1. In "Measuring Democratic Backsliding," Little and Meng (2024) survey "objective" indicators of democracy and argue that trends in electoral competitiveness and executive constraints show that, "in general" (p. 150), there is little evidence of backsliding during the past decade. In other words, their argument is that, based on these measures, the global pattern is one of democratic stability or stagnation rather than democratic backsliding per se.

2. I plan to reproduce the authors' construction of a "simple aggregate objective index," which they created to compare trends in democracy among different sets of countries. By so doing, they sought to challenge the argument about global democratic backsliding. I will focus on Figure 8 in the article (p. 157).[1] The authors created this index by normalizing a list of individual variables between 0 and 1 and taking the average for each country-year (p. 156). The authors add a clarification note, according to which this simple index is meant only to summarize the aggregate trends described in earlier parts of their article and that it is not an appropriate substitute for existing democracy indices on a country-year level. Yet, both creating and analyzing the index serve the authors to draw conclusions regarding the current state of democracy worldwide.

3. The authors seek to challenge the conventional narrative that the world is in a period of global democratic decline. They argue that, in order to support this narrative, scholars need reliable ways to measure democracy. This leads to their distinction between objective and subjective measures of democracy. While the former are factual, such as whether the incumbent party loses and accepts defeat in an election, the latter rely on the judgment of expert coders to answer questions such as whether a particular election can be considered "free and fair" (p. 149). Indeed, the authors note that recent studies pointing at global democratic backsliding use subjective indicators almost entirely. According to the authors, this approach might be problematic, because the more it relies on human judgment, the more it can be systematically biased. In other words, it may well be that a perceived global

[1] Throughout the replication paper, I use the words "in the article" to emphasize when I am referring to the Figure presented in the article itself. When I do not use these words, I am referring to Figures that are presented here, in the replication paper.

decline of democracy might in fact be driven by changes in coder bias rather than actual changes in regime type on the ground.

Therefore, the authors collected data from arguably "objective" variables (as they define them) by country-year. They emphasize that, for simplicity and consistency with existing literature, they include all countries in the sample (when possible) for the main analysis and focus on average levels of variables (p. 151). Practically, they compare trends at the global level over time between the patterns revealed by objective variables and the patterns revealed by coder-based measures (V-Dem and Freedom House). As noted, their main analysis is based on the construction of a "simple aggregate objective index." The idea underlying their research design is rather simple: If the world is experiencing major backsliding in the aggregate, they expect to see empirical evidence of this phenomenon also in the objective measures data, but they fail to do so.

In sum, this research design helps them address rival explanations pointing at democratic backsliding. Their comparison between objective and subjective measures of democracy aims to corroborate or refute the core claim regarding a widespread trend of democratic backsliding.

4. In the first part of their article, the authors present data descriptions based on multiple objective variables. Later, in the main part of their article, they summarize these objective indicators and compare trends among different sets of countries. To that end, the authors construct a "simple aggregate objective index" by normalizing all individual variables between 0 and 1 (based on data availability, which is a key aspect I will discuss later on) and taking the average for each country-year. Specifically, those indicators include: (1) the average of percentage of the population with suffrage from V-Dem; (2) presidential vote shares; (3) winning-party seat shares; (4) incumbent-party time in office (limited to 20 years); (5 and 6) legislative and executive competitiveness; (7) whether the incumbent party lost the last election; (8 and 9) the multiparty index and the process-violations index from NELDA; and (10 to 12) the presence of term limits, succession rules, and dismissal rules. According to the authors, rather than to replace existing democracy indices on a country-year level, they chose to create this simple index only to summarize aggregate trends and to provide a general sense of how indicators related to the quality of democracy

have changed over time. Indeed, since the input variables are measured on different scales, the magnitude of the index does not have a clear interpretation (p. 156).

The main comparisons that I will focus on here are presented in Figure 8 in the article, which shows the trends of their aggregate objective index over the past 40 years in a thick line, compared to those of V-Dem polyarchy index score (thin black line) and Freedom House democracy score (the thin gray line). While the figure shows the objective index to be higher than V-Dem and Freedom House, this is not substantive by itself since, as noted above, it simply reflects the fact that many of the simple index's components have a "low bar" (for example, multiparty elections). Instead, the authors focus on the change over time, with particular attention to subtle changes in recent years and contemporary direction. For all three indices, the main change occurred at the end of the Cold War when there was a noticeable increase (many new democracies were created). Since 2000, however, trends have been different, when V-Dem and Freedom House modestly declined and the authors' objective index generally continued to increase. Moreover, in 2020, the objective index showed the highest value so far, while the other two lines show declines.

Because some scholars have argued that weighting by population is better for capturing the experience of the average citizen rather than the country (p. 156), the authors also include the middle and right panels of Figure 8 in the article which present the same trends, weighted by population with and without China and India, respectively. Both cases reveal a more negative trend for all three lines, albeit volatile. The authors note that, in the case of their objective index, the negative trend when weighted by population could be driven by specific elections or other sharp shifts on one of the underlying indicators. The right panel of Figure 8 in the article illustrates that the trend is much smoother when the two most populous countries, India and China (which have had noisy trends in recent years), are omitted. This somewhat weakens the decrease for Freedom House and V-Dem's polyarchy index scores and leads to a similar trend as the unweighted graph for the objective index.

In sum, the authors chose to use "objective" measures, and then to summarize them into an aggregate index, so they could test the claim regarding a widespread trend of democratic backsliding. They rationalized their decision to use averages by arguing that this is the common way of addressing trends of democracy around the world (p. 151). They

also chose to compare descriptions when the data are weighted by population size (with and without China and India), because this is another key approach taken in extant literature when it comes to trends of democracy performance (p. 156).

5. The authors do not report elements of statistical inference. It is clear from their objective, explanations, and approach why: They did not view elements of statistical inference as necessary for their analysis. Rather than addressing some kind of uncertainty in the data, in the article they focus on patterns over time at the aggregate level. Interestingly, V-Dem does include elements of statistical inference. As noted in its recent codebook, V-Dem uses statistical models to estimate underlying but unobserved concepts like those coder-based measures of democracy. V-Dem's models adjust for differences among experts, estimate uncertainty, and rely on iterative estimation procedures whose stability is checked using convergence diagnostics. Per this logic, when the models converge, it indicates that the estimated country scores are stable and reliable, meaning the final indicators are statistically inferred rather than directly observed.

   Yet, as the authors' goal was to measure variables that include only objective and factual data and to compare cases using these observed values, they viewed elements of statistical inference as unnecessary. More generally, though, one may argue that they could use elements of statistical inference. For example, to account for potential measurement error or data missingness in some of their variables and when they combined multiple indicators that may not be perfectly comparable across countries or time.

