Dichotomizing democracy: Surprisingly similar cutpoints in historical measures of democracy

Samuel Baltz[*], Fabricio Vasselai,[†] and Allen Hicken[‡]

October 14, 2020

Abstract

We show that by dichotomizing the two most prominent numerical classifications of the democratic level of political regimes, *The Polity Project*'s polity score and the *Varieties of Democracy* polyarchy index, it is possible to create a binary dataset that closely agrees with many previous dichotomous classifications of democracy. To match either the polity or polyarchy variables to a binary dataset, we pick the cutpoint that dichotomizes the many-valued measure so that it optimally matches the binary measure; when we do that for a series of diverse binary measures of democracy, we find highly similar optimal cutpoints. We show that behind decades of sophisticated disagreements over how to classify regime types there is actually a highly consistent picture of which countries have been considered democratic over time. We propose a consensus classification based on dichotomizing V-Dem's polyarchy score, and we show that it can replicate major results from the binary classifications.

1 Democracy: A matter of degree or of type?

Since the beginnings of political science as a field, identifying which political entities were democratic has been a major concern (Popper, 2020, Schumpeter, 1942). Even Dahl's 1973 seminal theoretical book on polyarchy was largely based on the author's classification of countries' democratic status circa 1969. As a consequence, over the decades, a great deal of the debate on the processes of democratization – and recently, of de-democratization – has centered around the details of how to classify political regimes.

Among the many methodological disagreements (for a review, see Boix et al. (2012)) perhaps the most critical has been whether democracies are a matter of type or of degree. Early work approached regime type as a largely all or nothing proposition, classifying some regimes as democracies and others as non-democracies at a given point in time (Dahl, 1973, Lipset, 1960). As efforts to chart countries' experiences with democracy over time became more common, we see the development of numerical proxies for democratization, like the

[*]PhD candidate, Department of Political Science, University of Michigan (Email: sbaltz@umich.edu)
[†]PhD candidate, Department of Political Science, University of Michigan (Email: vasselai@umich.edu)
[‡]Professor, Department of Political Science, University of Michigan (Email: ahicken@umich.edu)

percentage of the population voting in national elections (Lerner, 1958). Cutright (1963) was probably the first to propose a many-valued scale of democracy, which he calculated by combining his subjective scores for a series of institutional characteristics within a given country. For nearly two decades Cutright's measure served as the standard empirical variable employed in cross-country comparisons and econometric analyzes. Subsequently, Bollen (1980) pointed how much Cutright's subjective measure was endogenous to the very political and economic processes it was often used to investigate. Bollen proposed, instead, a score based on country characteristics more exogenous to institutional idiosyncrasies, like the presence of free press and free opposition, the fairness of elections, and whether the legislative and executive branches (when existent) were elected. While similar in spirit to measures of democracies used today, Bollen's measure covered only a few countries for a limited number of years. In part the work of Bollen and others was a reaction to the perceived limits of dichotomous classification schemes. In later work with Jackman (1989) Bollen summarized the common criticism of dichotomous measures: "Dichotomizing democracy lumps together countries with very different degrees of democracy and blurs distinctions between borderline cases" (Bollen and Jackman, 1989: p. 612).

Despite this common criticism, dichotomous classifications of democracy have not only re-emerged, but become more and more common in empirical analyses involving democracy. Alvarez et al. (1996), who offered their own dichotomous classification scheme, responded to Bollen and Jackman (1989), claiming that "it is one thing to argue that some democracies are more democratic than others and another to argue that democracy is a continuous feature over all regimes" (p. 21). While one might grant that there are different degrees of democracy and autocracy, there still exists a discontinuity. Or in their words, a country year cannot be "half-democratic: there is a natural zero point" (*ibidem*). Alvarez et al. (1996) drew on minimalistic notions of democracy, as established by Schumpeter (1942) and Dahl (1973), defining democracy as political regimes where governing offices are filled via competitive elections.[1] Specifically, for each country-year they identify whether elections were used in choosing current executive and legislative branches, whether those elections had more than one party, and whether there had ever before been an alternation in power (the turnover rule). Relying only on basic historical information about countries, they were able to classify many more countries and years than any previous study, covering a total of 141 countries through 41 years. Their classification became one of the most widely used, and was subsequently updated multiple times (Bormann and Golder, 2013, Cheibub et al., 2010, Przeworski et al., 2000) and their basic scheme, with some alteration, became the basis of multiple future classifications (Boix et al., 2012, Geddes et al., 2014, Svolik, 2008).

Alongside these new dichotomous measures, two ordinal indices were also frequently used by researchers: Puddington et al. (2018) (Freedom House) and Marshall et al. (2017) (the Polity Project). These two large-scale projects attempt to combine some of the most appealing characteristics of previous efforts to classify regime type. First, at their core they are both ordinal measures of country-year democracy levels, following in the tradition inaugurated by Bollen (1980). Second, like Alvarez et al. (1996) these measures sought comprehensive country coverage and extensive over-time coverage, with continual updates

[1] They also claimed that their dichotomous classification was based on observable indicators, and thus more objective than previous measures.

each year. Third, despite their ordinal nature, they are nearly always described as having a specific cutpoint in their ordinal scales that would define when a country-year could be considered democratic – akin to the natural zero point mentioned by Alvarez et al. (1996).

More recently, a new non-dichotomous alternative has emerged with an emphasis on comprehensiveness and transparency (Coppedge et al., 2019). The Varieties of Democracy project (V-Dem) relies on a very large team of experts selected from around the world to classify, among other things, the degree of political freedoms, electoral fairness, and electoral competitiveness of countries in each year since the French Revolution. Using a Bayesian composition approach V-Dem produces country-year measures corresponding to different conceptions of democracy, each with estimated confidence intervals. The primary measure of levels of democracy in V-Dem, called polyarchy, offers for the first time a non-ordinal many-valued measure (it ranges from 0 to 1, with 3 decimal places), with error estimation, clear classification rules, a large international team of expert classifiers, and covering a long historical period.

Interestingly, while each of these non-dichotomous initiatives have each been highly successful, often researchers choose to convert measures into dichotomous variables at some cutpoint along their scale (see Bogaards (2010)). This has long been common practice with Freedom House and Polity, and the same seems to be starting to happen with V-Dem (Kasuya and Mori, 2019). Naturally such a practice of dichotomizing many-valued measures of democracy, that is, of transforming political regime degrees into types, has also been criticized by many. Boix et al. (2012: p.1528) argue that such a dichotomization decision "ignores the fact that it is generally impossible to interpret a country's movement across a particular threshold in a substantive way." Gleditsch and Ward (2006) note that, even more problematic, each given value of those many-valued measures of democracy could result from a large number of combinations of their underlying components. Bogaards (2010) shows that altering the choice of where to cut the commonly used Freedom House and Polity ordinal measures of democracy level to create a binary classification of democracy actually changes empirical results of previously published models. And many are uncomfortable with the apparent arbitrariness of the choice of cutpoint. As Boix et al. (2012: p. 1528) lament, "none of the authors offer a concrete reason for the thresholds other than claiming they are intuitive or citing another study that uses the same threshold." As an alternative to dichotomizing the many-valued variables, some highly sophisticated methods have been developed to transform a large number of democratic classifications into a consensus classification (Pemstein et al., 2010).

Nevertheless, to discipline a choice of cutpoints, there is a simple option: we could simplify the many-valued measures of democracy using whichever cutpoint produces a binary variable that agrees with previous binary measures as much as possible. But this requires a compromise, because we might expect a many-valued measure of democracy to have a different optimal cutpoint for each binary dataset that we wish to match to (Boese, 2019, Boix et al., 2012, Kasuya and Mori, 2019).

However, using a series of datasets that have been central to the comparative study of democracy, we show that such a compromise might not be necessary. For nine previous classifications of democracy, we find the optimal cutpoint for both the Varieties of Democracy (V-Dem) polyarchy variable (Coppedge et al., 2019) and the Polity Project's polity variable (Marshall et al., 2017). For the polyarchy variable, which ranges from 0 to 1 with three digits of precision, we show that of the 1001 possible optimal cutpoints, *multiple different binary classifications of democracy share exactly the same best value*, and all of the values are close to each other. We observe a similar pattern for polity, though it only has 21 possible values. Crucially, *we argue that this represents a form of agreement between the binary classifications of democracy, and is not just a property of the variable being dichotomized.*

We identify the optimal cutpoints for matching polyarchy and polity onto the classifications by Anckar and Fredriksson (2019), Bernhard et al. (2001), Boix et al. (2012),[2], Cheibub et al. (2010), Gasiorowski (1996), Geddes et al. (2014), Puddington et al. (2018), and Skaaning et al. (2015). and Svolik (2008). These classifications of democracy were produced by several groups (with some overlap in authors) over decades of research. They aim to capture substantively distinct notions of democracy, and they apply those notions to different cases. Among their shared cases, they can have pairwise agreement as low as 63%. Indeed, the only consensus in the study of how to classify democracies is the uniform agreement that there is no consensus on how to classify democracies. Boix et al. (2012: p. 1525) write that the differences among "dozens of distinct measures of democracy and theorized in dozens of other articles" are "not merely academic", given that "empirical results can depend on the specific measure of democracy used". For Cheibub et al. (2010: p. 67) "differences across regime measures must be taken seriously" as "existing measures of democracy are not interchangeable".

And yet, despite the active disagreements that have produced such a variety of regime classification datasets, the empirical picture is extremely consistent. In Figure 1, we compare the level of agreement between the major binary codings of democracy, by matching each pair of datasets and checking how often they agree about whether or not a given country-year is a democracy. This represents the universe of reasonably independent measures of democracy that are either dichotomous or have straightforward dichotomization rules.

[Figure 1 about here]

Figure 1 shows the pairwise agreement between each dichotomous measure of democracy. In the top subfigure, each cell shows the pairwise agreement between the corresponding dichotomous measures of democracy. In the bottom subfigure, each cell shows the proportion of the dataset named in that column that also appears in the dataset named in that row. So, for example, the top subfigure reports that Boix et al. (2012) has a 94% pairwise agreement with Geddes et al. (2014), and the bottom subfigure reports that 52% of the cases that appear in Boix et al. (2012) also appear in Geddes et al. (2014), while 96% of the cases that appear in Geddes et al. (2014) also appear in Boix et al. (2012).

The pairwise agreement between these datasets puts numbers to the common refrain that there is no consensus in this area of study. To the contrary, datasets in this literature typically disagree only about one in every ten or one in every twenty cases.

However, there is a crucial limitation to this comparison. The number of cases in the intersection of any two datasets shows that pairwise comparisons can rarely tell the whole

[2] All of the results presented here refer to the original 2012 version of this dataset, because our analysis will include empirical replications of portions of that paper. Note however that the dataset was updated to include country-years through 2015 (Boix et al., 2018). When our analyses are run on the 2018 version of the dataset instead of the 2012 version, none of our substantive conclusions change.

story; commonly, for large datasets like Boix et al. (2012), pairwise comparisons can only be made on the basis of a quarter or a third of the dataset, leaving out thousands of country-years. It is also easy to imagine that the country-years in the intersection of any two datasets will systematically tend to be the country-years that the field is most certain about, since these will be the country-years that are more commonly studied and classified than the country-years which are present in one dataset but not in another. If this is true, the intersection of two datasets will tend to have a misleadingly high pairwise agreement, whereas the authors' coding rules might be more likely to disagree about the cases that appear in one dataset but not the other.

