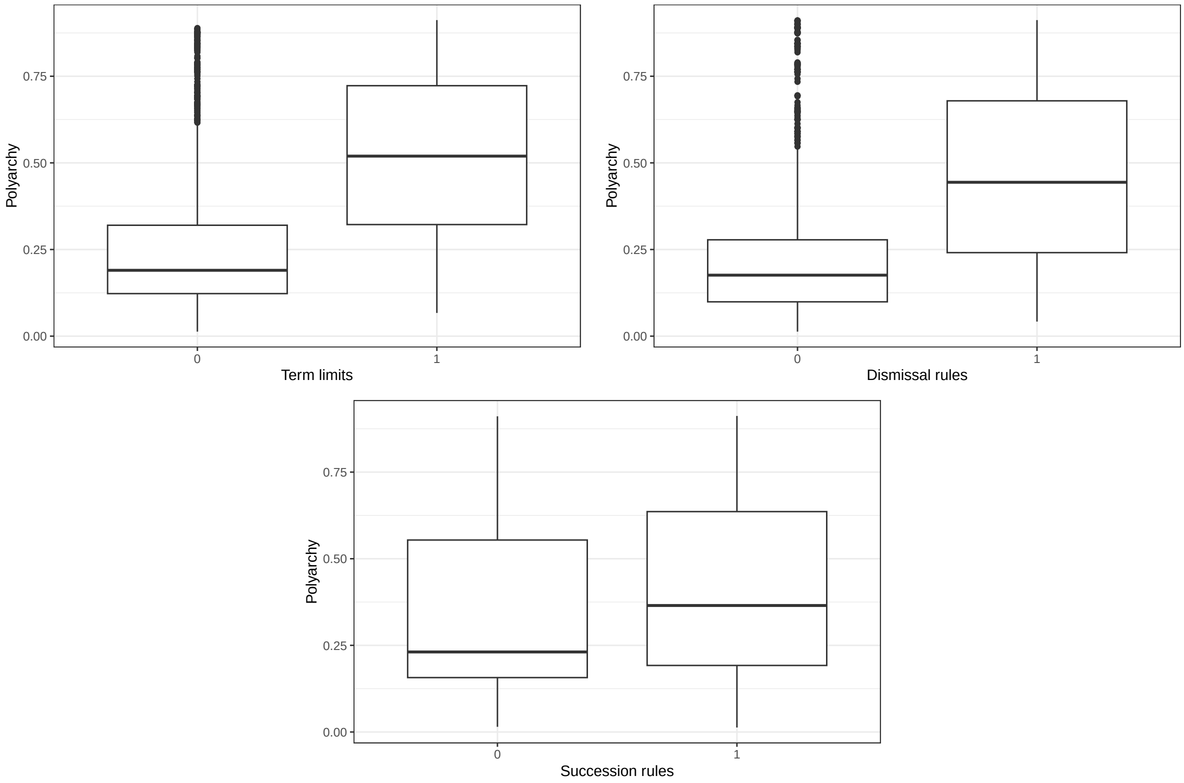

Figure 7: Relationship between executive constraints data and EDI

largest party won less than 75% of the seats). Accordingly, and as the upper-left cell of Figure 9 illustrates, there is limited variation in level of democracy between countries with values between 1 and 5.5; countries with scores of 6 and 6.5 have substantially higher scores, while countries with scores of 7 have even higher. The DPI executive index (upper right cell) shows similar issues, having "cut points" between values of 4 and 5 and then 6 and 7.

These issues extend to L&M's NELDA-derived indices (bottom row). For example, the multiparty elections index contains three dichotomous indicators: whether 1) opposition parties were allowed to contest the election, 2) opposition parties were legal, and 3) there was a choice of candidates on the ballot (Hyde and Marinov, 2019). A score of 0 represents a situation where none of these conditions were met, and a score of 1 is where they all were. The bottom left cell indicates that this index scales irregularly onto democracy: countries with none or one of these conditions (values 0 or 1) show similar levels of EDI, and countries with two conditions (2) have only slightly higher EDI scores. In contrast, countries with all four conditions in place (1) have much higher EDI scores. Similarly, the process violations index shows two apparent cut points with a moderate increase in EDI between 0 and 1 and a large increase between 2 and 3.

As the preceding descriptions demonstrate, all of these indicators have noisy relationships with EDI. Insofar as we consider the latter a high-face-validity measure of democracy, the former indicators seem to lack the reliability and sensitivity necessary to capture even

1. No legislature
2. Unelected legislature
3. Elected, 1 candidate
4. 1 party, multiple candidates
5. Multiple parties are legal but only one party won seats
6. Multiple parties DID win seats but the largest party received more than 75% of the seats
7. Largest party got less than 75%

Figure 8: DPI legislative index of electoral competitiveness

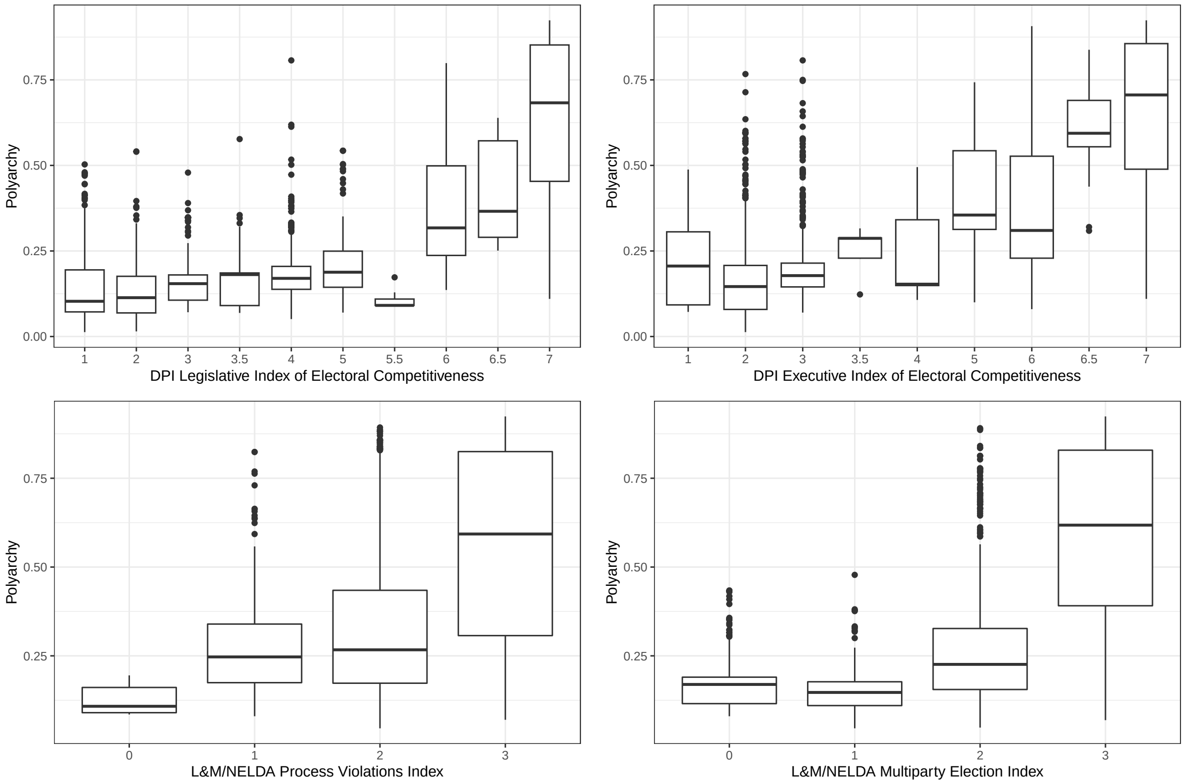

Figure 9: Relationship between ordered categorical data and EDI

middling changes in democratic quality. Moreover, none of these indicators can safely be treated as even approximately interval-valued, making L&M's analyses of backsliding, which generally rely on global *averages* of these measures (or their aggregates) and rest on this assumption, problematic. Figure 9 also illustrates that the category with the greatest spread in democracy values across all four indices is the top category. Cumulatively, this pattern in L&M's categorical indicators means that the overall index will have a *very low threshold* for considering a case to be highly democratic. As a result, would-be democratic backsliders have substantial room to maneuver before they receive a substantially lower democracy score.

Figure 10 presents scatterplots of the relationship between L&M's four continuous indicators of democracy and V–Dem's EDI. While one of these indicators (suffrage) is conceptually clearly related to democracy, it does not distinguish well between countries in the present era since most countries, even autocracies, have universal adult suffrage. The other three continuous indicators—party years in power, incumbent vote share, and incumbent seat share—have conceptual and empirical issues in L&M's use. 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 reason 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 variation in levels of democracy for all values of all three variables, indicating that the correlation is incredibly noisy.[26]

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 little sense, at the country level, as a direct indicator of democracy.[27] If an incumbent wins an election, this automatically results in a *minimum* .08 drop in a country's score on the L&M democracy index.

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

[27] This variable makes more sense in L&M's analysis of global trends since, on average, a greater proportion of incumbent losses across countries is evidence of greater electoral competition. Similarly, within a single country, this value can help predict democracy when aggregated over time. It might also play a role in a non-linear predictive model.