6. To create their aggregate index of objective measures of democracy, the authors merged multiple sources of observational data by country-year. It means that both their unit of observation and unit of analysis is county-year. Overall, they had 6,987 observations in total that encompass 179 countries for the years 1980 to 2021. However, as shown in Table 1, in many of their variables, there is a large pattern of missingness that might affect their results. Specifically, the authors used the following 12 variables to create their aggregate index (in parentheses I mention variable name; data source):

I. Percentage average of the population with suffrage (V-Dem suffrage; V-Dem).
II. Presidential vote shares (Press share; DPI).
III. Winning-party seat shares (Govt seat share; DPI).

IV. Incumbent-party time in office (truncated at 20 years) (Legislative index; DPI).
V. Legislative competitiveness (Legislative index; DPI).
VI. Executive competitiveness (Executive index; DPI).
VII. Whether the incumbent party lost the last election (Incumbent party lost; NELDA).
VIII. The multiparty index (Multiparty election; NELDA).
IX. The process-violations index (Electoral process; NELDA).
X. The presence of term limits (Term limits indicator; The authors' original data).
XI. The presence of succession rules (Succession indicator; The author's original data).
XII. The presence of dismissal rules (Executive dismissal indicator; The author's original data).

As shown in their replication code, the authors applied normalization only when a continuous variable had a scale other than 0 to 1, with binary indicators already coded as 0/1 left unchanged. Table 1 reveals patterns of missingness in the data. Most of the variables had more than 15% of missingness, with "Press share" even showing more than 65% of missingness. Because they use average, it means that for those country-year observations where the data is not available, the aggregate index is still computed, but it is based on fewer components. So, if and when some of the remaining components for that country-year are extreme in their value, results might change substantially.

The standard deviations show some interesting patterns as well. A substantial heterogeneity across country-years could be decentered for most index components, particularly for term limits, executive dismissal, and party competition. On the other hand, variables like suffrage and electoral process display relatively low dispersion, which reflects their widespread adoption today as well as their limited variation over time. The table thus suggests meaningful cross-regime variation in some of the data.[2]

[2] The negative minimum value in "Govt seat share" arises because it is computed as one minus the ratio of government seats to total seats, and in a small number of cases this ratio exceeds one due to differences in how government and total seats are recorded in the DPI dataset. This produces slightly negative values that simply indicate extreme legislative dominance.

Table 1: Objective index inputs: coverage, missingness, and summary statistics

Variable (index input)	Non-missing	Missing	% Missing	Mean	SD	Min	Max
Press share (reversed, 0–1)	2,427	4,560	65.3%	0.405	0.226	0.000	0.990
Party institutionalization (norm., 0–1)	4,559	2,428	34.8%	0.561	0.347	0.000	0.950
Executive dismissal indicator	5,065	1,922	27.5%	0.766	0.424	0.000	1.000
Succession indicator	5,187	1,800	25.8%	0.854	0.353	0.000	1.000
Govt seat share (reversed, 0–1)	5,212	1,775	25.4%	0.334	0.235	-0.125	1.000
Term limits indicator	5,326	1,661	23.8%	0.549	0.498	0.000	1.000
Executive index (0–1)	5,562	1,425	20.4%	0.794	0.290	0.000	1.000
Legislative index (0–1)	5,567	1,420	20.3%	0.833	0.270	0.000	1.000
Incumbent party lost	5,849	1,138	16.3%	0.289	0.453	0.000	1.000
Electoral process	6,590	397	5.7%	0.895	0.190	0.000	1.000
Multiparty election	6,590	397	5.7%	0.850	0.283	0.000	1.000
V-Dem suffrage (0–1)	6,987	0	0.0%	0.969	0.165	0.000	1.000

7. I attempted to replicate Figure 8 in the article. In order to do so, I had to, first, load all datasets and merge them into one dataset. Gladly, the authors included in their article a link to a replication code that could be loaded into R. Second, based on the merged-dataset, I could compute the aggregate index. This went well: I indeed managed to get the same results and graph. Below is a picture of Figure 8 from the article. Figure 1 is my own replication for Figure 8.