This motivates the idea of using a many-valued classification of democracy that is part of a very large dataset, like V-Dem or Polity, and identifying the optimal cutpoint for matching it to a binary dataset. When we perform this on the datasets in Figure 1, we find that the optimal cutpoint for matching V-Dem's polyarchy to those datasets only ranges from 0.382 to 0.423. More remarkably, multiple distinct datasets yield precisely the same optimal cutpoint values. Polity, which ranges from -10 to 10 at intervals of 1, shows a similar level of agreement, around its values of 5 and 6, though a less striking one since there are many fewer possible values.

Such strong agreement would not be substantively meaningful if the classification suggested by these similar cutpoints produced substantively different results than the original datasets did. So, using the modal best cutpoint to produce a consensus binary variable, we replicate one substantive result that originally motivated each of the binary measures. This demonstrates that the literature on measuring democracy can be represented by a single binary measure that is highly similar to each of the existing datasets without losing many of the major substantive results of those datasets to date.

With these empirical replications, our argument is not that all empirical analyses that an author might wish to conduct using a binary coding of democracy will be completely redundant with the consensus dataset used in this section. Of course, for any datasets which are not exactly identical, there is no guarantee that an analysis of interest will yield the same results using both datasets. Rather, our point is that in order to avoid the risk of redundancy with the consensus dataset, future datasets should either use a starkly different coding rule from the other datasets, or should simply use this consensus dataset. Indeed, even adopting a starkly different binary coding rule might not be enough; we will demonstrate that extremely different coding rules can result in extremely similar optimal cutpoints.

2 Finding similar cutpoints in different datasets

There are many advantages and disadvantages to using a variable with a very large number of possible values to classify democracies, but one advantage is that a classification with many values can be turned into a classification with few values. Figure 2 illustrates the basic idea of a cutpoint, using a fictional example of a country that became exponentially more democratic during the 20th century.

[Figure 2 about here]

Figure 2 shows that the country began as a full autocracy in 1900, and then became monotonically more democratic until it attained perfect democracy in the year 2000. The question is: in what year did it become a democracy? Is it when the country became "half democratic", as in the left panel in the bottom row of Figure 2? Or should we allow it to vary from case to case by drawing the line at some descriptive statistic of the distribution like the median value, as in the right panel in the bottom row of Figure 2? We could draw the line at – or even in between – any of the 101 points in Figure 2.

To confront this problem, we first precisely state what we mean by a "cutpoint". Call our measure of democracy level $\mathbf{x}$, and say that it is bounded between a lower bound x_L and an upper bound x_H. Then we translate $\mathbf{x}$ into a binary classification $\mathbf{b}$ of democracies ($b = 1$) and autocracies ($b = 0$) using a cutpoint $\tau \in [x_L; x_H]$ according to the following rule, where b is the element of $\mathbf{b}$ corresponding to some $x \in \mathbf{x}$:[3]

$$b = \begin{cases} 1 & \text{if } x \geq \tau \\ 0 & \text{if } x < \tau \end{cases}$$

This classification idea makes two claims. First, it claims that if a value $x \in \mathbf{x}$ is a democracy, then x' is also a democracy $\forall x' > x$, and conversely that if $x \in \mathbf{x}$ is an autocracy, then x' is also an autocracy $\forall x' < x$. This follows directly from the definition of $\mathbf{x}$: it holds if $\mathbf{x}$ is defined so that higher values of $\mathbf{x}$ mean that a case is more democratic. Second, the classification also makes the much stronger assertion that τ is the least democratic democracy, so that any value $x \in \mathbf{x}$ such that $x \geq \tau$ is a democracy, and any value $x \in \mathbf{x}$ such that $x < \tau$ is an autocracy. The difficult question in this method is therefore to decide where exactly to set τ.

In a vacuum, picking a cutpoint requires making a contentious (and perhaps even incoherent) substantive decision: we have to decide what proportion of the total quantity of possible democracy is a sufficient amount of democracy to be the minimum possible democracy. Thankfully, the fact that we have many different measures of democracy offers a simple shortcut for picking a cutpoint: previous authors have picked the cutpoint which best matches the many-valued measurement to a much simpler measurement, like a binary measurement (Kasuya and Mori, 2019). Anybody who has classified countries into either democracies or autocracies has already made a substantive judgment about what is the minimum requirement for a country to be a democracy, so rather than make a substantive judgment ourselves, we can simply try to match previous judgments as well as possible.

Consider the situation in Figure 3. This plot imagines a binary classification for the same burgeoning democracy in Figure 2, produced by a separate author, according to another rule for what constitutes a democracy.

[3] According to this rule, when the elements of $\mathbf{x}$ have a precision of $n \in \mathbb{N}$ digits, τ is actually an equivalence class: it is the set of real numbers which have an identical first n digits, since any such real will produce an identical classification $\mathbf{b}$. For example, if $\mathbf{x}$ has 4 digits (3 decimal places) like the polyarchy variable of Coppedge et al. (2019), there are infinite possible choices for τ that will yield an identical $\mathbf{b}$, obtained by setting the first 4 digits to be equal and then varying the subsequent digits. For clarity in this paper we refer to cutpoints in the singular – such as "a cutpoint" and "the optimal cutpoint". We really mean "the optimal cutpoint of precision n". Note however that an optimal cutpoint of precision n is still not necessarily unique, since multiple cutpoints of precision n could still yield a $\mathbf{b}$ which matches the binary coding equally well.

[Figure 3 about here]

This author has made their own substantive judgments about what exactly is the minimum requirement for a democracy; we can now take advantage of that existing judgment by setting a cutpoint so that our many-valued measure of democracy from Figure 2 most successfully matches the binary measure of democracy in Figure 3. The bottom-left panel of Figure 3 shows that a naive cutpoint set at $\tau = 0.5$ classifies many cases incorrectly. In the bottom-right panel, we identify the "optimal cutpoint". This is the point at which any increase or decrease in the value of τ will cause us to classify fewer years correctly. We denote the optimal cutpoint τ^*.

Note one feature of optimal cutpoints: they are rarely perfect. Figure 3 illustrates a common situation in which a country becomes democratic, then backslides into autocracy, and then returns to democracy. We can never capture this feature of the binary variable by applying a cutpoint to Figure 2, because the level of democracy in the many-valued classification is monotonically increasing every year. This simply represents a disagreement between the two measures; in some cases the optimal cutpoint might classify every case correctly, but in other cases the optimal cutpoint τ^* might only classify, say, 90% of cases correctly.

Behind this simple picture there is a thorny problem: no two binary classifications in the literature make exactly the same judgment about what counts as a democracy. If they did, the datasets would be redundant. How can we resolve the problem when cutpoints disagree? In the next section we demonstrate a highly convenient fact: the literature on democratic classifications, which includes many different notions of what exactly constitutes a democracy, consistently agrees about the best cutpoint for dichotomising the main many-valued measures of democracy.

3 Optimal cutpoints for democratic classifications

3.1 Finding the optimal cutpoint

In this section we find the optimal cutpoint for matching V-Dem's polyarchy variable and the Polity Project's polity variable onto each binary classification of democracy. We will primarily focus on five datasets that we take to be representative of the major ways that authors have categorized democracies. Those five datasets, and their stated coding rules, are listed in Table 1.

[Table 1 about here]

We focus on these datasets because they represent the full range of methods for categorizing democracies. Cheibub et al. (2010) explicitly updates Alvarez et al. (1996) and its implementation in Przeworski et al. (2000). Boix et al. (2012) is an original update of the Boix and Rosato (2001) dataset, which employs a different definition than the literature which takes Przeworski et al. (2000) as its starting point. A third explicit rule for what consitutes a democratic country-year is used by Geddes et al. (2014), who codes democracy as the absence of non-democratic features rather than the presence of democratic institutions. These three datasets represent very large cross-national and historical coverage of the history of democracy using three different explicit rules for whether or not a country-year represents a democracy. In contrast to all of these, Puddington et al. (2018) uses a discussion and review process rather than categorizing by means of explicit rules. For the purposes of finding cutpoints, we also deliberately include a quite different case: we include the coding in Svolik (2008) for whether or not a democracy is consolidated, but when finding cutpoints we treat transitional democracies as being not fully democratic. We do so to be as conservative as possible, because we make absolutely no *a priori* judgments about where the line should be drawn between a democracy and an autocracy, and it would be reasonable to expect that such a different coding rule and selection of cases might produce a very different cutpoint value.[4] However, for the purposes of simply comparing the agreement among the binary datasets, we follow Svolik (2008) in considering transitional democracies to be democracies, because it is not accurate to say that Svolik disagrees with other authors in considering these to be democracies. Among all of these authors' varied coding procedures, some are also in strong contrast to the way that the many-valued measures are produced; polity is drawn from a discussion and review process, while polyarchy is the result of a classification model produced from an expert survey.

The level of pairwise agreement between these five datasets was included in Figure 1. Which cases drive their disagreements? Of the 199 distinct countries that appear in at least one of these five binary datasets, 101 of them (a majority) have no disagreement among any author who classifies a year as either democratic or autocratic. For the other 98 countries, there is at least one disagreement over the classification for at least one year. The frequency of countries which have a certain number of disagreements is shown in Figure 4.

[Figure 4 about here]

Figure 4 shows that nearly all of the countries that have a disagreement have only a small handful of country-years under dispute. Among the countries that are the subject of at least one disagreement, about half of them have fewer than 5 country-years that at least one author classified differently from at least one other author. Rather than being substantive disputes over whether or not a particular case represents a true democracy, many of these disagreements are simply differences in when authors count changes. For example, does a country-year represent the status of the country on January 1 of that year, or December 31, or something else? Disagreements over a large number of country-years only affect a small number of the countries in the dataset. Only 26 of the 199 countries have more than 10 years

[4] There is also a structural problem that motivates this decision. If we followed Svolik in treating every case in the Svolik (2008) dataset as being a democracy, then we would classify both transitional and consolidated cases as democracies. This would produce a dataset with more than 3,000 democracies and no autocracies. Svolik (2008: p. 156) reports that the full universe of democracies and autocracies that he coded is a temporal extension of the data in Boix and Rosato (2001), extended back to 1789 and forward to 2001; however, our analysis already includes an updating of the Boix and Rosato (2001) dataset, namely Boix et al. (2012). So rather than reconstructing Svolik's coding of democracies and autocracies to produce a dataset that we should expect to be highly redundant with the Boix et al. (2012) dataset, we instead pursue the more conservative idea of taking Svolik's binary coding of transitional or consolidated democracies as its own type of democratic classification and finding the optimal cutpoint for that distinction.

which were coded as a democracy in at least one dataset while being coded as an autocracy in at least one other dataset. Table 2 records these countries and the years in question.

[Table 2 about here]

Simply counting up the appearances of country-years as democracies or autocracies across the binary datasets is one way of checking the datasets' agreement. In Figures 5 and 6, we present an alternative method for checking similarity between many-valued classifications and dichotomous classifications: the optimal cutpoints for dichotomising polyarchy and polity to match with each of the dichotomous classifications.