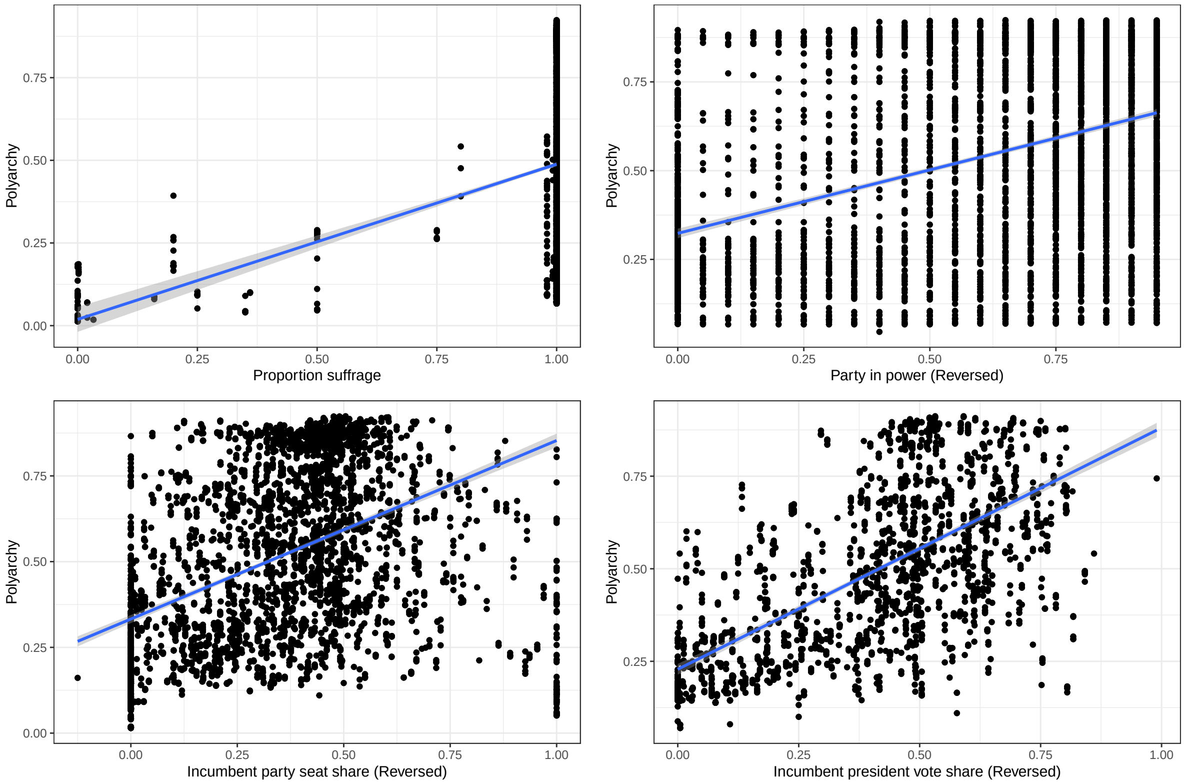

Figure 10: Relationship between L&M's continuous indicators and EDI

4.2 Aggregation

The variables L&M include in their index reflect a very specific conceptualization of democracy, which L&M neither explicitly discuss nor justify. We formalize this conceptualization as follows:

$$\text{Democracy} = \frac{\text{Exec. Constraints}}{4} + \frac{\text{Competition}}{2} \times \left(\frac{\text{General}}{2} + \frac{\text{Exec.}}{2} + \frac{\text{Legis.}}{3}\right) + \frac{\text{Suffrage}}{12}$$

Even if we accept the premise that an index of democracy does not require data on more difficult-to-measure concepts (e.g., media freedom), this conceptualization involves several debatable elements. In particular, by taking the unweighted average, L&M assume that all included variables have the same relationship with the underlying concept ("democracy").[28]

While developing a consensus on the "correct" theoretically-derived weights to elements of the L&M index would be a contentious endeavor, the lack of weights is itself only defensible in its mathematical simplicity. Scholars may debate the relative importance of term limits to democracy, but it is difficult to argue that the absence of term limits is

[28] L&M run robustness checks where they assign random weights to different indicators in the index construction. This check does not demonstrate that aggregation technique is irrelevant, but rather that their findings are robust to equally arbitrary decisions.

conceptually as important for democracy scores as having 0% suffrage. These situations are empirically equivalent in L&M's index. Moreover, by including three indicators of executive constraints in their index, L&M weights this concept highly (i.e., three times as much as suffrage), despite the potentially limited relevance of these indicators to otherwise more democratic societies.

Moreover, even if different indicators are equally important to democracy, they may still scale onto the concept differently. For instance, having high levels of suffrage may be more relevant to differentiating between countries with low levels of democracy than between those with high levels. Similarly, all previously-discussed issues with ordered categorical variables are directly relevant for index aggregation. For example, including the DPI legislative competitiveness index directly in a measure of democracy has several debatable implications. To illustrate, the difference between having an unelected legislature vs. an elected legislature has the same weight on democracy score as the difference between only one party winning seats and that party winning 75-99% of the seats. While deciding which differences in the DPI index should have what weight would be very difficult, making no decision about such weights is itself a decision with very difficult-to-justify implications.

4.3 Missingness in the Data

By treating missingness as "as-if" random in their index construction and descriptive analyses, L&M introduce two problems. Most generally, there is a great deal of missingness in the data they use: the median country-year observation in L&M's index has two missing indicators; 43% are missing a quarter or more of the indicators, and 11% are missing half or more. In other words, many country-year scores only represent the average of a subset of L&M's indicators, which means that these scores do not necessarily represent even their conceptualization of democracy. The second problem is that data are not missing at random but are systematically missing from countries with low values on certain indicators. For example, countries that do not hold national-level elections have missing data on, e.g., electoral win shares. As a result, these countries receive erroneously high scores because they only have data for indicators on which they perform well.

The missingness in these data is problematic, even if one eschews aggregation altogether. Missing data make it difficult to use each indicator—in isolation—to draw meaningful conclusions about global trends in democracy over the last decade, especially when one describes such trends using global averages. The left-most panel in Figure 11 presents results from regressing V–Dem's EDI on dummy variables coding missingness in each of the L&M measures (except for the V–Dem suffrage measure, which has similar coverage as EDI). These coefficients are mostly negative and statistically significant: less democratic country-years are more often missing. While this selection issue is concerning, even more worrying is that similar trends appear when regressing one and five-year *changes* in

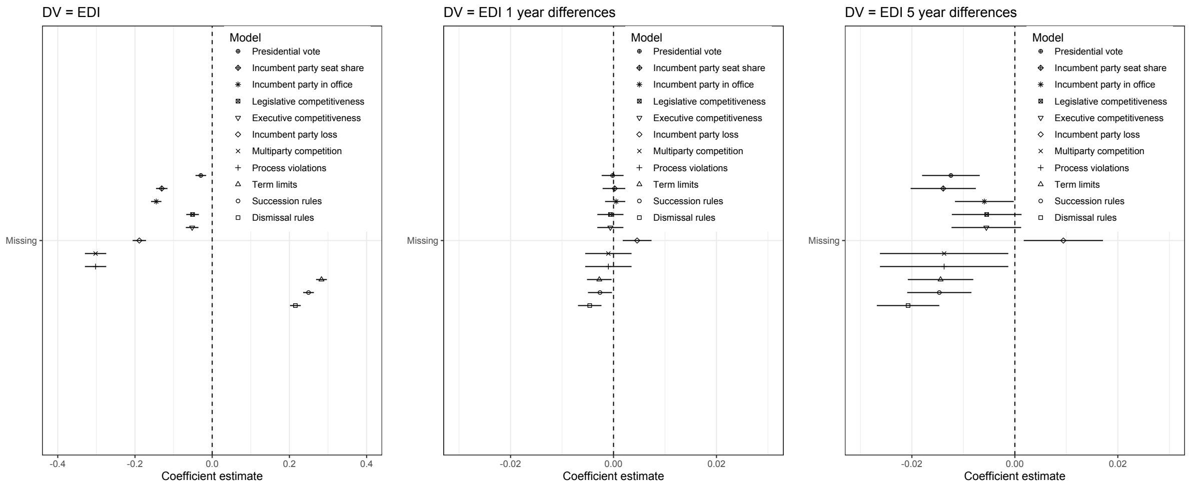

Figure 11: Results of regressing EDI levels and differences on indicators of missingness for each L&M indicator

EDI on missingness indicators for each L&M measure (see two rightmost panels, Figure 11). Overall, missingness in L&M's indicators predicts backsliding on EDI, especially for 5-year changes. In other words, the measures that L&M rely on for their analysis are systematically biased against finding backsliding; their descriptive analyses present world average scores for measures that have been pruned of country-years that are, on average, more likely to exhibit such backsliding.

Figure 12 further illustrates the patterns unearthed by these regressions. Each panel in the figure 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), and 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. Therefore, much of the recent downward action in the EDI world average is 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 13 illustrates another problem with the L&M approach to missingness: it creates face validity problems when assessing country-year trends at the index level. For example, across all country-years (1980-2020) all of L&M's indicators save four (suffrage, term limits, succession rules, and dismissal rules) are missing for China. For the period 1982-2017, the L&M index therefore reports China as a perfect democracy (score of 1) because the country scores 1 on all four indicators for which it has data.