Figure 8

Unweighted and Weighted Average Indices

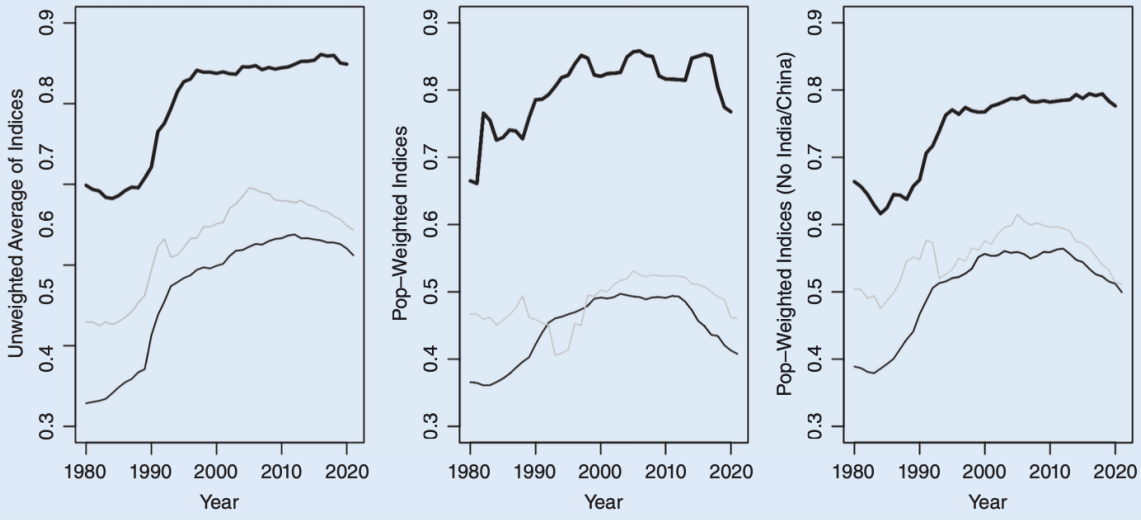
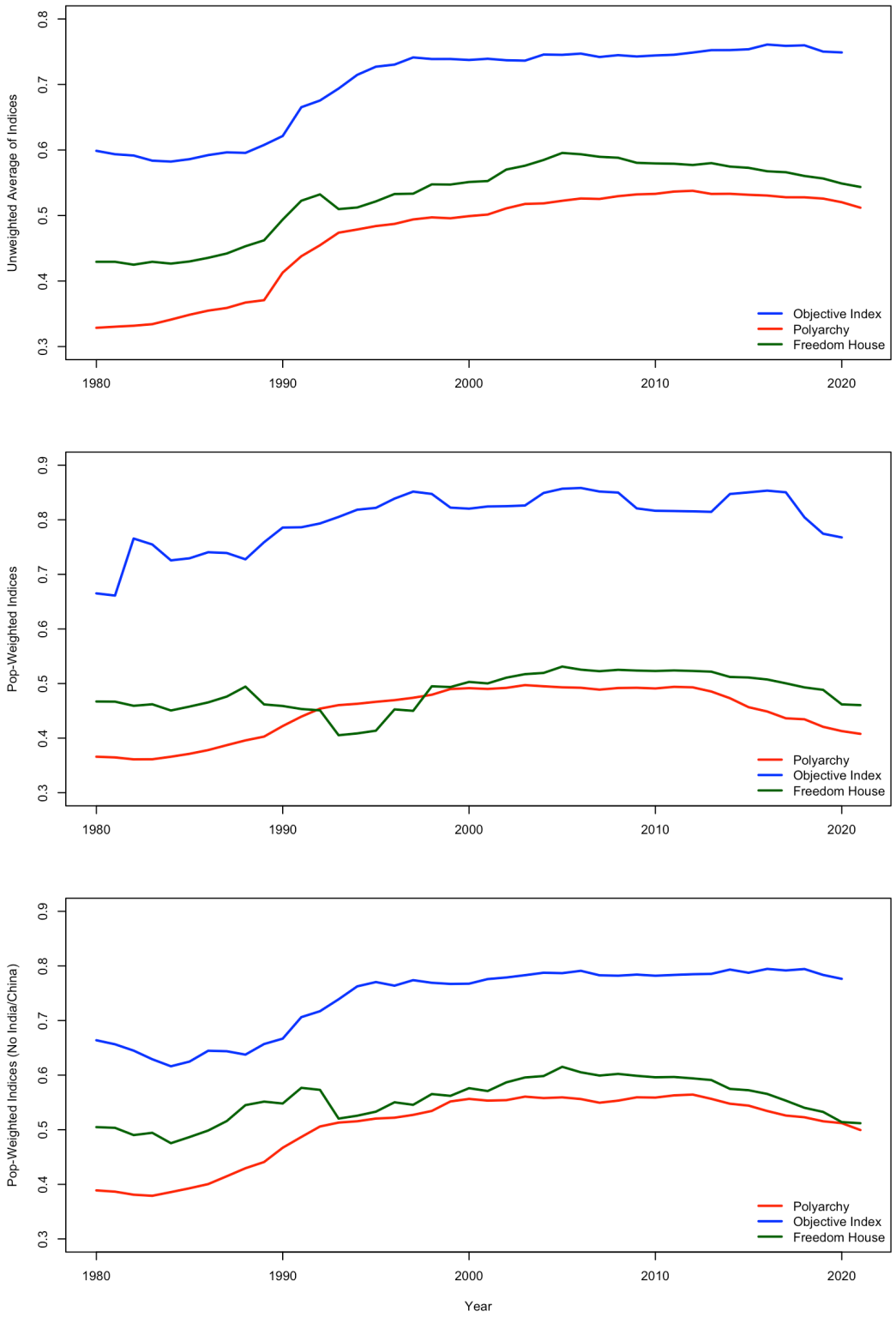