[Figure 5 about here]

[Figure 6 about here]

Figure 5 plots the quality of the match between polyarchy and each binary dataset when τ takes on every possible value in the range of polyarchy, and Figure 6 shows the same data for polity. In these figures we include every dataset from Figure 1, and in Table 3 we record the cutpoints and level of agreement for just the five core datasets that we will focus on.

[Table 3 about here]

Figure 5 presents a remarkable regularity: the optimal polyarchy cutpoints for all of the dichotomous measures are in deep agreement with one another, differing by only 0.041, which is only 4% of the total range of polyarchy. In Figure 5, multiple extremely different datasets with radically different cases and coding methodologies share exactly the same optimal polyarchy cutpoint. We find exactly identical optimal cutpoints between Anckar and Fredriksson (2019) and Boix et al. (2012),[5] and also between Svolik (2008) and Freedom House.[6] These two pairs of identical cutpoints differ only by 0.002, and a fifth dataset by Skaaning et al. (2015) lies exactly in between them, off from each by only 0.001. There is a second grouping of three datasets from 0.382 to 0.393. In this group, Geddes et al. (2014) and Cheibub et al. (2010) almost exactly agree with one another, with optimal cutpoints of 0.382 and 0.384 respectively. This method of checking every dataset against the optimal polyarchy cutpoint in V-Dem shows that expert agreement on what is and what is not a democracy is quite remarkably strong.

Is this an exceptional feature of the polyarchy variable, or does it rather reveal something about the agreement between the binary datasets themselves? We will explore this question more deeply in Section 3.2, but as an initial investigation we can exploit the fact that there is more than one extremely large dataset with an original many-valued measure of democracy. Figure 6 shows the analogous values for polity, rather than for polyarchy. Polity also shows that the datasets closely agree about the value of τ^*, with many datasets sharing exactly the

[5] The updated dataset Boix et al. (2018) has precisely the same optimal cutpoint value as Boix et al. (2012).

[6] This is a particularly close agreement since we treat the transitional democracies in Svolik's data as autocracies.

same optimal polity cutpoint, but this is less remarkable because polity has only 21 values rather than 1001 values.[7]

It is intriguing enough that the values of τ^* are always nearly identical – and sometimes exactly identical – when there are 1001 possible values of τ^* in the case of polyarchy and 21 possible values in the case of polity. But the figures show an even stronger result. The fact that most of the panels in those figures have the same characteristic shape demonstrates that not only is the optimal cutpoint always near the same value, but if we pick a cutpoint other than the optimal cutpoint, it is similarly worse than the optimal cutpoint across all of the datasets. We see a similar pattern in Figure 6, though less striking because it has many fewer values.

There is one important caveat pertaining to the optimal cutpoint of three of the nine datasets in Figures 5 and 6. While six of these datasets are binary measures, Freedom House, Gasiorowski (1996), and Skaaning et al. (2015) have more than two possible values, but many fewer than our many-valued measures of polyarchy and polity. So, a decision has to be made about how to dichotomize them. In Figures 5 and 6 we dichotomize them by simply dropping those cases which are neither fully democratic nor fully non-democratic. An argument could also be made that we should instead classify all ambiguous cases as non-democratic, since by virtue of being ambiguous they are not fully democratic. Figures 7 and 8 show the implications of this decision for the optimal polyarchy and polity cutpoints respectively.

[Figure 7 about here]

[Figure 8 about here]

In the cases of Gasiorowski (1996) and Skaaning et al. (2015) the decision turns out to be mostly unimportant for the optimal cutpoint, at most changing the second decimal place and roughly keeping the results in line with the optimal cutpoints seen in Figures 5 and 6. In the case of Freedom House, however, the decision of how to treat middle values is hugely important. Because Freedom House has approximately as many "Partially Free" countries as it has "Free" or "Not Free" countries, merging the "Partially Free" category with the "Not Free" category causes the optimal cutpoint to rocket up to 0.635, dramatically away from every other optimal cutpoint. For the remainder of the analysis we will drop the middle cases, noting that this decision only has a meaningful effect in the case of Freedom House.

Having checked the optimal polyarchy and polity cutpoints for each binary dataset, in Table 4 we address the related question of how to optimally match polyarchy to polity. Since polity ranges from -10 to 10, there are 19 possible (nontrivial) ways to dichotomize polity, each with a corresponding τ^*.

[Table 4 about here]

To handle the fact that there is no one substantively correct dichotomization of polity, one naive strategy is to simply dichotomize the variable in the middle of the range; Table 4

[7] It is also worth noting that exclusively regional classifications like Mainwaring et al. (2007) and Bowman et al. (2005), which can be clearly dichotomized into democratic versus non-democratic cases, yield slightly higher optimal V-Dem and Polity cutpoints.

shows that if the dichotomization of polity is around the middle of polity's range, then the optimal cutpoint for matching polyarchy to the resulting dataset ranges between 0.3 and 0.4.

Another strategy might be to pick the modal result for polity's optimal cutpoint for matching the binary datasets, and use that as its best binary classification (the one which is most similar to the literature). In Figure 6, polity's optimal cutpoint was found to be in the range from around 4 to 6. Conveniently, 4 is also the value that has been suggested by Marshall et al. (2017) as a potential democratic cutpoint. At this value, the optimal polyarchy cutpoint for matching to polity also closely resembles the optimal polyarchy cutpoint for matching to the other datasets in Table 3, with a value of either 0.396 or 0.42. Interestingly, the latter is within 0.03 of the five tightly grouped values in Figure 5. Even more intriguing, these are the highest agreement levels in Figure 5: the optimal polyarchy cutpoints for matching onto the binary dataset obtained by dichotomizing polity at values 4 and 5 produce a better agreement than the optimal polyarchy cutpoints at any other value of polity.

So, the result is another unexpectedly strong relationship. Pick the most common optimal polity cutpoint from Figure 6, around 4 or 5. The former turns out to also be the common cutpoint that is recommended by the *Polity Project*. Then dichotomize polity in every possible way, and check what the optimal polyarchy cutpoint is to match with each dichotomization of polity. The very best polyarchy cutpoint for matching to any possible dichotomization of polity happens to a) correspond to the optimal polity threshold for matching to the binary datasets, b) correspond to the recommended polity cutpoint, and c) also be the best polyarchy cutpoint for matching to the binary datasets. This reveals an exceptional level of agreement between the binary datasets, polyarchy, and polity.

Despite the consistent rhetorical agreement in the literature that there is no consensus on how to classify regime types, the extremely strong agreement that we have uncovered in the optimal cutpoints might have been expected after seeing in Figure 1 that very different datasets often have very high pairwise agreement. But it is worth bearing three things in mind. First, even exceptional pairwise agreement does not guarantee identical optimal cutpoints; notice that Cheibub et al. (2010) and Boix et al. (2012) only disagree to the tune of about one in every 20 cases, and yet their optimal cutpoints differ by nearly 4% of the range of polyarchy. Second, low pairwise agreement does not guarantee different optimal cutpoints. Freedom House and Svolik (2008) disagree about 1 in every 5 cases,[8] and yet they have precisely identical polyarchy cutpoints. Third, the pairwise comparisons are often based on small slivers of large datasets. In the pairwise comparison of Boix et al. (2012) and Freedom House, those datasets share only 7,643 cases. We can very easily imagine that the 10,555 cases in Boix et al. (2012) but not in Puddington et al. (2018) – or the 1,082 cases in Freedom House but not in Boix et al. (2012) – could have dragged the optimal cutpoint of one dataset away from the optimal cutpoint of the other. And yet, their optimal cutpoints differ only by 0.002.

Having found that the optimally dichotomised polyarchy variable only disagrees with the optimally dichotomised polity variable in about one out of every twenty cases, we will focus primarily on polyarchy for the rest of our analysis. We focus on just one of the two for

[8]Interpreting Svolik's transitional democracies as democracies, and blanking Freedom Houses's middle values.

the sake of communication, and we focus on polyarchy rather than polity because polyarchy has many more possible values than polity, and the recurrence of exactly the same cutpoint in the 1001 possible values of polyarchy is more striking than the recurrence of the same cutpoint in the 21 possible values of polity.

3.2 Validating the cutpoints

Perhaps these regularities are somehow a consequence of comparing cutpoints to each other, and we would get strikingly similar shapes no matter what the distributions of polyarchy or polity were (although the fact that both polyarchy and polity have tightly grouped cutpoints that are different from each other does already guard against this). To concretely assure us that this isn't just some unseen statistical regularity that is unrelated to the substantive meaning of polyarchy, Figure 9 shows how each dataset compares to a column of random numbers drawn from $\mathcal{U}(0, 1)$.

[Figure 9 about here]

When testing what cutpoint best dichotomizes a uniform random number to match the binary datasets, the shape of the panels is determined entirely by one detail: the frequency of zeroes and ones in the dataset. The optimal cutpoints are also trivial. Those datasets which have more democracies than autocracies are best dichotomized very close to 0. Those datasets which have more autocracies than democracies are best dichotomized very close to 1. These dichotomizations yield agreement levels of the same proportion as the distribution democracies and autocracies in the dataset. This pattern is quite an instructive, because it reveals what happens in the absence of any similarity in the data-generating process of the fake "democracy" measure and the dichotomous values. The reason that polyarchy peaks around 0.4 is that at this value it is systematically likely to classify the same countries as ones and the same countries as zeroes as the dichotomous datasets are, and it does so at a rate greater than the base frequency of un-systematically classifying them all as zeroes. Notice that this can also be seen in Figure 5: the optimal polyarchy cutpoint always comfortably beats the highest or lowest polyarchy cutpoints, often doing so by several dozen percentage points. So we can be assured that the strong regularity we find in Figure 5 and Figure 6 was not just some automatic distributional oddity; polyarchy and polity really are picking up a latent similarity between the dichotomous measures.

To validate that the consistent appearance of a threshold just above 0.42 is not just a meaningless artifact of the shape of polyarchy, the complementary check is to dichomotise the real polyarchy variable to optimally match a uniform random draw of zeroes and ones. The distribution of optimal cutpoints when polyarchy is matched against a column of zeroes and ones 1,000 times is shown in Figure 10. Indeed, the area around 0.4 is seen to be a very uncommon result. The bounds of polyarchy are the most likely values to be optimal cutpoints when it is dichotomized to match a binary variable that it is not systematically related to, while optimal cutpoints in the middle of the range of polyarchy are highly uncommon.

[Figure 10 about here]

Another important question is how often democracies and autocracies appear in each dataset for each value of polyarchy. We should expect an increasing trend, with very few cases at low polyarchy values being counted as democracies in the five dichotomous datasets, and very few cases at high polyarchy values being counted as autocracies. Figure 11 confirms this expectation.

[Figure 11 about here]

Figure 11 asks the following question. For the country-years that have a given polyarchy value in V-Dem, how often do the authors of the dichotomous datasets consider those country-years to be democracies? To obtain the proportion for each polyarchy value in Figure 11, we first check the cases that appear in V-Dem at each polyarchy value. Then, we count the number of times that those cases were classified as a democracy in the dichotomous datasets. Finally, we divide that by the total number of times that any observation at that polyarchy value was included in any binary dataset.