Turkmenistan illustrates another issue with this method for dealing with missingness. In 1991, the L&M index only includes data from one indicator: proportion suffrage, which has a value of one. As a result, Turkmenistan is a perfect democracy in 1991. In

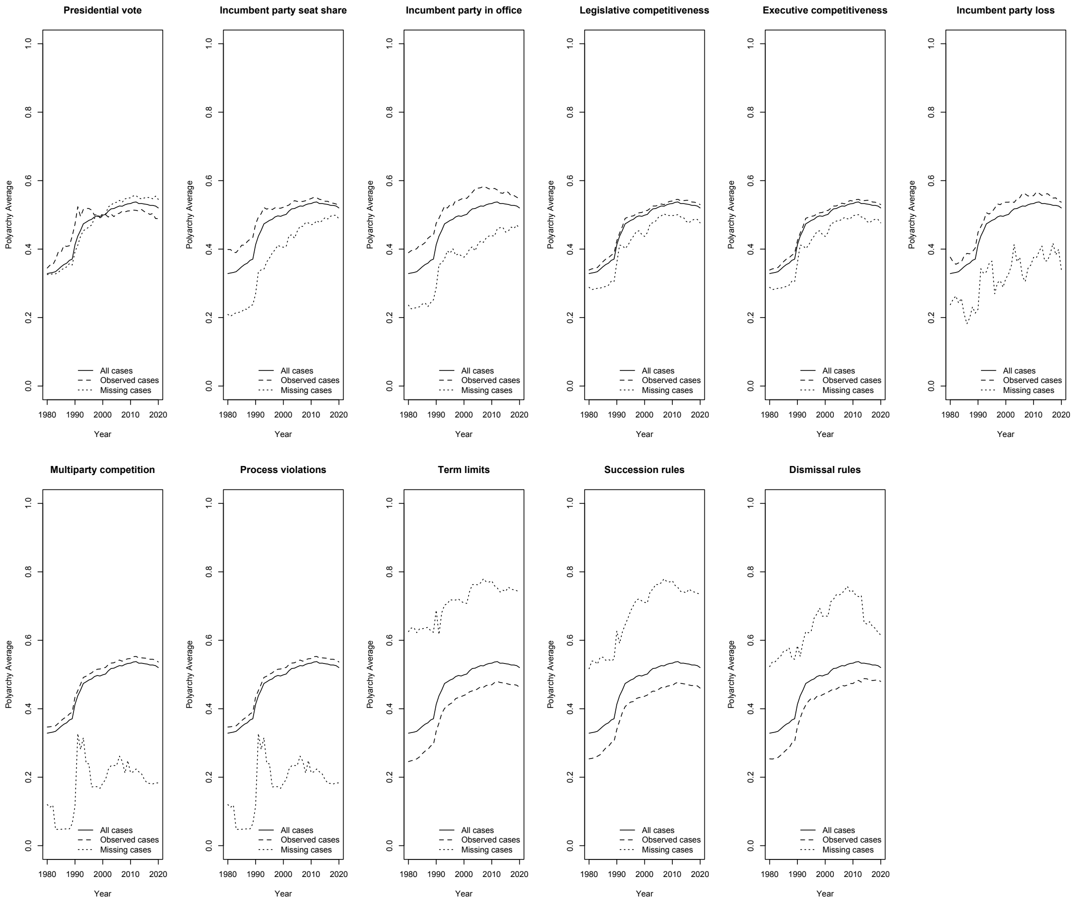

Figure 12: V–Dem's EDI world averages across L&M indicator case coverage

1992, when the L&M index includes data from all 12 indicators, Turkmenistan's score immediately drops to 0.59. This drastic change is due entirely to missing data suddenly being included, as opposed to political changes in the country.[29]

4.4 Implications of L&M's Coding Decisions

The implications of L&M's subjective measurement decisions are best illustrated visually. As Figure 1 in the introduction illustrates, the relationship between this index and V–Dem's EDI is especially weak for countries with moderate and high levels of democracy; scores also diverge for some important cases (Brazil, Denmark, North Korea, and the United States). In this section, we discuss how L&M's coding decisions led to these divergences, using both these cases and an additional four in Figure 13.[30]

First, highly autocratic Turkmenistan has very high scores on L&M's measure, much as North Korea in Figure 1. It has an average value of 0.56, and a minimum value of 0.44 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. Similarly, Equatorial Guinea has a high score of 0.61 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 illustrate 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 illustrates important issues with L&M's coding of relatively stable democracies. Unlike the US, Turkmenistan, and China; Norway never becomes a perfect democracy in the L&M index because the data 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 in Figure 1, 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

[29] Equatorial Guinea provides a final example of how patterns of missingness can affect country-year trends. In 1980-2, the L&M index only contains data from four indicators: suffrage and the three executive constraint indices. From 1980-1981, all four had values of zero, resulting in an index-score of zero. In 1982, Equatorial Guinea adopted universal suffrage and succession rules, resulting in a score of .5. This score immediately dropped to 0.33 in 1983, due to the addition of three other NELDA indicators with relatively low scores.

[30] Appendix Figure D.2 shows trends for all countries that have received a perfect 1-score on L&M's index during the coding interval, highlighting that 1) L&M's index diverges substantially from other indices of democracy and 2) has much lower face validity in many notable cases.

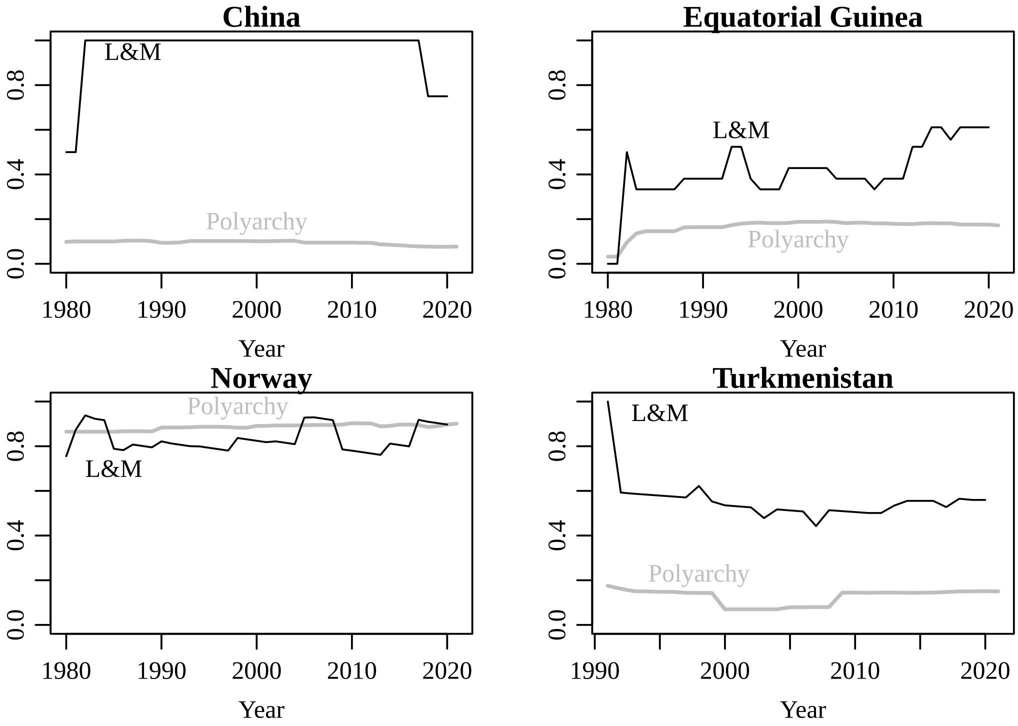

Figure 13: A selection of cases with substantively different assessments under EDI and L&M.

12 points return. This latter indicator actually explains all the variation in the United States scores. The US only includes 7 indicators, all of which universally have a perfect score of 1, except for the indicator of the 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.

Finally, Figure 14 highlights the key implication of these coding decisions for the study of backsliding, presenting the particular countries that contribute the most to the noted differences between EDI and L&M in measured (average) global democratic backsliding from 2010 to 2020.[31] In brief, there are two types of countries that contribute to the divergence: countries that show 1) little or no backsliding on the L&M index but substantial backsliding on EDI (e.g., Hungary, India, Poland, Turkey, and Venezuela); 2) either backsliding or no change in EDI, but where the L&M index finds clearly positive democratic change from 2010 to 2020 (e.g., Bolivia, Yemen, Brazil, Egypt, Sudan, Equatorial Guinea, and Denmark). For both groups of countries, changes in EDI have higher face validity than those on the L&M index.

We elaborate on the importance of face validity checks for evaluating democracy measures through discussions of two important cases: China and Turkey.

[31] Appendix C provides more details.