Figure 2: Replication of Figure 8 in the Article

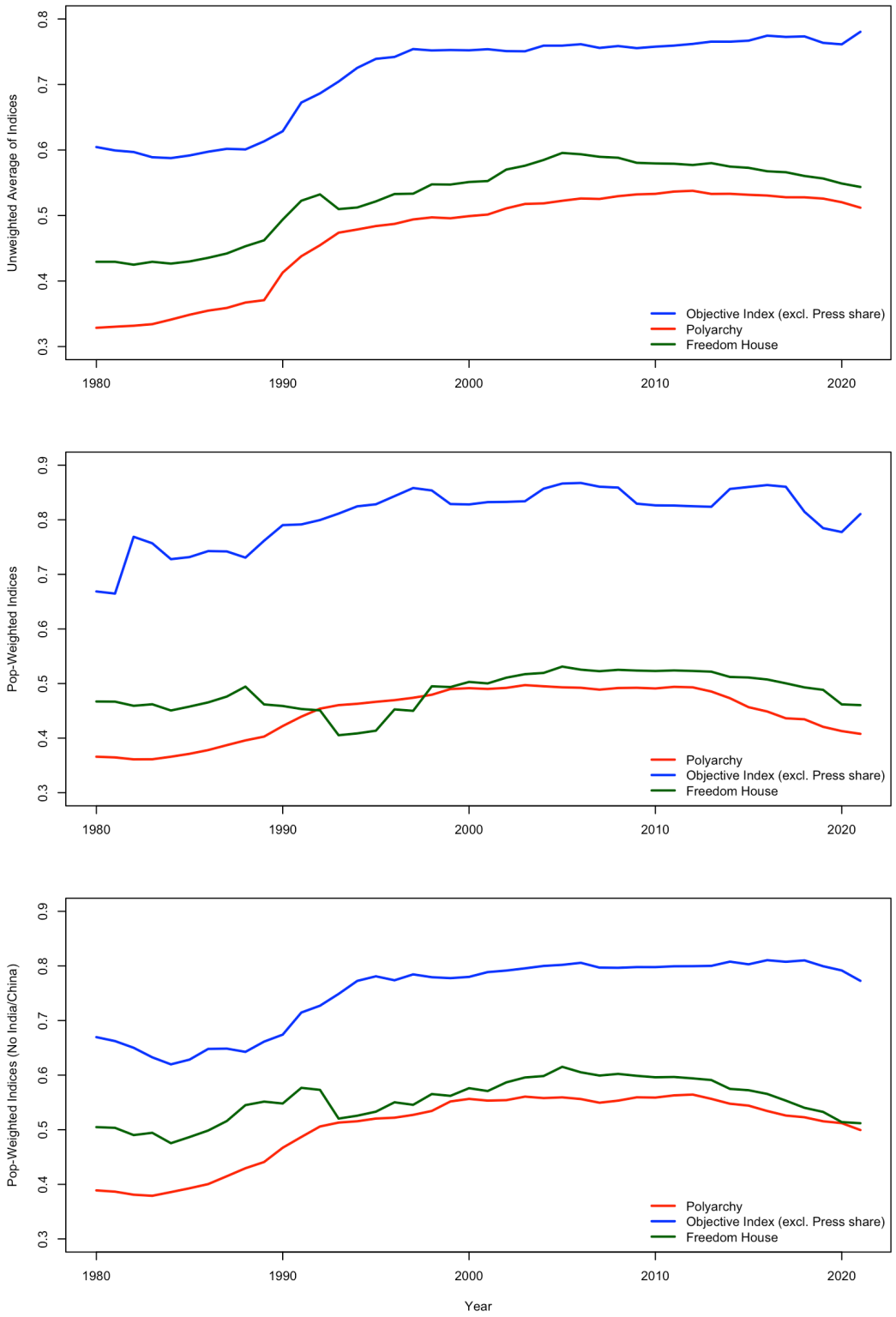


8. Since I did get the same results, I will change one key analytic choice and explain how and why the results change. As we saw in Table 1, there is one variable in particular that has a very high rate of data missingness: Presidential vote shares (Press share) from The Database of Political Institutions (DPI). This variable shows yearly winner vote share in presidential elections. Yet, it does not have data for all country-years; in fact, it does not have data for most of the country-years that are included in the data (about 65% of them). Nevertheless, the authors include this variable in their aggregate index. To test the consequences of this decision, I will present now what happens to results when this variable is excluded from the index's construction.

   Figure 2 show the left, middle, and right panels of Figure 8 in the article when the aggregate index is computed without the variable "Press share." As we can see, the decision whether to include or exclude this variable changes the results in a substantial manner. In the upper panel, the line of the aggregate index turns from a pattern of stagnation at the end, to a positive direction (meaning, more democracy). The same change occurs in the middle panel, where a decline turned into an increase. But, surprisingly, excluding this particular variable changed the bottom panel (correspondent to the right panel of Figure 8 in the article) from a rather modest decline to a steeper decline. In other words, excluding this variable and weighting by population without China and India shows that the authors' objective variable depicts a more similar pattern to that of V-Dem and Freedom House. Namely, a noticeable democratic backsliding which starts a bit later in time but points at the same direction with a similar slope. Contrarily, excluding that variable without weighting by population reveals a positive trend in the objective index, which, theoretically should have provided more evidence for the authors' claim on a more positive than expected trend in democratic backsliding overall.

   Either way, we can see that excluding or including one variable in the index changes the way we interpret its results in substantial terms. I should emphasize that the important change is only the one happening at the end of the trend line, meaning, recently. Yet, because this is also the focus of the authors in their article as well (discussing recent democratic backsliding and not historic trends), this does greatly impact their overall claim.

Figure 2: Replication of Figure 8 in the Article, without "Press share"



9. As my previous answer reveals, the data used by the authors to create the aggregate index seem to be volatile and noisy. If including and excluding one variable, out of 12, could change the direction of the trend, as well as making its slope noticeably steeper (depending on population weights) something might be worrisome in the way it is constructed. To be clear, the idea of creating an index, to begin with, is predicated on including important variables that could change the results over time. But the problem is, first, that the authors use this index to challenge two long-standing indices with wide reputation from scholars (V-Dem and Freedom House) while relying on data sources that were picked without resolute justifications. Second, it also seems that parts of the volatility in their data derives from large patterns of missingness, especially in several of the variables. Combined together, these call for a closer examination of the variables and data used.

   We should ask ourselves, for example: What would have happened if we decided to build two aggregate indices, rather than one, from the same 12 variables? In a similar fashion to our discussion on age in exploration 10, we can argue that some of the variables in their index actually capture different aspects of democracy than others in the same group of 12 variable. An immediate recategorization that comes to mind is that between variables which are purely institutional and those which relate to election outcomes. If we were to distinguish between these two categories, as I will attempt to do in this replication, we might be able to isolate variables that are very noisy (outcomes) from those which are more stable over time (institutions).

   In addition to revealing more interesting trends in the data, this could also add robustness to the authors' claim. If we saw similar trends on both indices pointing at the same direction, it would more likely mean that their claim is corroborated by a factual reality, rather than combination of variables' techniques. Distinguishing between these two categories would also help to obtain a clearer view into patterns of missingness in the data, which could lead to a better understanding as for which variables drive the noisiness in the data and at what particular point in time.

   Therefore, I will create two versions for the objective index representing different democratic concepts. One is "institutions," which describes to what extent democratic rules exist and includes suffrage, term limits, and electoral rules. The other is "outcomes," which includes dominant election victories, long party tenure, and lack of turnover. Theoretically,

a dominant ruling party can produce "bad" outcomes without institutional decay. The authors acknowledge that themselves by stating that "electoral performance can be influenced by factors other than how free and fair the elections are," for example in cases where incumbents perform differently "because they did a good job in office" or "a worse job in office" (p. 159). Thus, mixing the two embeds a conceptual assumption that dominant electoral outcomes are democratic erosion, even if democratic rules remain intact.

In addition to operationalization, another potential problem in the current analysis the authors are making is that of measuring the central tendency. As Wilcox (2012) observes, "[e]ven the population mean, if it could be determined exactly, can give a distorted view of what the typical participant is like" (p. 1). In our case, rather than "participants," we have country-year observations, with which the authors seek to learn about the typical case of a country's democratic trajectory. To do this, they rely heavily on (arithmetic) means. Yet, we know that this tool is often a rather dubious measure of central tendency. The mean might not always be indicative of the population (let alone when drawn from volatile data) primarily because of potential outliers and heavy-tailed distributions. Therefore, I will also run robustness checks for their results by trying different types of means we learnt about in class, to show whether different estimators of location yield different conclusions regarding the typical case in their data.

First, I will use the mean after trimming 10% of values from the lower and higher ends. Second is the Winsorized Mean, which I will calculate by replacing the lowest 10% values with the value at the 10^{th} percentile, and the highest 10% with the value at the 90^{th} percentile. Third is a one-step Huber M-estimator, which calculates the Median Absolute Deviation (MAD), identifying values which are away from the median to a certain cutoff and producing a mean that is also less influenced by extreme values. Lastly, I will use the median, which is another robust measure of central tendency.