For example, suppose we set 0.5 as the cutpoint and that the only two country-years in V-Dem that had a polyarchy value of exactly 0.5 were the Republic of Michigan in 2018 and the Theocracy of Ohio in 2017. Say that the Republic of Michigan 2018 was recorded in four datasets as being a democracy, and was missing from the fifth dataset, while the Theocracy of Ohio 2017 was recorded in all five datasets as being an autocracy. Then at the polyarchy value 0.5, the proportion of democratic classifications would be $\frac{4}{9}$. This plot shows the proportions for the entire range of polyarchy (1001 values) for $N = 16,328$ country-years that were contained in V-Dem and any one of the 5 dichotomous datasets, double-counting any observations that occurred in multiple datasets. So the values that go into the proportions are not distinct country-years, but rather distinct classifications of any country-year.

The strong right-skew of the points in Figure 11 also gives us a simple explanation for why the optimal cutpoint is around 0.42. We can clearly see in Figure 11 that the range in which the other five authors most consistently disagree with one another is between around 0.3 and 0.5. This is offset from the center of the polyarchy distribution, to the left. The driving force, and the most interesting regularity in this plot, is that authors begin to fairly consistently assign countries to be 1 when they have polyarchy values of around 0.65, so for most of the upper half of the range of polyarchy, country-years are consistently assigned a 1 by upwards of $\frac{2}{3}$ of the datasets that they appear in. By polyarchy values of 0.8, we see nearly 100% of all cases assigned a 1 by every single dataset they appear in. In contrast, authors only consistently identify those country-years as autocratic which have polyarchy values less than about 0.15, and around polyarchy 0.2 we begin to see quite inconsistent assignment. A number of cases around the lowest quarter of the polyarchy scale sometimes have as much as 40% disagreement over whether or not they are democracies. These asymmetries suggest, perhaps counter-intuitively, that there is more consensus about what makes a country a democracy than about what makes a country an autocracy. Authors appear to be more willing to label a potential autocracy as a democracy than to label a potential democracy as an autocracy. Conversely, notice that cases that have a proportion of democratic votes below 0.5 are almost always classified by the cutpoints method as autocratic (there are very few cases below $y = 0.5$ and to the right of $x = 0.42$), but there are more cases which more than half of the binary variables believe are democratic but which the cutpoint method classifies as autocratic (dots above $y = 0.5$ but the left of $x = 0.42$).

3.3 Agreements and disagreements

We have consistently found strong agreement among the datasets, but two important pieces of the puzzle are missing from Figure 11. First, our method for calculating the proportion of agreement does not consider how many cases appear at each polyarchy level. It cannot say whether a conspicuous outlier is a stray classification by one author considering a case that nobody else included in their dataset, or if it is a mismatch between polyarchy and a consensus decision of all 5 binary datasets. Second, it does not identify which cases are contentious and which are the subject of strong consensus. In this section we identify which cases drive the disagreements in the literature, and we also identify where the consensus classification using $\tau = 0.42$ actually departs from previous binary classifications.

In Figure 12, we treat every appearance of a country-year in one of the five core binary datasets as a "vote" by a dataset in favour of that case being either democratic or autocratic.[9] A dot at the top of Figure 12 represents a unanimous judgment that a certain country-year which has that polyarchy value is a democracy; a dot at the bottom of Figure 12 represents a unanimous judgment that some country-year at that polyarchy is an autocracy; and a dot in between 0 and 1 on the y-axis of Figure 12 means that multiple binary datasets disagreed on how to classify some country-year with that polyarchy.

[Figure 12 about here]

Figure 12 shows that the proportion of datasets which classify each country-year as a democracy is indeed strongly increasing in polyarchy. Because the density of dots is difficult to visually assess, we plot a trend line, which is a simple linear regression where the dependent variable is the proportion of times that each country was classified as a democracy and the independent variable is polyarchy. The trend is sharply increasing. A vertical line marks the optimal cutpoint $\tau = 0.42$. The bulk of the disagreements (values closer to 0.5) can be seen to occur between polyarchy values of about 0.2 and 0.6, with hardly any serious disagreements at very high and very low values of polayarchy. Recall that we investigated which cases drive these disagreements in Figure 4 and Table 2. The few extreme outliers – the small number of dots which either have extremely high polyarchy and consistent agreement among the binary datasets that they are autocratic, or extremely low polyarchy and consistent agreement among the binary datasets that they are democratic – are typically disagreements over when a democratic spell began or ended. So there are some cases in which the authors all agree that the country switched between being a democracy and an autocracy, but they disagree about which year the switch happened in, or more precisely, they disagree how to code the year in which the switch occurred.

The idea that each binary classification can be considered a vote for or against a certain country being a democracy suggests that one idea for producing a consensus dataset would

[9]For this figure and and similar comparisons, because we are concerned with judgments not about the level of democracy however it is defined but much more specifically with the difference between democracies and autocracies, we treat any appearance of a democracy in Svolik (2008) as a judgment that it is a democracy, whether it is consolidated or not.

be to simply tally the number of times that each country-year is classified as a democracy. It might appear that this is a promising alternative to the idea of dichotomizing polyarchy by means of cutpoints. There are four heavy downsides to this idea compared to using cutpoints. First, this idea does not make any use of the many-valued variables polyarchy and polity, and any attempt to incorporate them will have to propose some sort of a cutpoint notion. Second, 39% of the cases that appear in at least one the five datasets actually only appear in one of those datasets. To turn the five datasets into a consensus dataset by means of voting, the cases that appear in only one dataset would have to either be dropped or be included on the basis of that one vote alone. So we could be no more confident in 39% of the cases in such a dataset than we are simply by checking the binary dataset they were originally coded in. Third, the union of the five binary datasets is only 65% the size of V-Dem; this idea would lose 8,667 country-years that we can capture by dichotomizing polyarchy. Fourth, voting alone cannot resolve ties. Across the 16,328 country-years that appear both in V-Dem and at least one of the five binary datasets, 261 of them have exactly as many datasets classifying them as democracies and autocracies. These 261 country-years come from 54 different countries. Those ambiguities are not resolvable on the basis of the binary classifications alone.

However, we can use this voting idea to get a sense of where the consensus dataset created by dichotomizing polyarchy at the $\tau = 0.42$ level disagrees with the pre-existing binary datasets. We simply check the cases which at least half[10] of the datasets classified as democracies but which hold polyarchy values less than 0.42, and the cases which more than half of the datasets classified as autocracies but which hold polyarchy values of at least 0.42. These are cases where our two consensus-finding methods, binary votes and optimal cutpoints, give contradictory results.

Across the 16,067 country-years that appear in V-Dem and at least one binary dataset (16,328 minus the 261 exactly tied votes), there are 992 disagreements, so about 6.2% of the dataset. 724 country-years which a majority of binary datasets classified as democracies are considered by the cutpoint method to be autocracies (false negatives in counting democracies), and 268 country-years which a majority of datasets classified as autocracies are considered by the cutpoint method to be democracies (false positives in counting democracies). So, compared to the binary datasets, the cutpoint method has a bias towards false negatives in counting countries as democracies, consistent with the situation seen in Figure 11. Table 5 shows the ten countries with the most false positives and the ten countries with the most false negatives.

[Table 5 about here]

False negatives – country-years classified as democratic by a majority of binary classifications, but classified as autocratic by the dichotomized polyarchy variable – tend to be countries that meet the minimalistic criteria used for binary classifications of democracy, while having polyarchy less than 0.42. The most frequent false negative is the United States from its founding through the late 1800s, which for 96 years is classified as a democracy

[10] Simply ignoring the 261 ties, as if they were not included in any datasets; just like country-years that were not included in any dataset, the ties are cases for which the collective body of datasets does not give us any reason to conclude that they are either democracies or autocracies.

by a majority of binary datasets while holding a polyarchy value under 0.42. There is a straightforward explanation for this disagreement: the early United States satisfies many of the simple requirements, like an elected executive and a minimum level of male suffrage, that were used to create most of the binary classifications (See Table 1), while also being rife with egregiously non-democratic institutions (Mickey, 2015) that might lower its polyarchy value.

False positives – country-years classified as autocratic by a majority of binary classifications, but classified as democratic by the dichotomized polyarchy variable – are countries that do not meet the minimalistic criteria used for binary classifications of democracy, while having polyarchy of at least 0.42. This is a much rarer situation than false negatives, but the two countries with the most false positives are Namibia and Senegal.

Namibia yields a false positive from 1991 until 2010, and notably the binary datasets are quite split on whether or not Namibia is a democracy in this period, with those who classify it as an autocracy narrowly "out-voting" those who classify it as a democracy. However, during these decades its polyarchy value ranges from 0.601 to 0.7, which is substantially larger than the cutpoint of 0.42. Taken together with the prominence of Botswana in Table 2, one feature stands out as the likely reason that the binary classifications are divided on these cases: the turnover rule. Can a country be a democracy if it has never experienced a transition from one party to another? Authors who follow Przeworski et al. (2000) in answering "no" will classify Namibia and Botswana as autocratic, since they have been governed respectively by the SWAPO Party and the Botswana Democratic Party since independence; authors who answer "yes" may or may not code these cases as democratic.

The prominence of Senegal in Table 5 lends further credence to the idea that alternation of power is the cause of many false positives. Senegal is classified as an autocracy by three binary datasets through the Senegambia period of the 1980s until the end of the 1990s, but consistently has a polyarchy greater than 0.42. Notably, from 2000 onwards, four binary datasets classify Senegal as a democracy. Senegal must have failed one of the binary datasets' requirements for democracy throughout the 1980s and 1990s, but suddenly satisfied it in 2000; the substantive interpretation is clear, since the 2000 election in Senegal was the first time that the country experienced a peaceful transfer of power between parties. In this case, V-Dem's polyarchy variable was a leading indicator of a democratic spell; the binary datasets could only "detect" democracy after the change in government in 2000. The many-valued nature of polyarchy may have permitted V-Dem's country expert coders to record a rise in democracy before the change was observable through strict binary classifications.

Comparing the countries that have the most false negatives to the cases that have the most false positives reveals an interesting picture of the differences between polyarchy and the majority opinion of the binary classifications. Of the ten countries that have the most false negatives, five are Latin American, and three are periods in the distant history of long-standing European or North American democracies. Evidently, the binary datasets are more generous in classifying these cases as democracies than V-Dem is in assigning them a high polyarchy value. In contrast, of the ten countries with the most false negatives, seven are in Africa. V-Dem is substantially more likely to assign high polyarchy to African countries than the binary datasets are to classify them as democratic.

Now that we have studied the properties of the consensus cutpoint dataset and its relationship to the binary classifications of democracy, we will use it to replicate some past analyses.

4 Using the consensus dataset

The tremendously similar cutpoints would be, in the words of Boix et al. (2012), "merely academic" if they did not enable us to produce a dataset that replicates the major empirical results of the datsets individually. In this section we pick a central empirical result from each of Boix et al. (2012), Cheibub et al. (2010), and Svolik (2008), and we show that several core results can be replicated with the V-Dem dataset dichotomized using the optimal cutpoint.[11]

Recall that we are certainly not arguing that all binary codings of democracy are redundant, or that a dichotomized version of polyarchy can replace any binary coding of democracy in any statistical analysis and produce the same results. Our point is simply that the field already contains a consensus dataset of such flexibility that it can roughly replicate the main results that motivated a series of different datasets.