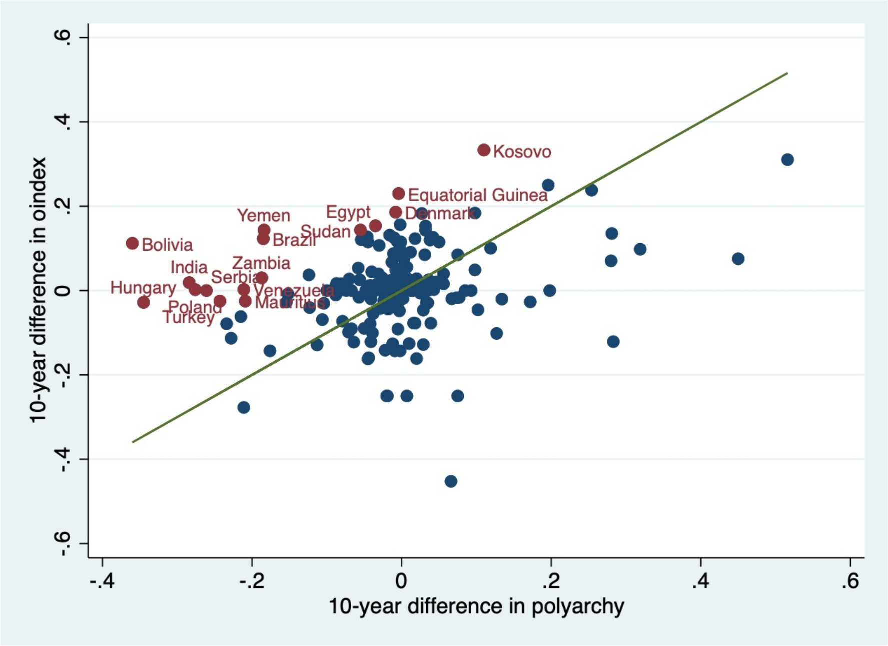

Figure 14: 10-year differences in V–Dem's EDI and Little and Meng's "Objective" index from 2010 to 2020

4.5 The Importance of Face Validity Checks

Given the complexities inherent in designing an unbiased set of measures to capture democracy and backsliding, we emphasize the importance of rudimentary face validity checks to any measurement enterprise. In brief, no amount of discussion regarding measurement principles in the abstract obviates the relevance of assessing how one's measure performs in practice, using other pieces of information, for example on country-specific developments, as points of reference. Specifically, there is valuable insight to be gained by looking at histograms, trends over time, and scatter plots of key measures—at both the country level and in aggregate. In this subsection, we demonstrate this point with two brief case studies.

Figure 15 shows trends since 1980 in L&M's democracy measure as well as V–Dem's EDI and Polity scores (re-scaled 0-1 for comparison) for China and Turkey. China is a relatively stable one-party autocracy (despite recent personalization under Xi Jinping, e.g., Shirk (2018)). The Chinese Communist Party controls all political offices, which are not subject to competitive elections. Thus, even with the most procedural definition of democracy, China should score very low for the period 1980 to 2020. By contrast, Turkey is a prominent case in the recent backsliding literature (e.g., Mechkova, Lührmann and Lindberg, 2017; Andersen, 2019; Cleary and Öztürk, 2022). Throughout the 1980s to early 2000s, Turkey engaged in a piecemeal reform process that addressed democratic deficits in its 1982 constitution, which had been established by a military junta that ruled from 1980-1983. Many of these reforms also involved changes to ordinary laws to limit military involvement in politics, protect individual freedoms, and ensure the rule of law

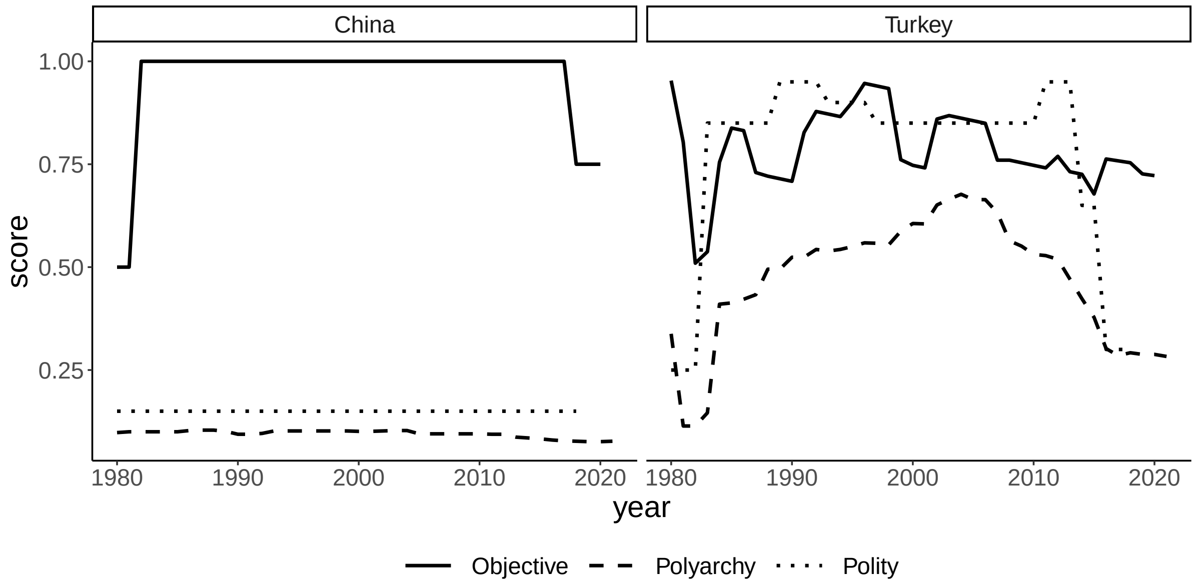

Figure 15: Trends in democracy scores for China and Turkey

(Özbudun, 2007). However, since the 2000s, Recep Tayyip Erdoğan has been accused of dismantling Turkey's democracy by repressing the opposition and pushing through changes that expanded the executive's power (Esen and Gumuscu, 2018). Thus, the two cases give us variation on what we should *expect* to see in terms of scores and trends in measures of (common definitions of) democracy.

As shown in Figure 15, according to L&M's objective index, China scored 0.50 out of 1.00 for 1980-1981 and then achieved a perfect 1.00 from 1982 to 2017, before it "backslid" to 0.75 in 2018 to 2020. These changes in values are marked solely by the implementation of executive term limits in the 1982 Chinese Constitution and their subsequent removal under Xi Jinping in 2018. More importantly, the L&M index departs drastically from the consistently low rating for China on both the V–Dem EDI and Polity index. Overall, the index lacks face validity for this case. For example, in 1989, the year of the Tiananmen Square massacre, China has a perfect score on the L&M index, whereas V–Dem's EDI is just 0.10 and the Polity score is 0.15 (or -7 on the original scale). Furthermore, the L&M index might overstate the degree to which China has "backslid" under Xi Jinping, with a drop of 25% of the scale due to the removal of term limits. By contrast, V–Dem's EDI registers a more modest decline from a peak of 0.10 in the 1990s and early 2000s to 0.08 by 2020.

The trends for Turkey on the L&M index show much more fluctuation around an average of 0.76. Oddly, Turkey's highest score – 0.95 – is in 1980, the same year that the armed forces staged a military coup and established a junta that ruled by decree under martial law for the next three years. While the L&M index does decline to 0.80 in 1981, this score still appears much higher than one would expect for a military dictatorship and is *higher* than Turkey's scores from 1987-1990, 1999-2001, and since 2007. In other words, the last military junta in Turkey is deemed more democratic than most of the years after it returned to civilian rule and embraced widespread democratic reforms (Özbudun, 2000; Altunisik, 2005). Furthermore, L&M's index appears to follow no recognizable pattern since the 1980s; although there is a slight downward trend under Erdoğan. By contrast, the Polity scores show a large improvement after the end of the military junta in 1983, with some subsequent improvements coinciding with reforms until a rapid decline under Erdoğan. V–Dem's EDI is less optimistic about the 1982 Constitution; it shows a more gradual upward trend throughout the 1980s, 1990s, and 2000s, in line with the piecemeal reform process that unfolded during this period. And, also in line with expectations based on the Turkish politics literature, EDI shows a gradual deterioration of democracy under Erdoğan.

5 Conclusion

In this piece, we engage with the question: how can we best measure (backsliding in) the state of global democracy? We first argue that such measurement requires careful conceptualization of democracy and democratic backsliding, the selection of indicators to fully operationalize these concepts, and an evaluation of bias across data sources (i.e., documents, analysts, experts), across cases (i.e., countries), and over time, in these indicators.

Most scholars operate with democracy concepts that are not adequately captured by a handful of indicators. This is true also for scholars defining democracy mainly as "electoral democracy;" contested multi-party elections require not only the presence of such elections but also that they are free and fair, that several political rights are guaranteed, that opposition parties are allowed to form, that there is a media environment informing voters of different policy position, etc. (see, e.g., Dahl, 1998). Hence, we need a broad set of measures to adequately assess the state of democracy in the world. We argue that the various indicators and indices from the V–Dem project therefore provide effective measures for studying the state of global democracy.