10. Recreating the Index with Two Categories: As discussed in my previous answer, I distinguished between two categories of variables: institutions and outcomes. The categorization process is described in Table 2. Then, to observe patterns of missingness more clearly, I created Figure 3 that shows percent missing, by variable, with the upper

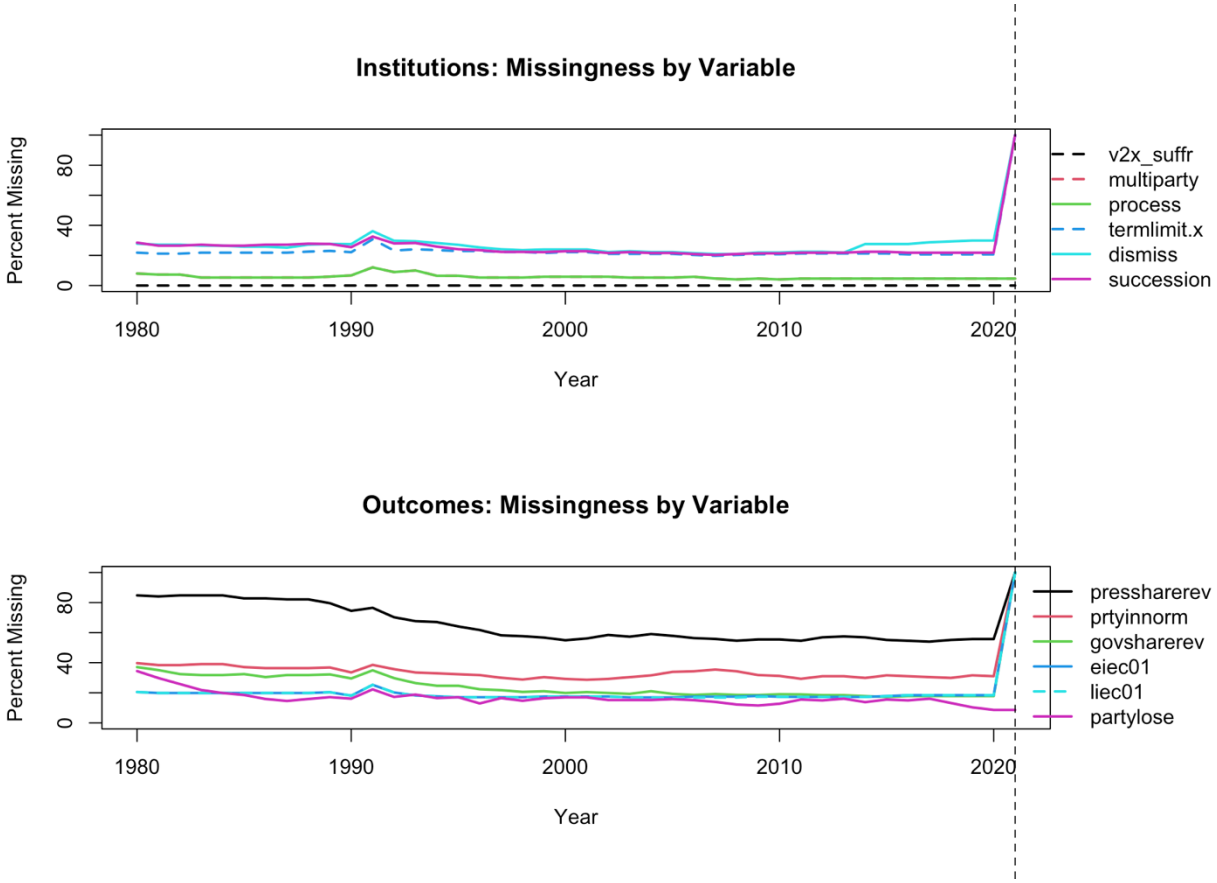
panel focusing on the institutions category and the bottom panel on the outcomes category. It turns out that in both categories, but to a greater extent in the outcomes category, we see a large share of missingness, specifically in 2021. Nevertheless, Figure 8 in the article does show that year for the aggregate index, and the authors also confirm that they use and show patterns until 2021 (p. 151). Because this large portion of missing data might affect result, I will replicate Figure 8 in the article again, this time by using strict versions of both my newly created categories. This means that the institutions and outcomes indices are only calculated for a country-year when all underlying component variables are observed, so any country-year missing even one component is excluded.[3] Figure 4 shows the yearly averages for these strict versions, alongside those of V-Dem and Freedom House.

Table 2. Categorization of Variables

Variable	Category
v2x_suffr	Institutions
multiparty	Institutions
process	Institutions
termlimit.x	Institutions
dismiss	Institutions
succession	Institutions
pressharerev	Outcomes
prtyinnorm	Outcomes
govsharerev	Outcomes
eiec01	Outcomes
liec01	Outcomes
partylose	Outcomes

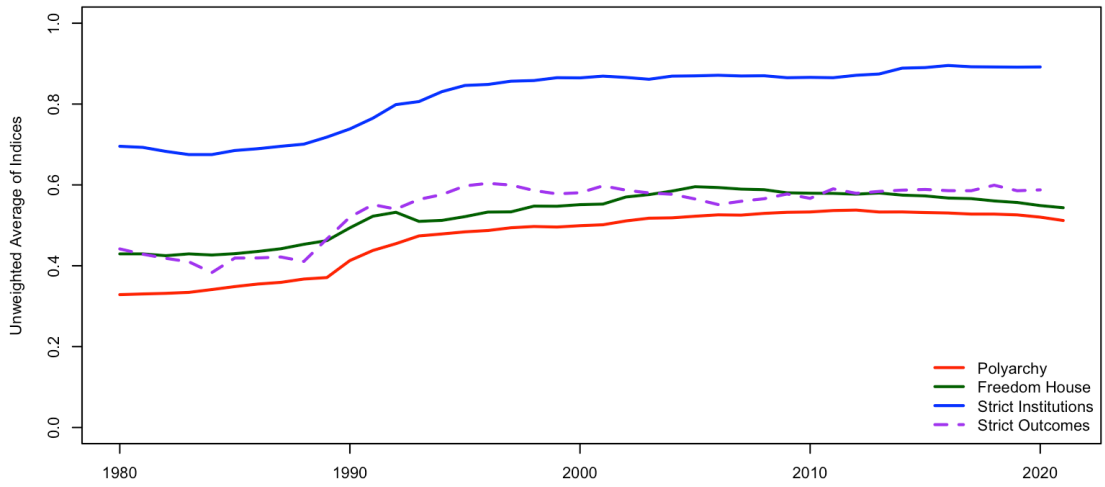
Figure 3. Missingness by Variable Over Time