For all of these replications, we continue to use a "consensus" cutpoint of 0.42, since it is around both the median and the modal value in Table 3. Note that we use the same cutpoint of $\tau = 0.42$ for every dataset, rather than using the optimal cutpoint for each individual dataset. The reason is that, if a researcher is adjudicating between using the "consensus" dataset and coding a new dataset, they cannot know what will be the optimal cutpoint for their new dataset before they code it, so they have to use the best existing consensus dataset. Our findings lead us to suggest that a very promising consensus dataset is found with cutpoint $\tau = 0.42$.

4.1 Replicating Boix et al. (2012)

The primary purpose of Boix et al. (2012) is to present a dataset, but together with that dataset the authors include descriptive plots of the association between their democratic measure and three core variables: GDP per capita, land equality, and latitude. They find that "economic modernization variables have steadily declined in their correlation with democracy over time".

In Figure 13 we replicate each of those main plots. In the interest of an independent replication, we use the GDP per capita value from Coppedge et al. (2019) rather than the one used in Boix et al. (2012), and rather than the land equality value used in Boix et al. (2012) we use the urbanisation value from Coppedge et al. (2019) (urbanisation and land equality being distinct but closely related ideas, this substition turns out to be sufficient to

[11] These replications are based only on the authors' publicly available datasets. In some cases we wrote our own code to replicate the authors' descriptions of their analyses. This is why the precise numbers that we report sometimes differ in substantively unimportant ways from the numbers that the original authors report. In a few cases, the publicly available datasets did not contain some of the covariates that were named in the reported analyses. In these cases, we found variables elsewhere that had the same substantive meaning and similar coverage. We were able to reproduce every substantive claim that any author reported using their datasets and our own implementation of their analysis. Using the dichotomized polyarchy dataset, we were able to replicate nearly every substantive claim that any author reported. This is a far stronger replication than if we had simply used the authors' code on a different dataset, because we were able to validate the original substantive results using our own independent implementation of the reported analyses.

replicate the exact same pattern and substantive conclusions as in the original paper). Since latitude is not sensitive to measurement decisions in the way that GDP per capita and land equality are, we simply use the same latitude value as Boix et al. (2012).

[Figure 13 about here]

We very closely replicate the pattern in Boix et al. (2012) that economic variables have declined in their capacity to predict level of democracy; the polyarchy dataset would give the same conclusion. Just one timespan is an exception: the figures presented in Boix et al. (2012) only begin in the 1830s, and with the V-Dem variables' time coverage we see that the polyarchy measure is less well-predicted by GDP/capita in the mid-1800s than the democracy measure of Boix et al. (2012).

Urbanization behaves slightly differently from land equality but in ways which do not affect the overall pattern of the results, and here polyarchy and the democracy variable of Boix et al. (2012) behave almost identically. Latitude also very closely resembles the figure in Boix et al. (2012), up to units, and the coefficients of the dichotomized polyarchy variable depart only slightly from Boix's classification.

The consensus dichotomization of polyarchy yields the same qualitative conclusions as the dichotomous variable in Boix et al. (2012) regarding these motivating concepts. In every case, the coefficient estimate for the dichotomized polyarchy variable is within the confidence interval of the coefficient estimate for the variable from Boix et al. (2012), and very often the two coefficients have almost exactly the same value.

4.2 Replicating Cheibub et al. (2010)

The presentation of the dataset in Cheibub et al. (2010) is organised around a series of replications of past work to study theoretical subtleties of their measure of democracy with respect to previous measures. The first of these is a replication of a main result in Fearon and Laitin (2003). In Table 6 we show that the dichotomized polyarchy dataset behaves exactly as the dataset in Cheibub et al. (2010) does: exactly the same variables are significant, with the same signs and almost exactly the same magnitudes.

[Table 6 about here]

The consensus dataset is just as successful as the Cheibub et al. (2010) dataset in replicating the results of Fearon and Laitin (2003).

4.3 Replicating Svolik (2008)

Replicating Svolik (2008) might seem to be a more difficult challenge for the consensus dataset, since in order to find an optimal cutpoint we departed from Svolik (2008) in considering transitional democracies to be non-democratic. But note that the dichotomization of polyarchy has no inherent meaning. We have been referring to this dichotomization as a dichotomization into democracies and autocracies, but really it is simply splitting polyarchy above and below a certain value to match whatever binary variable the author's substantive coding produced. In this case, our polyarchy dichotomization is matching Svolik's coding of consolidated and transitional democracies as closely as possible. So long as we can assume that consolidated democracies tend to be more democratic than transitional democracies, the dichotomization method should apply to this case as well as it does to dichotomization into democracies and autocracies. And that assumption is more or less justified; there is an 84% pairwise agreement among the cases shared between Svolik's classification and polyarchy dichotomized at 0.42.

In Table 7 and Table 8, we show a partial replication of Table 2 from Svolik (2008), with polyarchy dichotomized according to the optimal cutpoint $\tau^* = 0.42$.

[Table 7 about here]

[Table 8 about here]

One of the central findings in Svolik (2008) is "that the level of economic development, type of democratic executive, and type of authoritarian past determine whether a democracy consolidates, but have no effect on the timing of reversals in democracies that are not consolidated." Concerning the timing of reversals, every non-economic variable that is significant in Table 2 of Svolik (2008) is also significant when using a dichotomized polyarchy variable, and retains the same sign and a similar magnitude. Regarding economic variables, in several of the columns of Table 7 economic recessions are significant when using the dichotomized polyarchy variable, whereas in Table 2 of Svolik (2008) economic development was significant. The finding in Svolik (2008) is that "the level of economic development determines the extent to which a democracy is susceptible to the risk of a reversal, but the eventual timing of a reversal is only associated with economic recessions", and the dichotomized polyarchy finds the reverse.

However, in Table 8, we show that dichotomized polyarchy successfully retains the finding that the level of economic development determines whether or not a country consolidates, with the same sign and roughly the same magnitude as in Svolik (2008). But it does not retain significance for the type of executive and authoritarian past.

In the case of Svolik (2008), replication by the consensus dataset is a mixed success. Why is it not completely successful? Svolik's theory and hypothesis tests are tightly focused on the relationship between institutions and democracy *within democracies themselves*, and his consolidation variable is a model-based classification of types of democracies rather than a coding of the often clear-cut distinction between democracies and autocracies. It is easy to imagine that such a distinction tracks less closely with polyarchy than the difference between democracies and autocracies. This gives a useful insight into the limitations of dichotomizing V-Dem; to mimic situations in which the author is not making a simple judgment about the difference between democracies and autocracies, it may not be sufficient to dichotomize polyarchy. In situations like this, caution is warranted.

5 Conclusion

We demonstrated a strong numerical regularity across the major classifications of democracies and autocracies, by dichotomizing many-valued measures of democracy to match binary classifications of democracy. This strong regularity is compelling evidence that, despite the common description of this topic of study as one that broadly lacks agreement, there is actually a strong empirical consensus over which countries have been democratic in which years. We used this result to generate a binary variable that represents the consensus position of the field so far, since it agrees quite closely with a wide variety of past measures of democracy. This consensus dataset successfully replicated the core results that motivated the datasets of Boix et al. (2012) and Cheibub et al. (2010) and it partially replicated the results of Svolik (2008).

These results suggest that most new empirical analyses can simply use the consensus dataset rather than constructing a new classification. Unless a researcher has a sharply different coding rule when they create a new classification of democracy, there is compelling reason to expect that any new classification of democracies is likely to closely resemble the consensus classification, both in the decisions it makes about specific country-years and in the empirical results it yields.

6 Figures and Tables

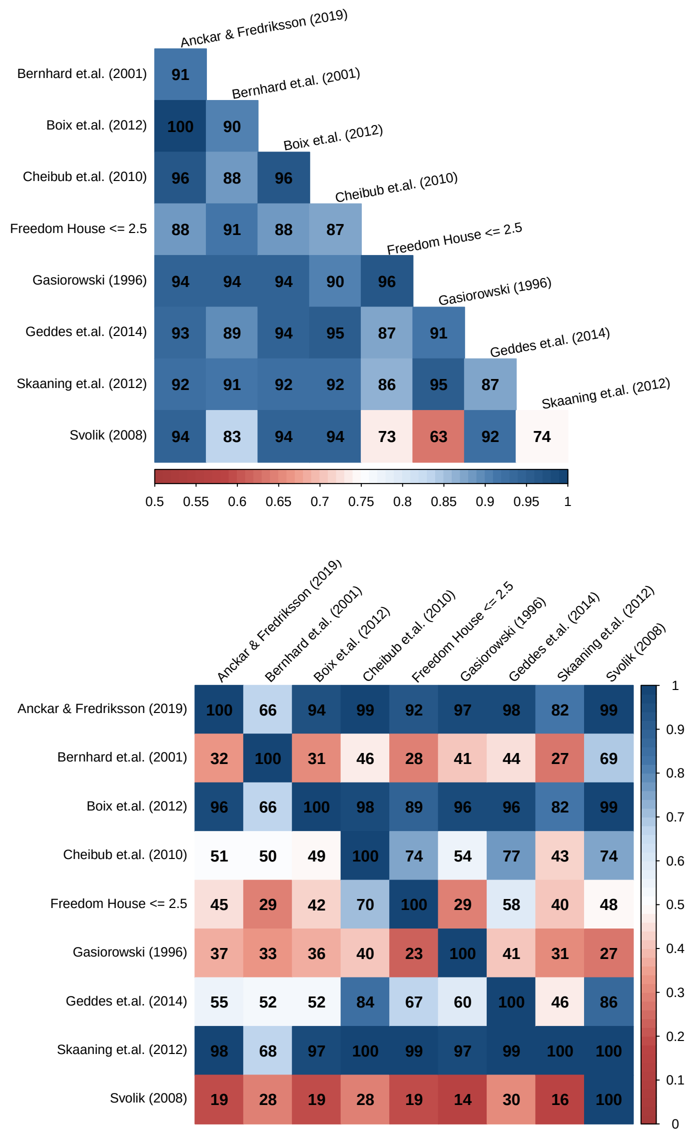

Figure 1: Top: Each row containing the name of a dataset shows the pairwise agreement between the dichotomous measure of democracy in that dataset and the dichotomous measure of democracy in the dataset in each column. Bottom: each cell represents the percentage of the dataset in that column which also appears in the dataset in that row. For example, Boix et al. (2012) has 94% agreement with Geddes et al. (2014), 52% of the cases that appear in Boix et al. (2012) also appear in Geddes et al. (2014), and 96% of the in Geddes et al. (2014) are in Boix et al. (2012). In the case of Svolik (2008) we treat transitional democracies as democracies for the purposes of comparing Svolik's coding to the other authors' codings.