We also emphasize that there is no such thing as a truly "objective" indicator of democracy, in the sense of being free of human judgment. In the absence of this unicorn, seemingly objective indicators can be even more problematic than their subjective counterparts. In particular, the lack of indicator availability and spotty conceptual breadth as well as restricted case coverage limit their value for the task of measuring democracy across the world. Moreover, if one tries to apply such measures to assess global democratic backsliding, doing so still requires a series of subjective decisions that require justification. While one might be able to create a sensitive, theoretically appropriate, and empirically effective strategy for assessing democratic backsliding using only low-subjectivity indicators (see Weitzel et al. (2023)), one should not blindly trust that a given analysis built around several such indicators is more trustworthy than one drawing on a broader set of indicators, including more subjective ones.

Yet, it is important to analyze the potential for different measurement errors and biases, including those in expert-coded indicators. Building on an extensive body of work from scholars engaged in democracy measurement, we discussed existing and new analyses of V–Dem's expert-coded indicators. We conclude that there is no compelling evidence of large and systematic bias across experts, countries, or time that could drive observed trends in global democracy, including the kind of "bad vibes bias" hypothesized by L&M.

Because we find the V–Dem measures most appropriate for the task of monitoring global democratic quality, we agree with the research building on these measures that finds evidence for global, multidimensional, multi-year democratic backsliding in the last decade. However, we emphasize that the validity of this conclusion—and especially conclusions about the extent of global democratic backsliding —relies on a series of interconnected critical decisions about conceptualization, measurement, and data collection. Drawing a conclusion based on an under-specified conceptualization of democracy, incomplete operationalization of the concept, systematically biased data, or problematic aggregation procedures will result in under-specified, incomplete, biased, and problematic conclusions about democratic backsliding. These latter points pertain, more specifically, to Little and Meng's (2023) proposed objective indicators and index of democracy, which we have discussed and analyzed in detail. We believe that the many and severe validity issues with their proposed democracy measure make it difficult to treat it either as a valid benchmark for describing global democracy trends or as yardstick for evaluating very different measures of democracy, such as those constructed by V–Dem.

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95:529–546.

Alizada, Nasifa et al. 2022. "V-Dem Institute Democracy Report 2022: Autocratization Changing Nature?" Gothenburg: V-Dem Institute. Report.

Altunisik, Meliha Benli. 2005. "The Turkish model and democratization in the Middle East." *Arab Studies Quarterly* pp. 45–63.

Andersen, David. 2019. "Comparative democratization and democratic backsliding: The case for a historical-institutional approach." *Comparative Politics* 51(4):645–663.

Beetham, David. 1999. *Democracy and Human Rights.* London: Polity Press.

Bermeo, Nancy. 2016. "On Democratic Backsliding." *Journal of Democracy* 27:5–19.

Boix, Carles, Michael Miller and Sebastian Rosato. 2012. "A Complete Data Set of Political Regimes, 1800-2007." *Comparative Political Studies* 46(12):1523–1554.

Cheibub, Jose, Jennifer Gandhi and James Vreeland. 2010. "Democracy and dictatorship revisited." *Public Choice* 143(1–2):67–101.

Cleary, Matthew R and Aykut Öztürk. 2022. "When does backsliding lead to breakdown? Uncertainty and opposition strategies in democracies at risk." *Perspectives on Politics* 20(1):205–221.

Coppedge, Michael et al. 2011. "Defining and Measuring Democracy: A new approach." *Perspectives on Politics* 9(2):247–267.

Coppedge, Michael et al. 2020. *Varieties of Democracy: Measuring Two Centuries of Political Change.* Cambridge: Cambridge University Press.

Coppedge, Michael et al. 2023*a*. "Varieties of Democracy (V-Dem) Codebook." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael et al. 2023*b*. "Varieties of Democracy (V-Dem) Dataset v.13." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael et al. 2023*c*. "Varieties of Democracy (V-Dem) Methodology." Varieties of Democracy (V-Dem) Project.

Cruz, Cesi, Philip Keefer and Carlos Scartascini. 2021. "DPI2020 Database of political institutions 2020: Changes and variable definitions." *Washington, DC: Inter-American Development Bank* .

Dahl, Robert A. 1998. *On Democracy.* New Haven: Yale University Press.

Esen, Berk and Sebnem Gumuscu. 2018. "The perils of "Turkish presidentialism"." *Review of Middle East Studies* 52(1):43–53.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide.* Princeton: Princeton University Press.

Haggard, Stephan and Robert Kaufman. 2021. *Backsliding: Democratic regress in the contemporary world.* Cambridge University Press.

Hyde, Susan D and Nikolay Marinov. 2012. "Which elections can be lost?" *Political analysis* 20(2):191–210.

Hyde, Susan D and Nikolay Marinov. 2019. "Codebook for national elections across democracy and autocracy (NELDA) dataset." 5:1–40.

Knutsen, Carl Henrik and Svend-Erik Skaaning. 2022. The Ups and Downs of Democracy: 1789-2018. In *Why Democracies Develop and Decline*, ed. Michaen Coppedge, Amanda Edgell, Carl Henrik Knutsen and Staffan I. Lindberg. Cambridge: Cambridge University Press.

Knutsen, Carl Henrik and Tore Wig. 2015. "Government Turnover and the Effects of Regime Type: How Requiring Alternation in Power Biases Against the Estimated Economic Benefits of Democracy." *Comparative Political Studies* 48(7):882–914.

Knutsen, Carl Henrik et al. 2019. "Introducing the Historical Varieties of Democracy Dataset: Patterns and Determinants of Democratization in the Long 19th Century." *Journal of Peace Research* 56(3):440–451.

Levisky, Steven and Daniel Ziblatt. 2018. *How Democracies Die.* New York: Crown.

Little, Andrew. 2023. "Replication Data for: Measuring Democratic Backsliding.".
URL: *https://doi.org/10.7910/DVN/G2SQ6Y*

Little, Andrew T and Anne Meng. 2023. "Measuring Democratic Backsliding." *OSF Preprints* .
URL: *osf.io/n32zk*

Lührmann, Anna, Marcus Tannenberg and Staffan I. Lindberg. 2018. "Regimes of the World (RoW): Opening New Avenues for the Comparative Study of Political Regimes." *Politics and Governance* 6(1):60–77.

Lührmann, Anna and Staffan I. Lindberg. 2019. "A third wave of autocratization is here: what is new about it?" *Democratization* 26:1095–1113.

Maerz, Seraphine F., Amanda B. Edgell, Matthew C. Wilson, Sebastian Hellmeier and Staffan I. Lindberg. Forthcoming. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *Journal of Peace Research* .

Marquardt, Kyle L. 2020. "How and how much does expert error matter? Implications for quantitative peace research." *Journal of Peace Research* 57(6):692–700.

Marquardt, Kyle L. and Daniel Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26(4):431–456.

Marquardt, Kyle L. and Daniel Pemstein. 2023. "Estimating Latent Traits from Expert Surveys: An Analysis of Sensitivity to Data Generating Process." *Political Science Research & Methods* 11(2):384–393.

Marquardt, Kyle L, Daniel Pemstein, Brigitte Seim and Yi-ting Wang. 2019. "What makes experts reliable? Expert reliability and the estimation of latent traits." *Research & Politics* 6(4).

Marquardt, Kyle L., Daniel Pemstein, Constanza Sanhueza Petrarca, Brigitte Seim, Steven Lloyd Wilson, Michael Bernhard, Michael Coppedge and Staffan I. Lindberg. 2017. "Experts, Coders, and Crowds: An analysis of substitutability." *V-Dem Working Paper* 53.

McMann, Kelly et al. 2022. "Assessing data quality: An approach and an application." *Political Analysis* 30(3):426–449.

Mechkova, Valeriya, Anna Lührmann and Staffan I Lindberg. 2017. "How much democratic backsliding?" *Journal of democracy* 28(4):162–169.

Medzihorsky, Juraj. 2017. "Election Fraud: A Latent Class Framework for Digit-Based Tests." *Political Analysis* 23:506–517.

Meng, Anne. 2020. *Constraining dictatorship: From personalized rule to institutionalized regimes*. Cambridge: Cambridge University Press.

Miller, Michael K. 2015. "Democratic Pieces: Autocratic Elections and Democratic Development since 1815." *British Journal of Political Science* 45(3):501–530.

Munck, Gerardo L. and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35:5–34.

Özbudun, Ergun. 2000. *Contemporary Turkish politics: Challenges to democratic consolidation*. Lynne Rienner Publishers.

Özbudun, Ergun. 2007. "Democratization reforms in Turkey, 1993–2004." *Turkish Studies* 8(2):179–196.

Pelke, Lars and Aurel Croissant. 2021. "Conceptualizing and measuring autocratization episodes." *Swiss Political Science Review* 27(2):434–448.

Pemstein, Dan, Eitan Tzelgov and Yi-ting Wang. 2015. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." Gothenburg: V-Dem Working Paper No. 1.

Pemstein, Dan et al. 2023. "The Varieties of Democracy Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." Gothenburg: V-Dem Working Paper No. 21.

Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub and Fernando Limongi. 2000. *Democracy and Development. Political Institutions and Well-Being in the World, 1950–1990.* Cambridge: Cambridge University Press.

Schedler, Andreas. 2012. "Judgment and measurement in political science." *Perspectives on politics* 10(1):21–36.

Shirk, Susan L. 2018. "China in Xi's" new era": The return to personalistic rule." *Journal of democracy* 29(2):22–36.