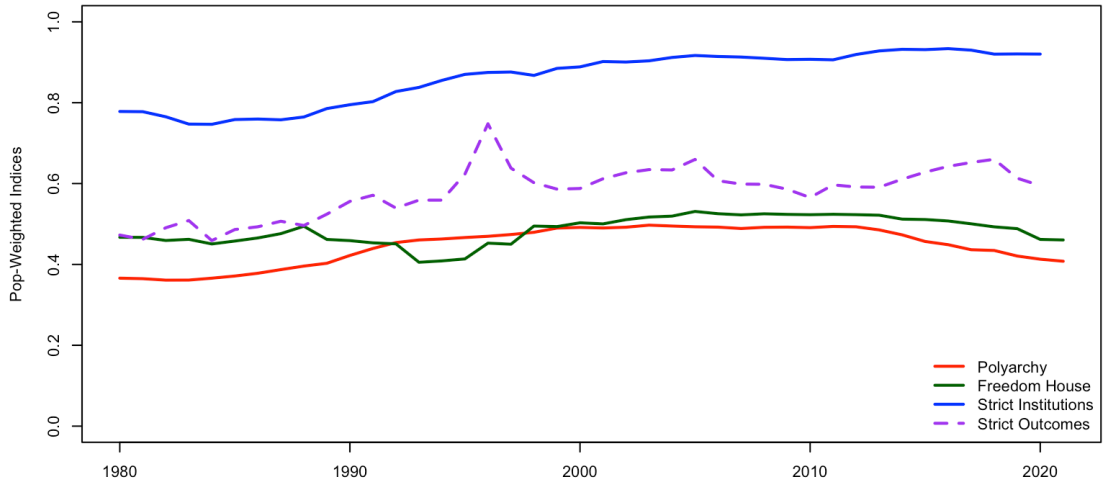
[3] This approach is not necessarily better or more accurate than the authors' approach. Rather, it serves the purpose of assessing missing data's consequentiality.

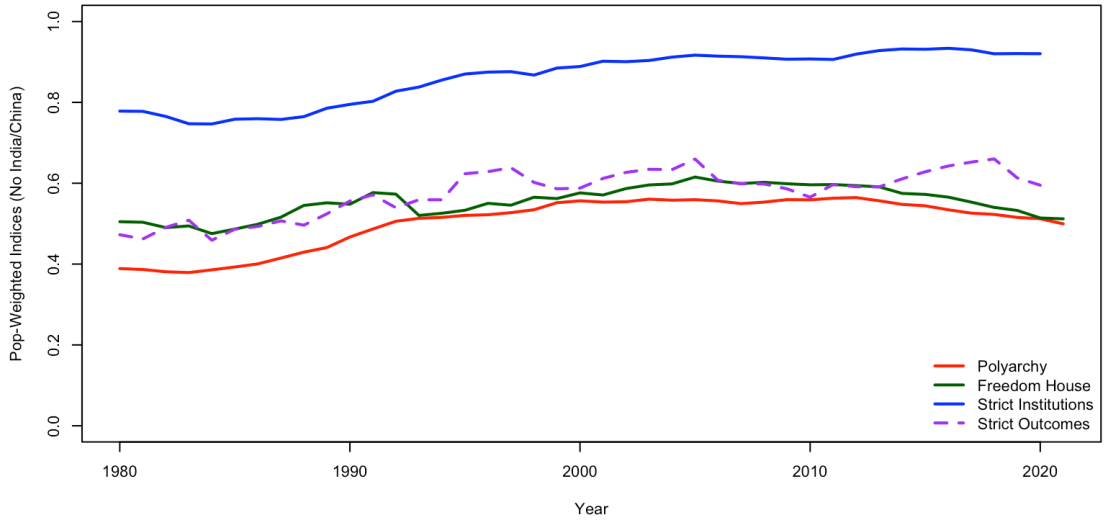


What all of this shows us, so far, is that differentiating between the institution and outcomes categories, as well as using their strict versions, both serve the authors' claim. Looking only at institutions-related variables in its strict version, Figure 4 reveals a very stable, and even positive, pattern over time. Contrarily, the outcomes category, even in its strict versions, is noisier. However, even for the outcomes category, using strict version reveals a much more stable pattern, when not weighted by population. When it is, the trend returns to being noisier.

Figure 4. Replication of Figure 8 in the Article with Strict Versions of Institutions and Outcomes


Unweighted Average of Indices
0.0
0.2
0.4
0.6
0.8
1.0
1980
1990
2000
2010
2020
Polyarchy
Freedom House
Strict Institutions
Strict Outcomes


Pop-Weighted Indices
0.0
0.2
0.4
0.6
0.8
1.0
1980
1990
2000
2010
2020
Polyarchy
Freedom House
Strict Institutions
Strict Outcomes


Pop-Weighted Indices (No India/China)
0.0
0.2
0.4
0.6
0.8
1.0
1980
1990
2000
2010
2020
Year
Polyarchy
Freedom House
Strict Institutions
Strict Outcomes

Using Different Measures of Central Tendency: After showing that getting rid of large patterns of missingness, as well as distinguishing between two categories of variables, could result in a more stable pattern and improve the data's description, we will now move to trying different measures of central tendency to see how much the trends discerned sustain them. To make comparisons fair, each measure of location will be applied to all four indices, including V-Dem and Freedom House.

Figure 5 shows the left panel of Figure 8 in the article with two categories for the aggregate index and four different measures of central tendency. The only noticeable difference between the original panel and these four possibilities is that, for V-Dem and Freedom House, the trend seems to be less negative. Except for using the median for polyarchy, in all panels of Figure 5 we see almost entirely flat lines. It means that when not weighted by population, all four measures of central tendency show that if democratic backsliding occurs, at occurs more on the fringes than at the center. Meaning, among countries with very high democracy score or very low democracy score (except for using the median for polyarchy; and still only at the very end).

Likewise, Figures 6 and 7 show the middle and left panels of Figure 8 in the article with two categories for the aggregate index and four different measures of central tendency. When weighted by population (with and without China and India), the patterns revealed are rather similar. While the institutions category I created for the authors' aggregate index is very stable over time and across measures of central tendency, the outcomes category remains noisy, but still stable in its overall direction. In other words, both categories show an overall trend of democratic stagnation, and not backsliding. Contrarily, V-Dem and Freedom House show the same pattern across the board: that of democratic backsliding, even if in a smaller slope.