RETURN TO TEXT

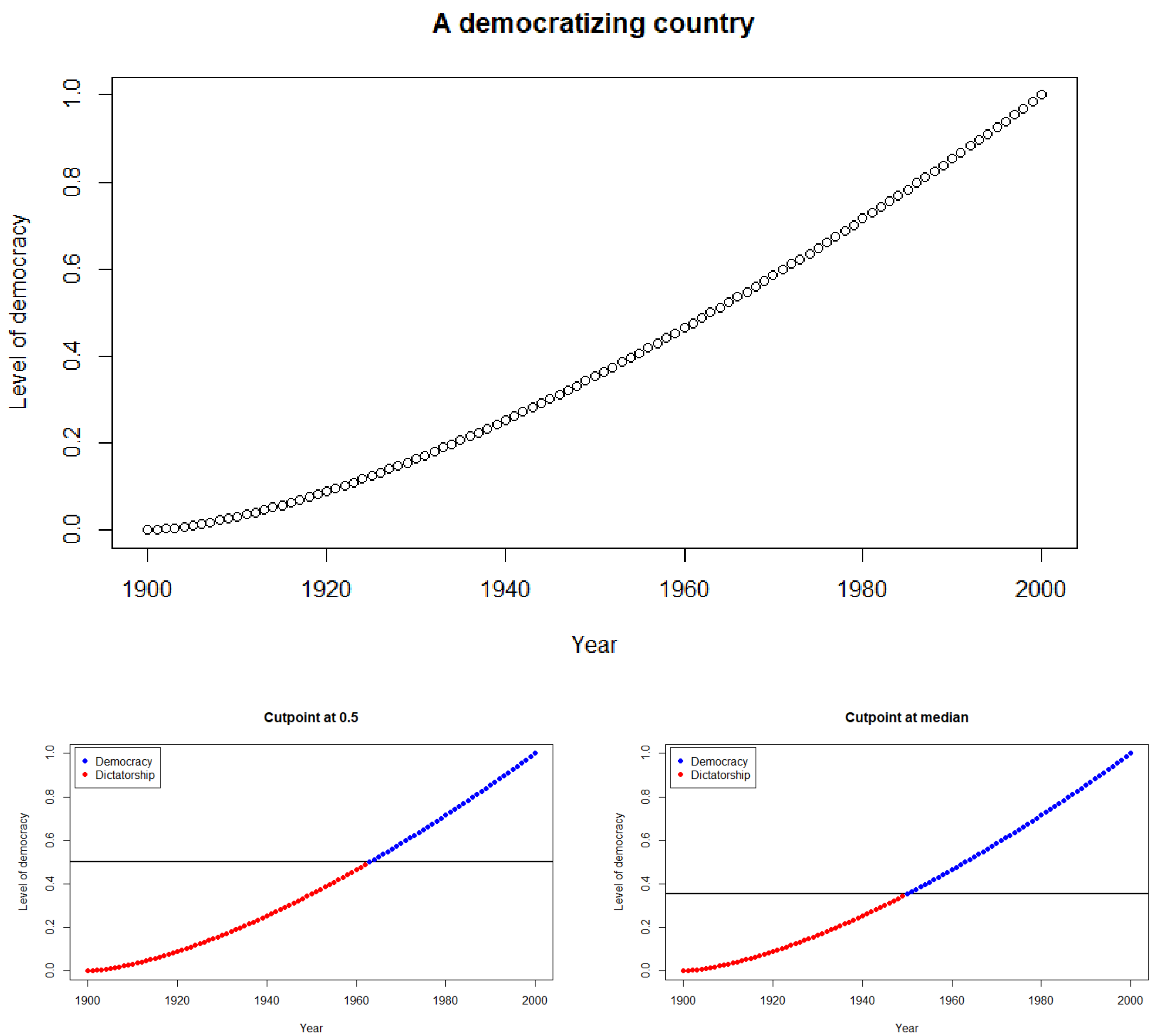

Figure 2: A fictional example to illustrate the idea of cutpoints.

RETURN TO TEXT

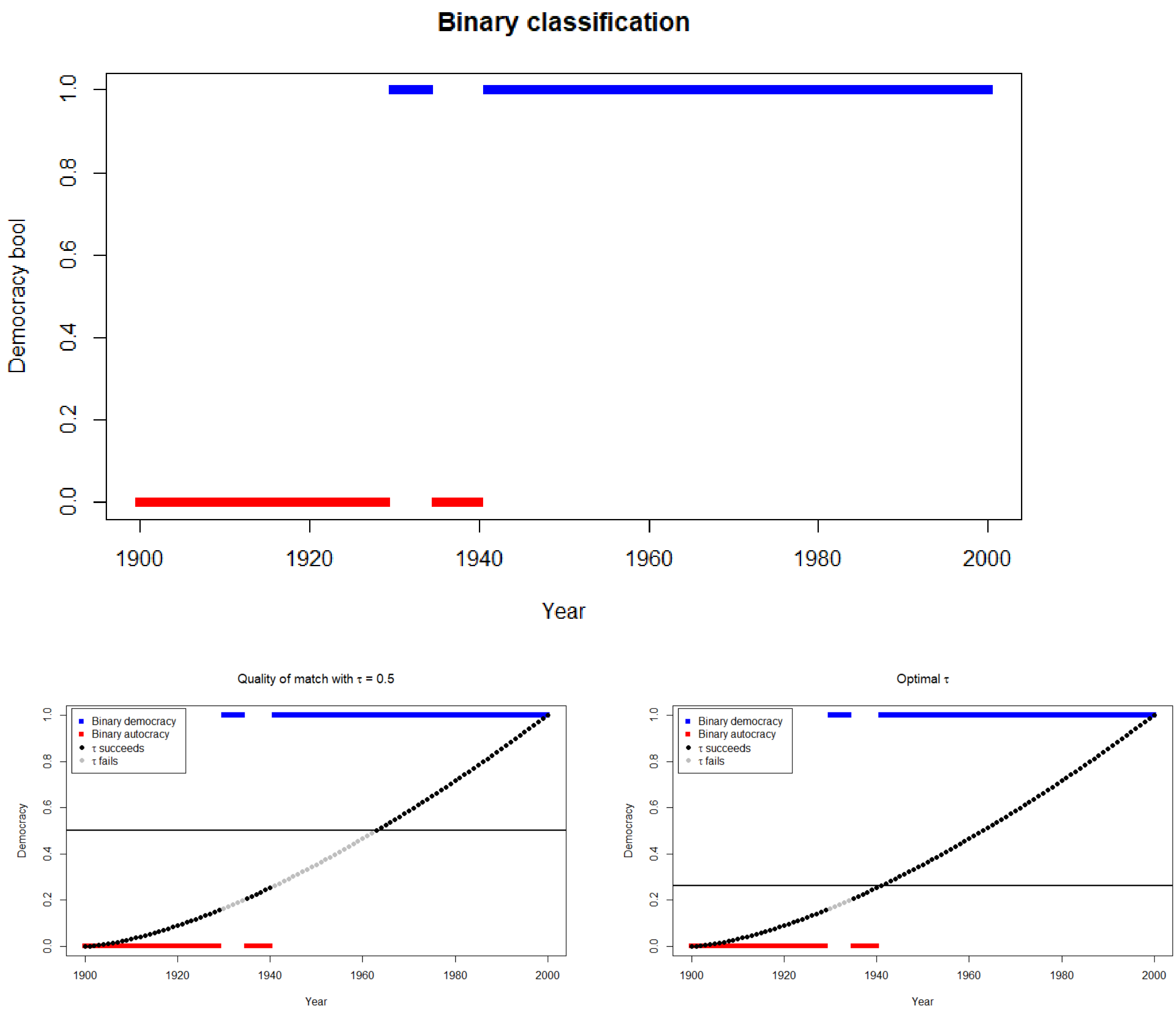

Figure 3: A fictional example of the optimal cutpoint for matching a many-valued classification to a binary classification.

RETURN TO TEXT

Dataset	Requirements for full democracy
Boix et al. (2012)	All of:
	↪Executive elected in popular elections
	↪Executive responsible to voters or legislature
	↪Legislature chosen in free and fair elections
	↪Most adult men have voting rights
Cheibub et al. (2010)	All of:
	↪Free and fair elections
	↪Minimum level of suffrage
Geddes et al. (2014)	None of:
	↪Executive in power without direct, fair, competitive election
	↪Rule change limiting electoral competition
	↪Major parties blocked from elections by military
Puddington et al. (2018)	Expert survey with discussion and review
Svolik (2008)	If backslid, then transitional
	If never backslid, estimate probability of transitional

Table 1: The binary datasets' rules for coding a country as a democracy. If it does not fulfill these criteria, then it is an autocracy. In the case of Puddington et al. (2018) no ruleset is given, so we instead state the process by which codings are done. This dataset is also originally several-valued, but we use a binarised version of it. In the case of Svolik (2008), the requirements stated are for a transitional rather than a consolidated democracy. Svolik takes these cases to be democratic, and we follow Svolik in considering transitional democracies to be democracies when comparing Svolik's codings to other authors'.

RETURN TO TEXT

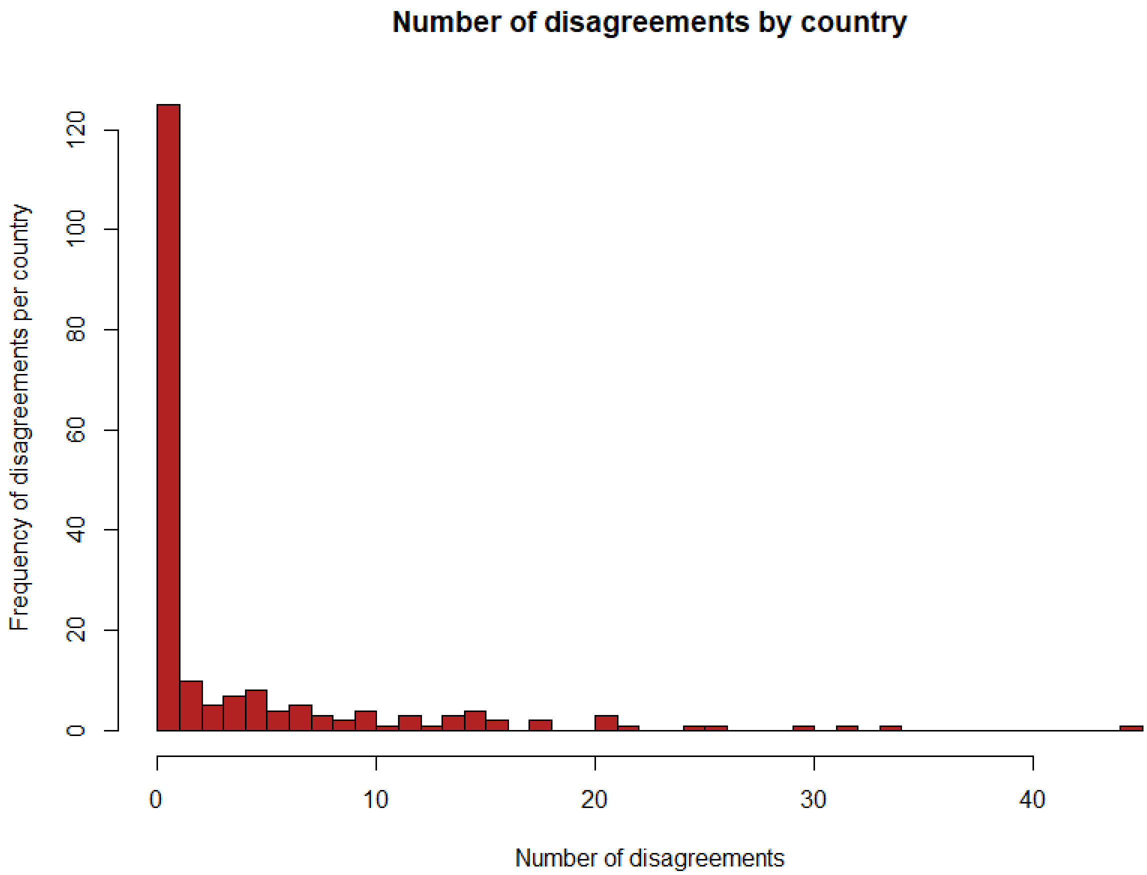

Figure 4: Histogram of the number of years for which each country had a disagreement over whether it was democratic or autocratic in that year. The majority of cases never had a year that one author classified as democratic and another author classified as autocratic.