Teorell, Jan, Michael Coppedge, Staffan Lindberg and Svend-Erik Skaaning. 2019. "Measuring Polyarchy Across the Globe, 1900-2017." *Studies in Comparative International Development* 54(1):71–95.

Treisman, Daniel. 2023. "How Great is the Current Danger to Democracy? Assessing the Risk With Historical Data." *Comparative Political Studies* OnlineFirst:1–29.

Waldner, David and Ellen Lust. 2018. "Unwelcome Change: Coming to Terms with Democratic Backsliding." *Annual Review of Political Science* 21:93–113.

Weidmann, Nils. 2023. "Recent Events and the Coding of Cross-national Indicators." *Comparative Political Studies* Forthcoming.

Weitzel, Daniel, John Gerring, Daniel Pemstein and Svend-Erik Skaaning. 2023. "Measuring Electoral Democracy with Observables.".
URL: *osf.io/preprints/socarxiv/9kzja*

Appendices

A Do country experts "pessimistically" update their scores?

As discussed in Section 3.2 of the paper, if the bias that Little and Meng (2023) 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 expert-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 (EDI) 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 and 5 in the paper as well as Appendix Figures A.4-A.7 do the same for the v9 to v13 comparison. Considering 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 Little and Meng (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 EDI. 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 EDI or the Liberal Component Index (LCI).

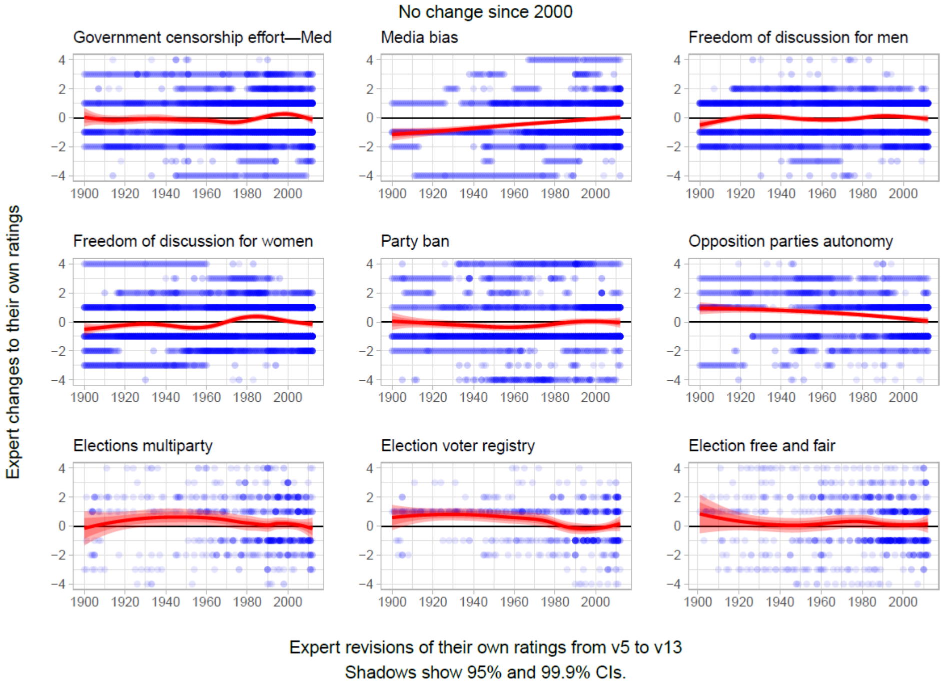

Figure A.1: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have not experienced any systematic change after year 2000 due to expert revisions (from V–Dem v.5 to v.13).

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 EDI 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 EDI.

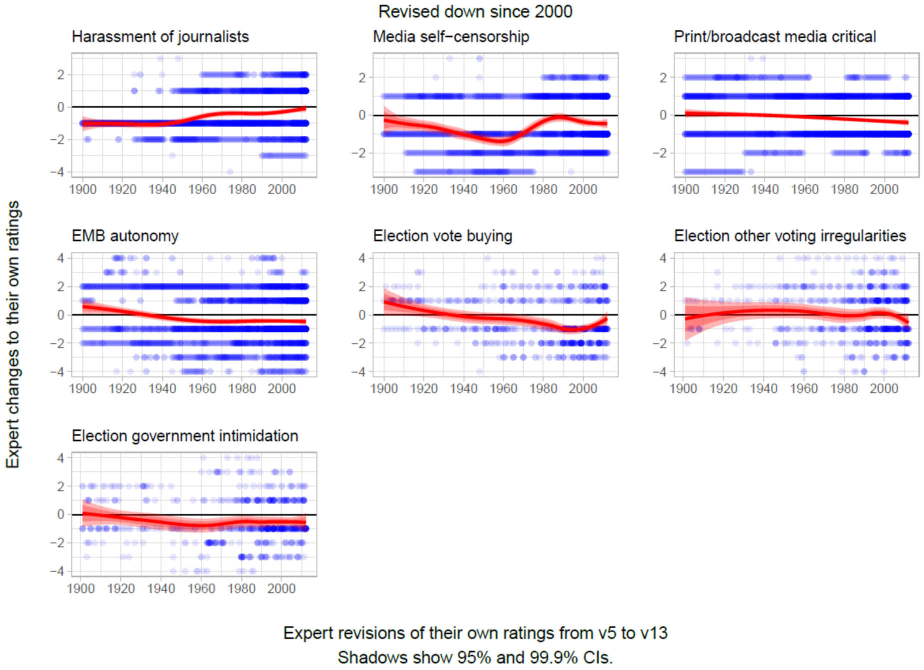

Figure A.2: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have have been revised down after year 2000 due to expert revisions (from V–Dem v.5 to v.13).

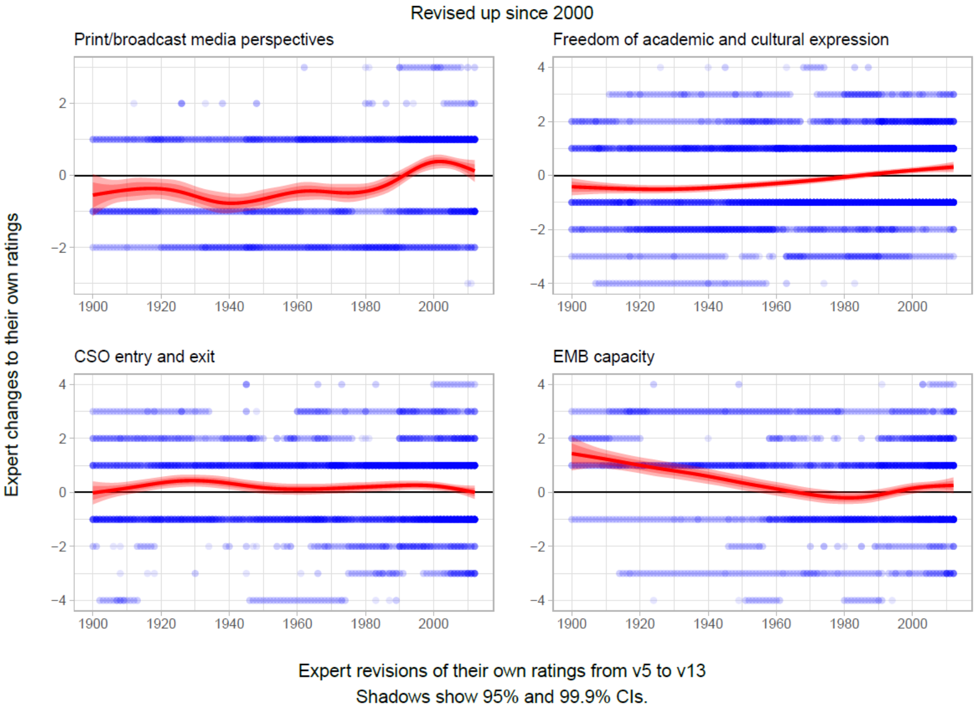

Figure A.3: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have have been revised up after year 2000 due to expert revisions (from V–Dem v.5 to v.13).