Figure 5. Replication of Figure 8's Left Panel with Strict Versions of Institutions and Outcomes

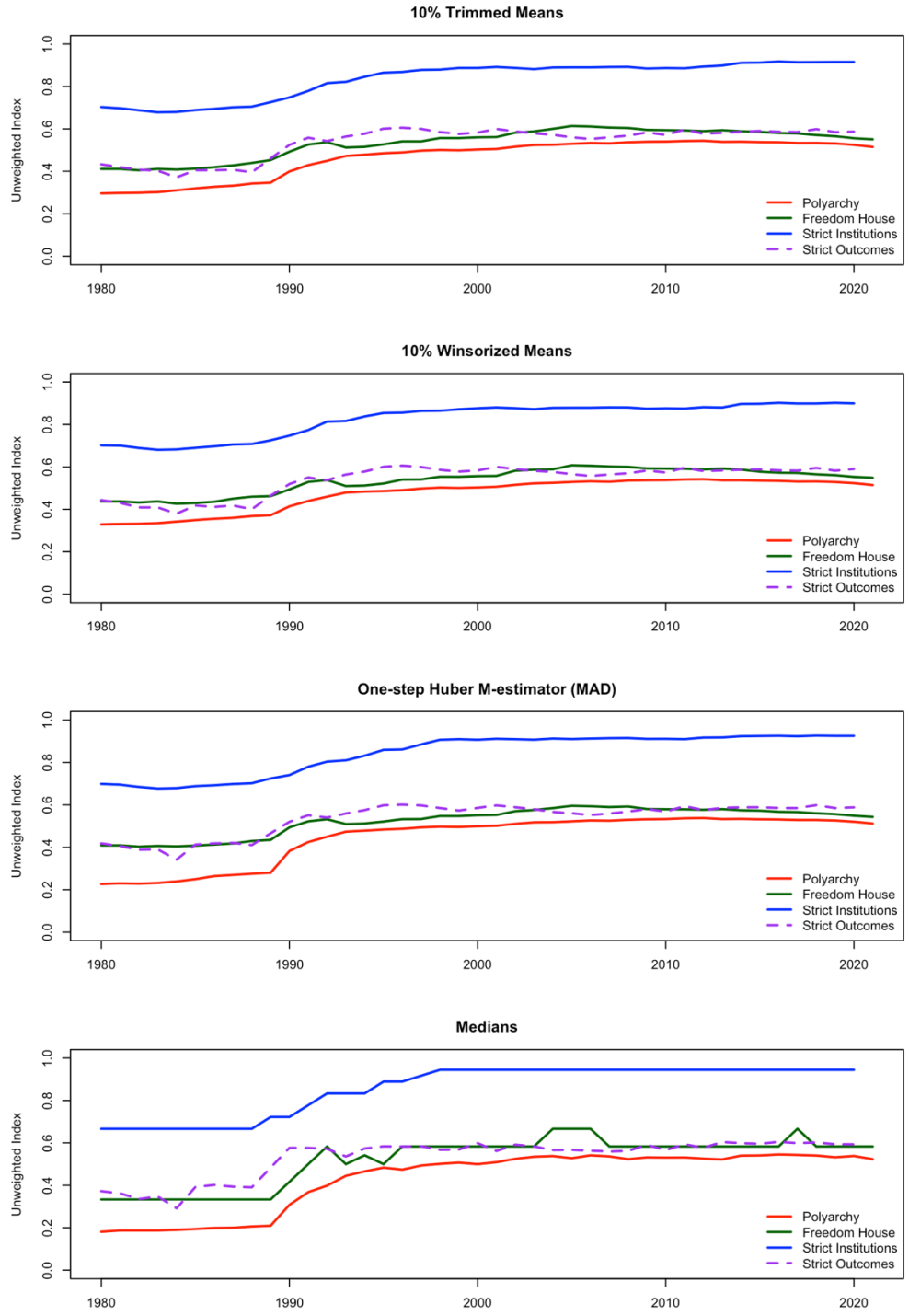


Figure 6. Replication of Figure 8's Middle Panel with Strict Versions of Institutions and Outcomes