RETURN TO TEXT

Country	Number of disagreements	Years
Botswana	45	1966 – 2010
Guatemala	34	1954, 1958 – 1963, 1966 –1982, 1986 – 1995
Lebanon	32	1944 – 1970, 1975, 1976, 2006 – 2008
Costa Rica	30	1920 – 1949
The Netherlands	26	1871 – 1896
Fiji	25	1970 – 1986, 1992 – 1999
The Gambia	22	1972 – 1993
Argentina	21	1946 – 1966, 1973, 1976, 1983
Lesotho	21	1966 – 1970, 1993 – 2008
Namibia	21	1990 – 2010
Armenia	18	1991 – 2008
Guyana	18	1972, 1994 – 2008
Ghana	16	1957 – 1960, 1969, 1972, 1979, 1981, 1993 – 2000
Pakistan	16	1947 – 1957, 1976, 1977, 1988, 1999, 2008
El Salvador	15	1972 – 1975, 1984 – 1994
Haiti	15	1947 – 1950, 1991, 1994 – 2001, 2007, 2008
South Africa	15	1994 – 2008
Thailand	15	1975, 1976, 1979 – 1988, 1991, 1992, 2006
Kyrgyzstan	14	1992 – 2001, 2005 – 2008
Paraguay	14	1989 – 2002
Zambia	14	1965 – 1967, 1991 – 2001
Peru	13	Many individual years
Nigeria	12	1966, 1979, 1983, 1999 – 2007
Panama	12	1946 – 1952, 1954, 1955, 1968, 1989, 1990
Sri Lanka	12	1977 – 1982, 1989 – 1994
Guinea-Bissau	11	1994 – 1997, 2000 – 2007

Table 2: Countries that have more than 10 years which were coded as a democracy in at least one dataset while being coded as an autocracy in at least one other dataset

RETURN TO TEXT

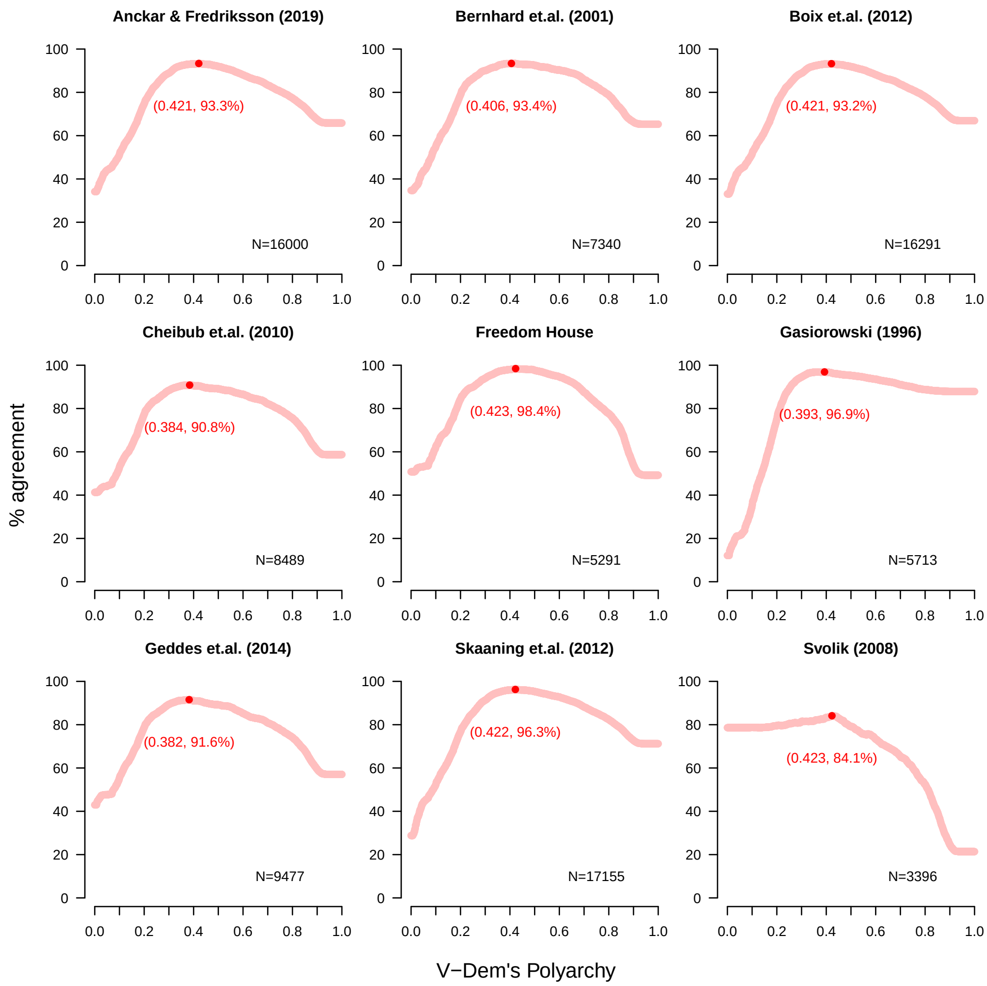

Figure 5: Accuracy of polyarchy using every cutpoint for each dataset (deleting ambiguous democratic categories, like "semi-democracies" and "transitions"). The dot represents the optimal cutpoint, and the cutpoint value is displayed, alongside the percent agreement that the binary dataset has with polyarchy binarized at that optimal cutpoint value.

RETURN TO TEXT

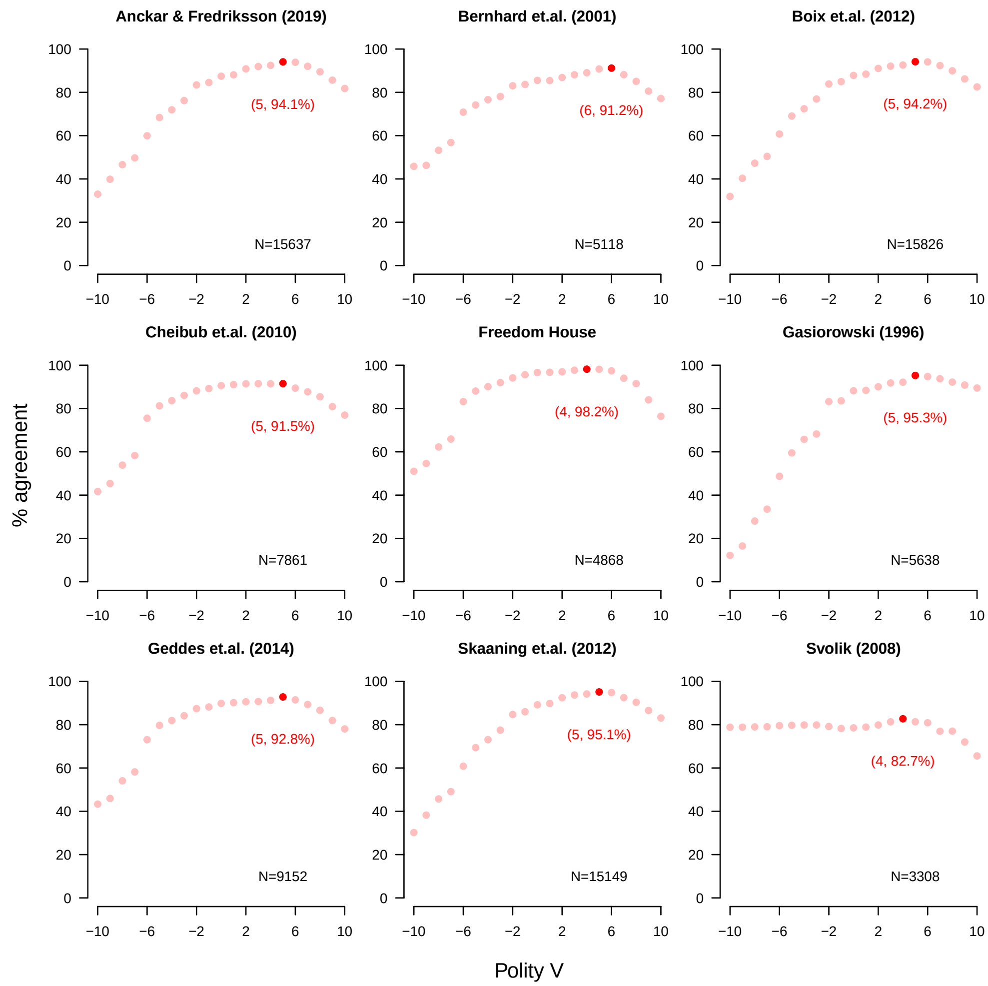

Figure 6: Accuracy of polity using every cutpoint for each dataset (deleting ambiguous democratic categories, like "semi-democracies" and "transitions"). The dot represents the optimal cutpoint, and the cutpoint value is displayed, alongside the percent agreement that the binary dataset has with polity binarized at that optimal cutpoint value.

RETURN TO TEXT

Dataset	Polyarchy			Polity		
	τ^*	% Agree	N	τ^*	% Agree	N
Geddes et al. (2014)	0.382	92%	9477	5	93%	9152
Cheibub et al. (2010)	0.384	91%	8489	5	92%	7861
Boix et al. (2012)	0.421	93%	16291	5	94%	15826
Svolik (2008)	0.423	84%	3396	4	83%	3308
Puddington et al. (2018)	0.423	98%	5291	4	98%	4868

Table 3: Optimal cutpoints for the core dichotomous measures. On the left, the cutpoint that dichotomizes the polyarchy score to best match each dataset, and the percent agreement that it achieves. On the right, the same information for polity instead of polyarchy.