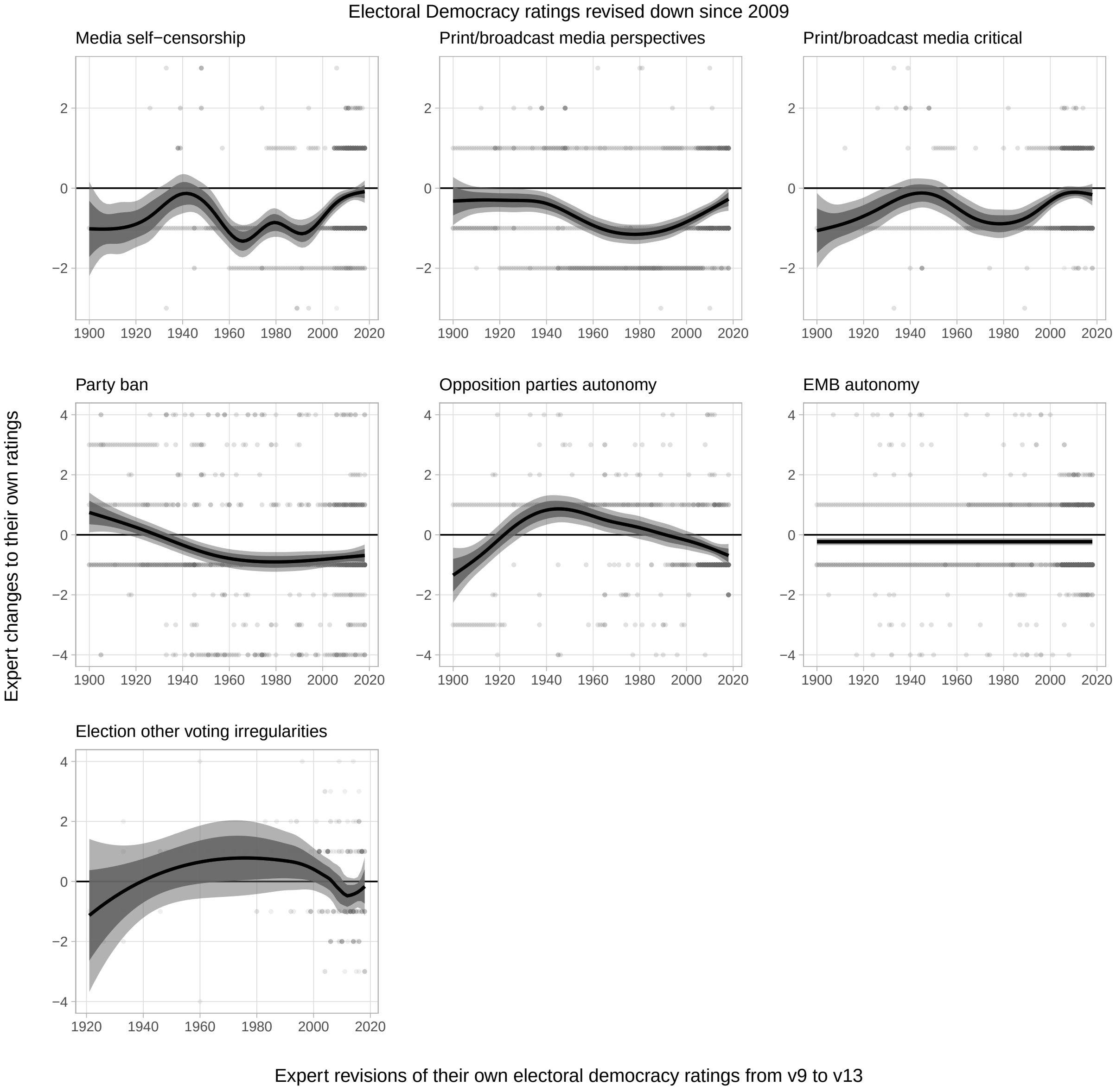

Figure A.4: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have have been revised down after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

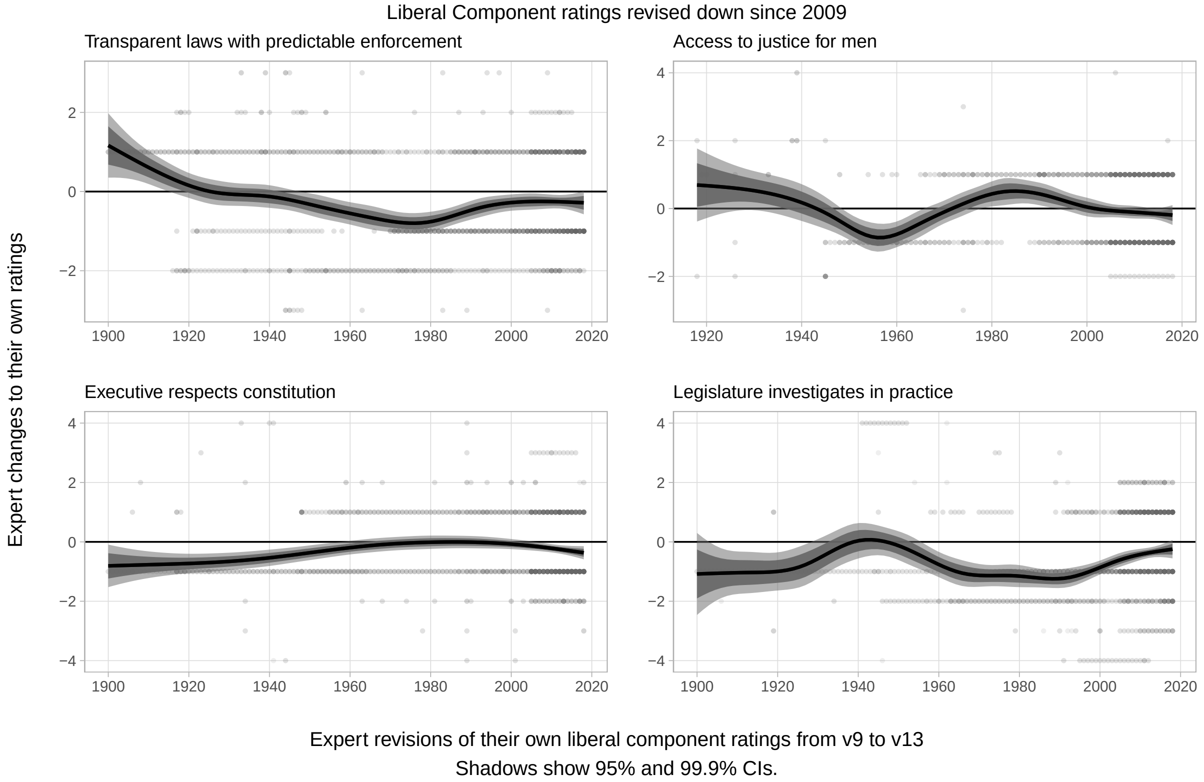

Figure A.5: Expert-coded indicators entering V–Dem's Liberal Component Index that have have been revised down after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

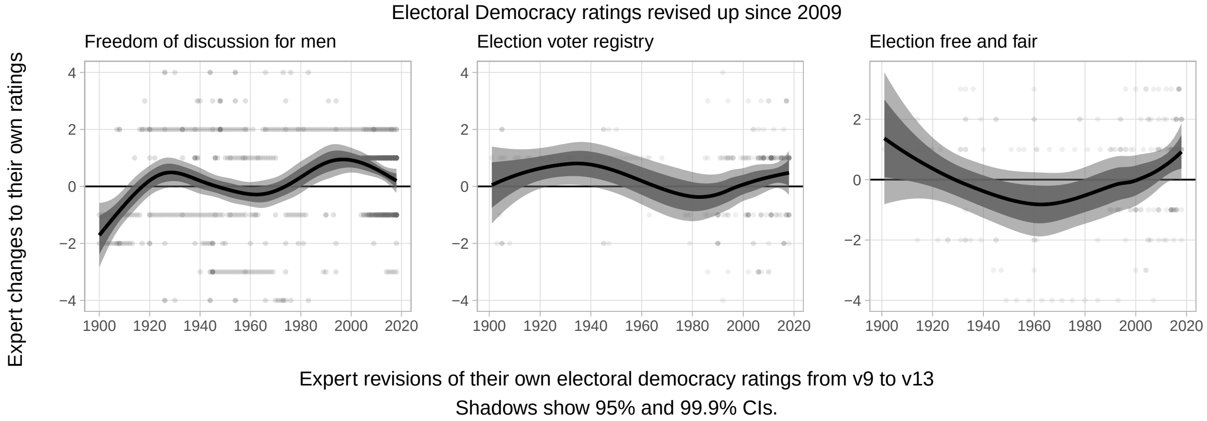

Figure A.6: Expert-coded indicators entering V–Dem's Electoral Democracy Index that have have been revised up after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