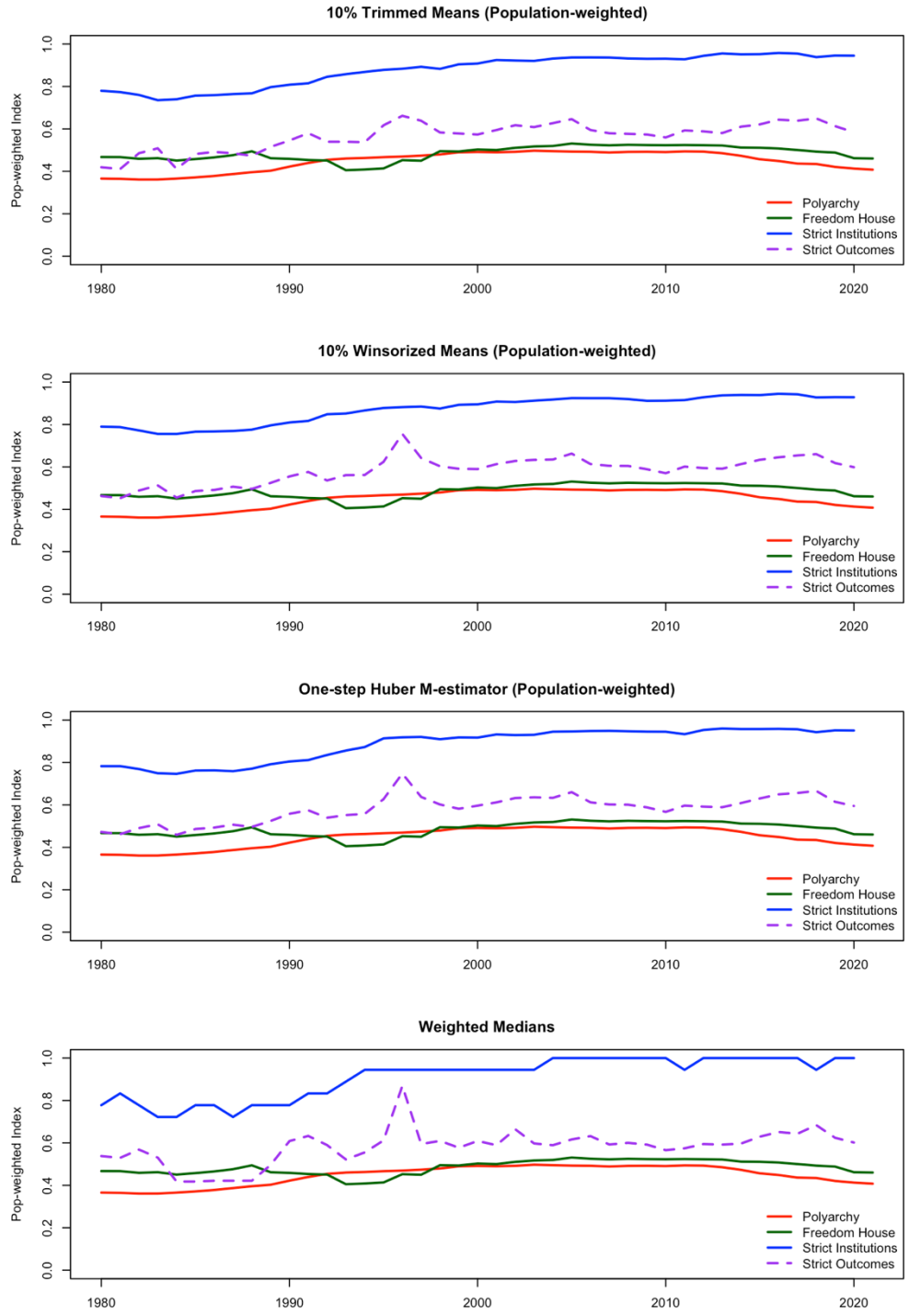
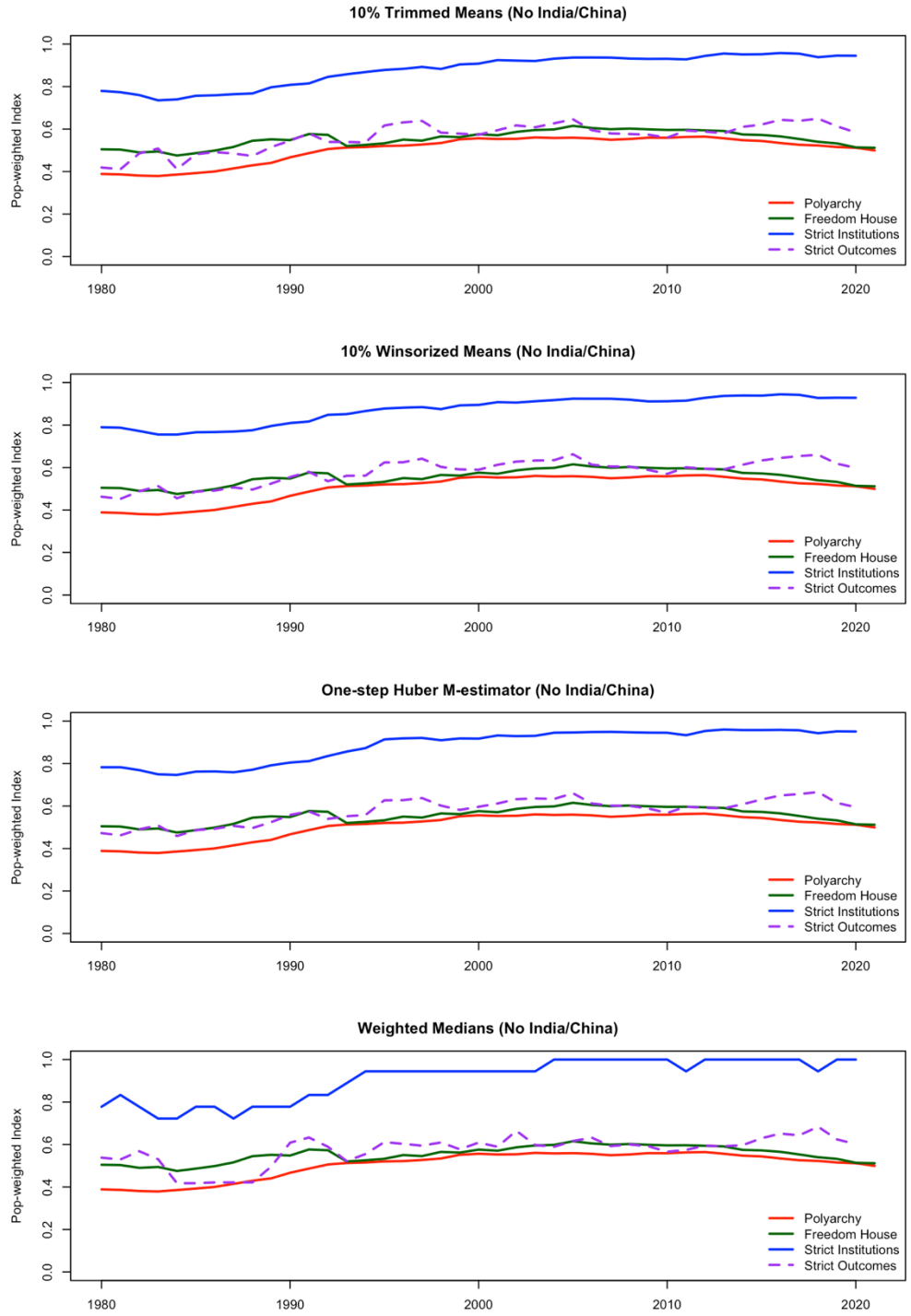


Figure 7. Replication of Figure 8's Right Panel with Strict Versions of Institutions and Outcomes



11. In writing this replication paper, I learned about the large number of untrivial decisions researchers make when they conduct research. To be sure, the term "untrivial" does not imply wrong or baseless. Rather, it implies that, when researchers seek to study a certain topic, they go through many intersections where they have to make consequential decisions. These decisions in research design, even if *prima facie* seem small, could change the study's overall conclusions. In the article I am focusing on here, the idea of decisions' consequentiality is even more apparent, because the authors constructed their own index, compiled their own merged dataset, and showcased many figures and descriptive statistics. Creating your own merged dataset and index exposes you to potential criticisms about your methodological decisions.

    My replication thus demonstrates the importance of replication ability. The fact that the authors attached a complete replication code to their article made me more confident about their findings overall. Yet, as my replication also shows, there still were some problems with the way the authors approached their data. The large portion of missing data affected, to some extent, their results. Since their focus was specifically on recent years and contemporary trajectory (including subtle changes), rather than long-standing trends over decades, the large quantity of data missingness (especially in 2021, which is included in their published graphs) did have an impact on their described results, even if not dramatic. Therefore, if I were to recreate this study today, I would limit the data described to 2020. I would also include different measures of central tendency to bolster the results even more, as my replication shows it does.

    In sum, my exploration seems to bolster the authors' claim that, based on their 12 objective measures of democracy, the trend is closer to that of democratic stagnation than to backsliding. This observation is sustained when their 12 measures are recategorized into two categories described separately, and when different measures of central tendency are applied. Moreover, when different measures of central tendency are applied, V-Dem and Freedom Houre (their "rivals") seem to be more sensitive. Then, both show democratic stagnation, rather than backsliding, or a more moderate backsliding than previously thought. This shows, of course, the consequences of deciding on a measure of central tendency to describe data.

Importantly, these conclusions are only relevant when we deploy the strict versions of both categories, which minimize the effect of large patterns of data missingness in the very recent time period used. In addition, they are only relevant, of course, if we accept the authors' operationalization of democracy as capturing these 12 institutions and outcomes variables to begin with.

This is a link to my GitHub repository for the paper:
https://github.com/idanf2/Replication-Paper-for-Q1

Works Cited

Little, Andrew T., and Anne Meng. 2024. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 149–61. https://doi.org/10.1017/S104909652300063X.

Wilcox, Rand. 2012. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.