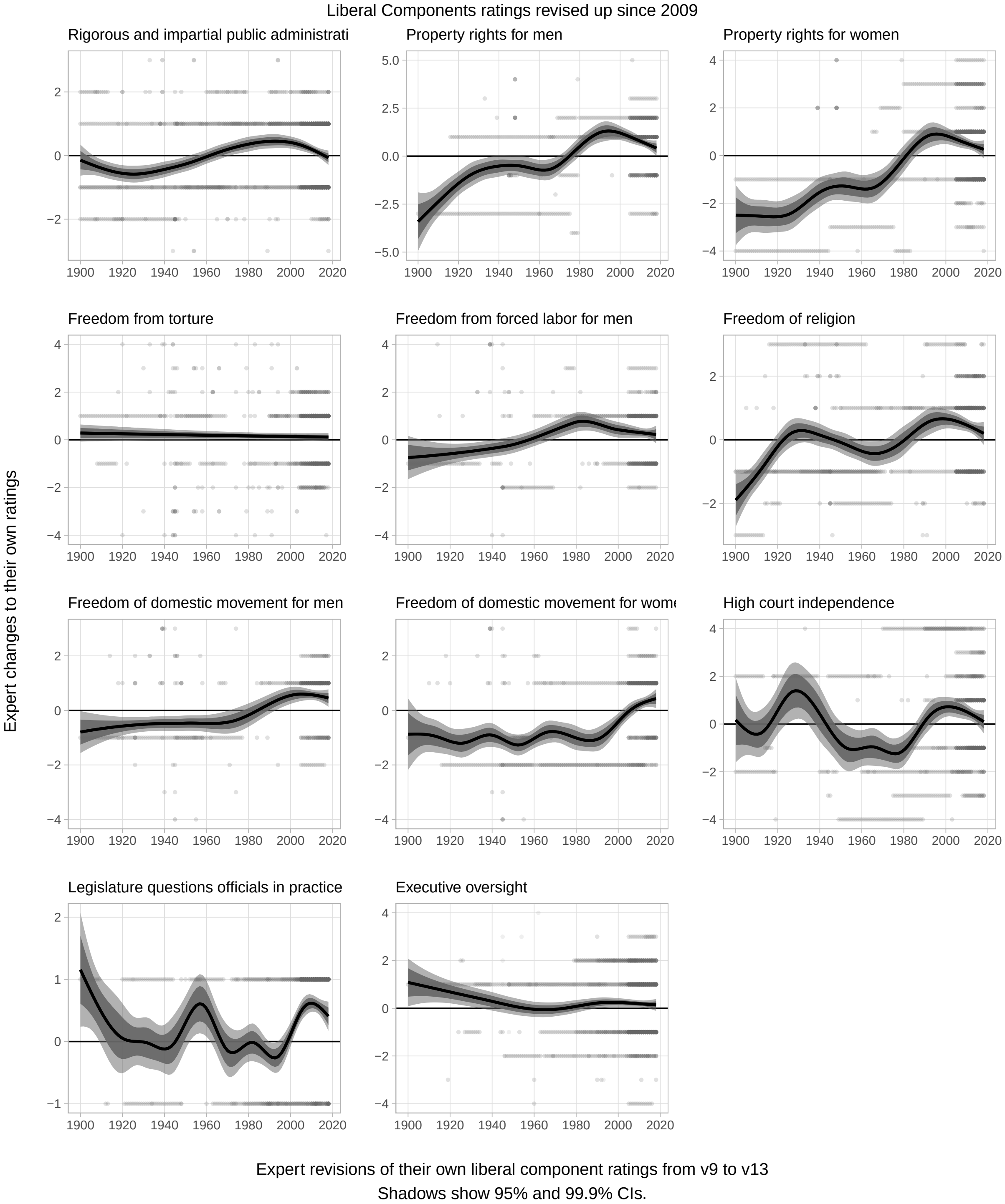

Figure A.7: Expert-coded indicators entering V–Dem's Liberal Component Index that have have been revised up after year 2000 due to expert revisions (from V–Dem v.9 to v.13).

B Relative question subjectivity and expert disagreement

In this section, we discuss the analysis summarized in Section 3.2 of the article. Contrary to what L&M argue, we emphasize that all democracy measures involve subjectivity, and that there is a spectrum of subjectivity even within V–Dem's indicators. We consider whether relative subjectivity is associated with more systematic expert disagreement. To do so, we assess patterns in expert disagreement for two expert-coded ("C") V–Dem indicators with different levels of subjectivity: 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 experts 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 - and broad - question. The latter asks experts 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. We do not claim that either of these variables are "representative" of high- or low-subjectivity questions in V–Dem more generally; rather, these questions simply provide a compelling contrast.

For each variable, we regress expert disagreement (measured using the standard deviation of expert ratings at the country-year level) on country and expert traits: the century of coding; level of freedom of expression; level of democracy; and number of experts. Table B.1 reveals that none of these variables predict expert disagreement for the low-subjectivity variable, *v2elboycot*. In contrast, expert 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 experts.

The findings, presented in Table B.1, suggest that expert 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 expert-level characteristics. Both (relatively) low-subjectivity and high-subjectivity questions are common among the expert-coded indicators in the V–Dem dataset (although we again stress that v2elfrefair is an outlier in its level of subjectivity). Our findings imply 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 affect all questions similarly.

Table B.1: Predicting Expert 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 Experts	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.

C Assessing recent backsliding across the world

Figure 14 in the paper plots the 10-year difference from 2010 to 2020 on L&M's democracy index on the y-axis and changes during the same interval for V–Dem's EDI on the x-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. Note that +.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 measure.

D Additional statistics regarding of L&M's democracy indicators

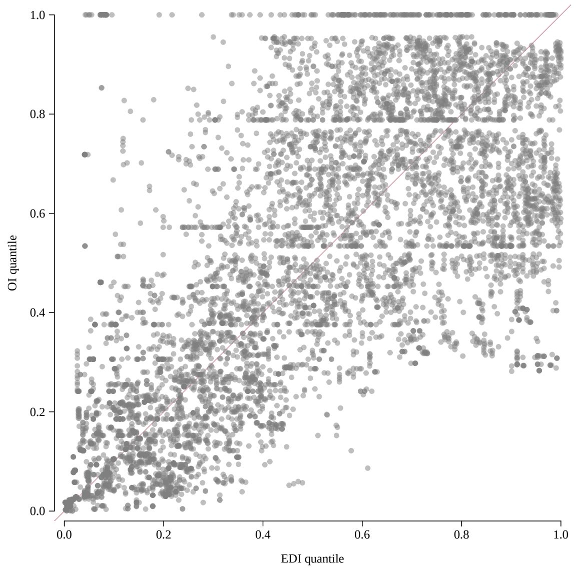

Figure D.1: Joint distribution of V–Dem's EDI and L&M's Index, 1980–2020

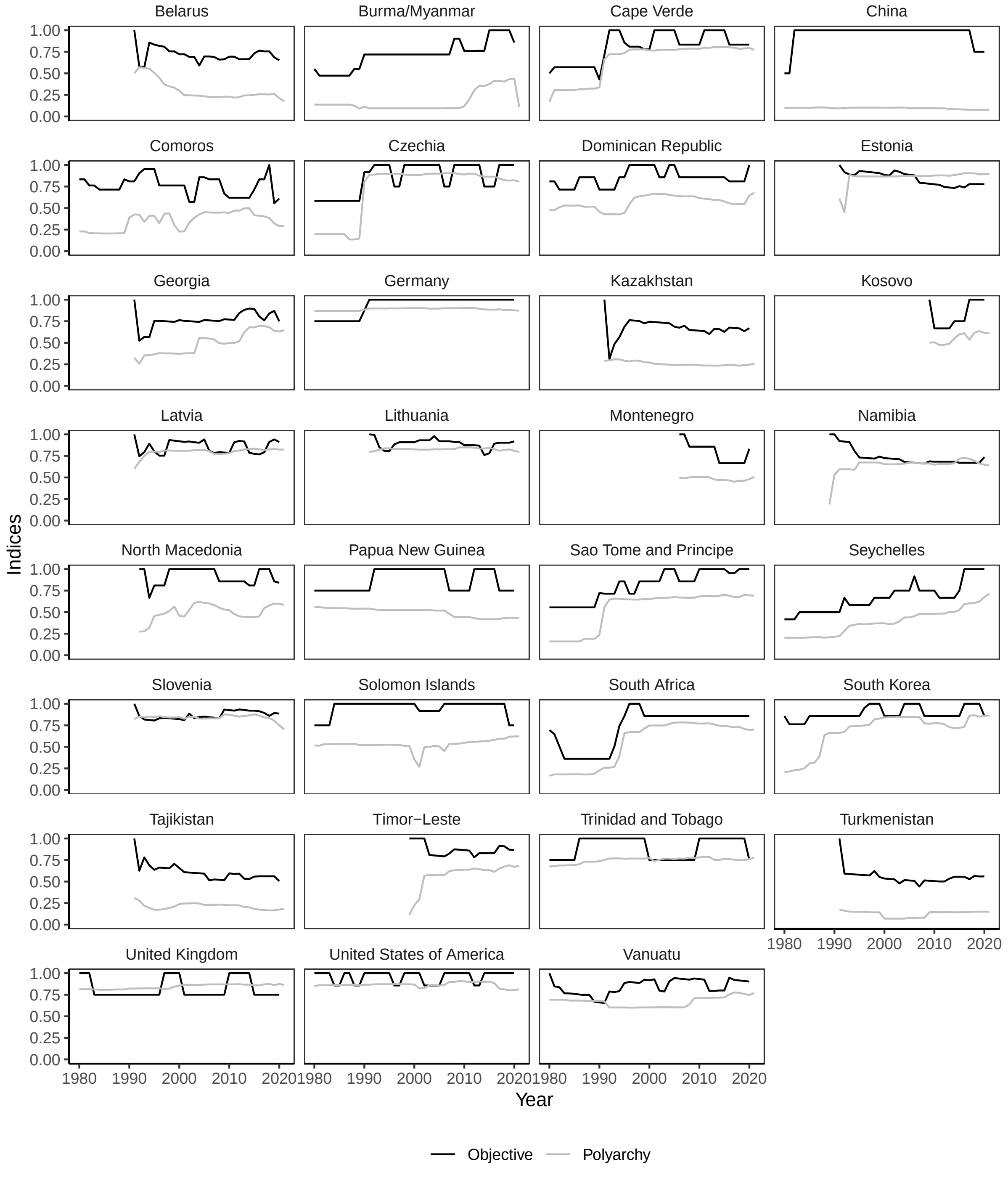

Figure D.2: Countries that earn a perfect score for at least one year on the L&M index