RETURN TO TEXT

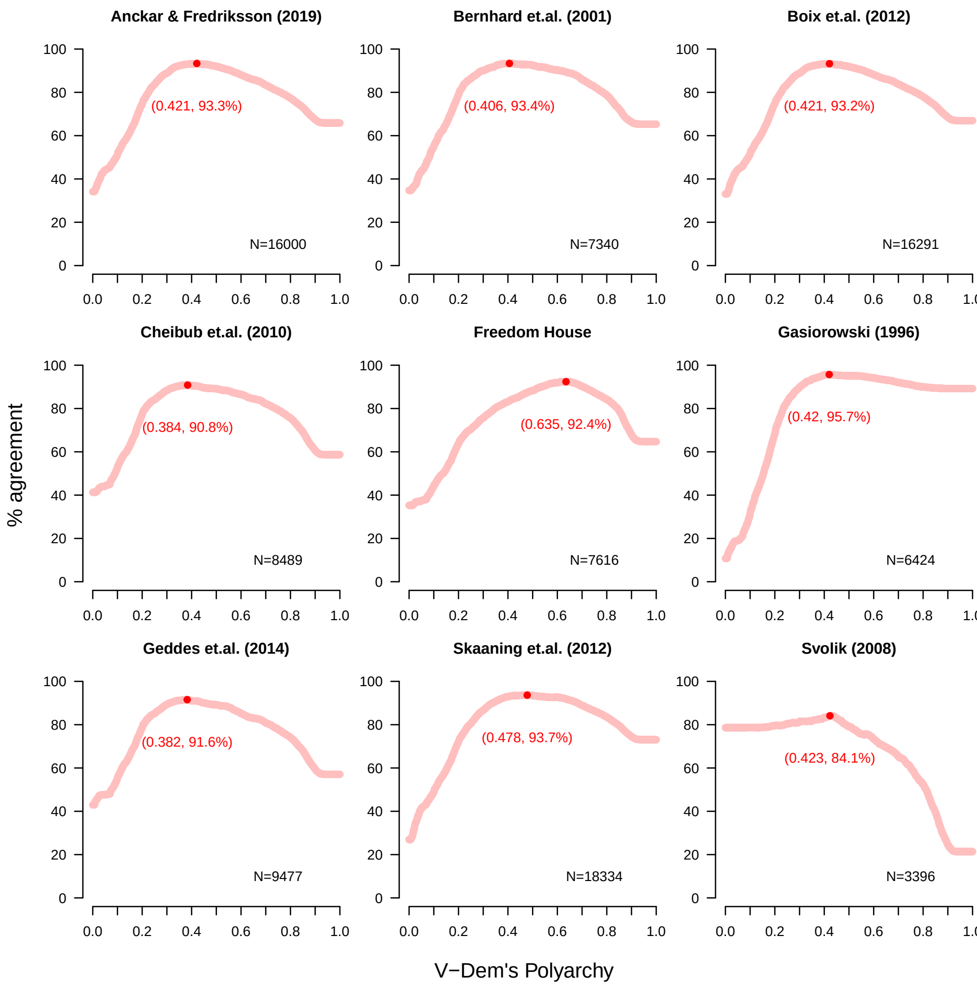

Figure 7: Accuracy of polyarchy using every cutpoint for each dataset (all pair-wise overlapping cases included). The dot represents the optimal cutpoint, and the cutpoint value is displayed, alongside the percent agreement that the binary dataset has with polyarchy binarized at that optimal cutpoint value.

RETURN TO TEXT

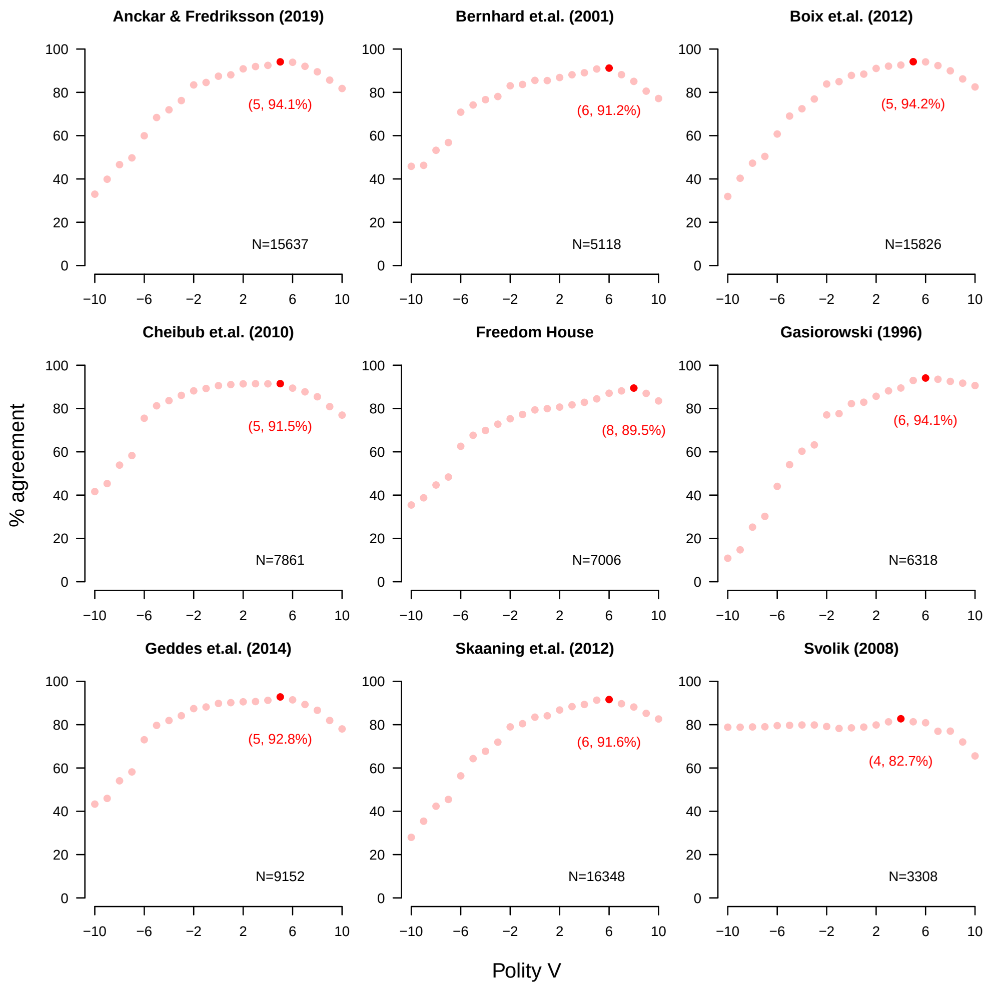

Figure 8: Accuracy of polity using every cutpoint for each dataset (all pair-wise overlapping cases included). The dot represents the optimal cutpoint, and the cutpoint value is displayed, alongside the percent agreement that the binary dataset has with polity binarized at that optimal cutpoint value.

RETURN TO TEXT

Polity dichotomization	Polyarchy τ^*	% Agree
-9	0.04	88%
-8	0.04	85%
-7	0.18	80%
-6	0.21	81%
-5	0.21	83%
-4	0.24	82%
-3	0.29	86%
-2	0.32	86%
-1	0.34	87%
0	0.34	87%
1	0.34	89%
2	0.4	91%
3	0.4	92%
4	0.42	93%
5	0.45	93%
6	0.47	92%
7	0.56	92%
8	0.69	91%
9	0.79	92%

Table 4: Optimal V-Dem cutpoints for matching onto the binary datasets that are created when we take every possible dichotomization of the polity score. We have normalised the polity score to run from 0 to 1, and excluded the bounds. The proportion of agreement is also provided, though a high proportion of agreement does not mean that the dataset corresponds to any substantive reality; in cases nearer to the bounds, it typically means that a polity dichotomization that considers every case to be an autocracy (or democracy) corresponds extremely closely to a polyarchy dichotomization that considers every case to be an autocracy (or democracy).

RETURN TO TEXT

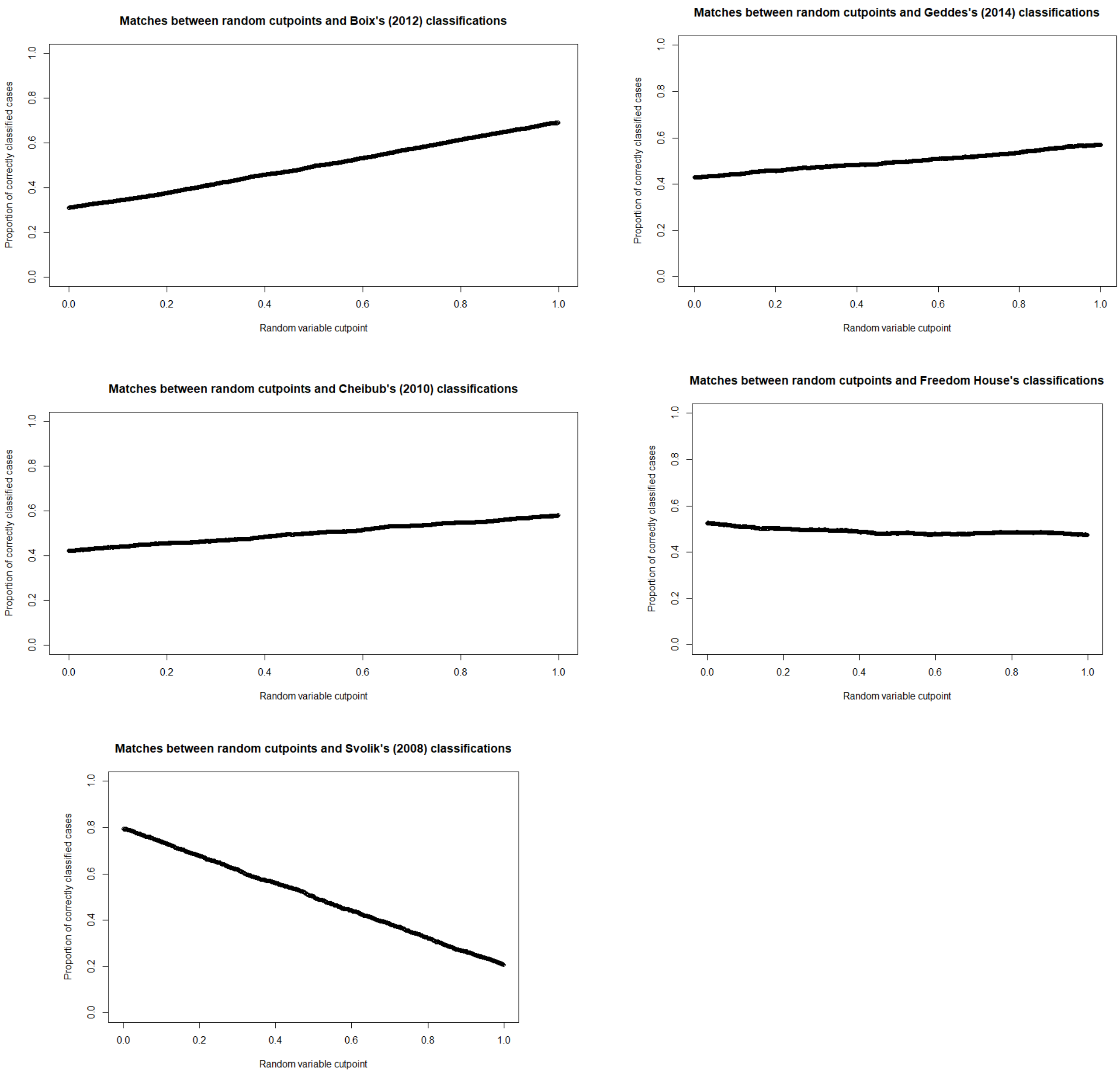

Figure 9: Accuracy of a uniform random variable every cutpoint for each dataset.

RETURN TO TEXT

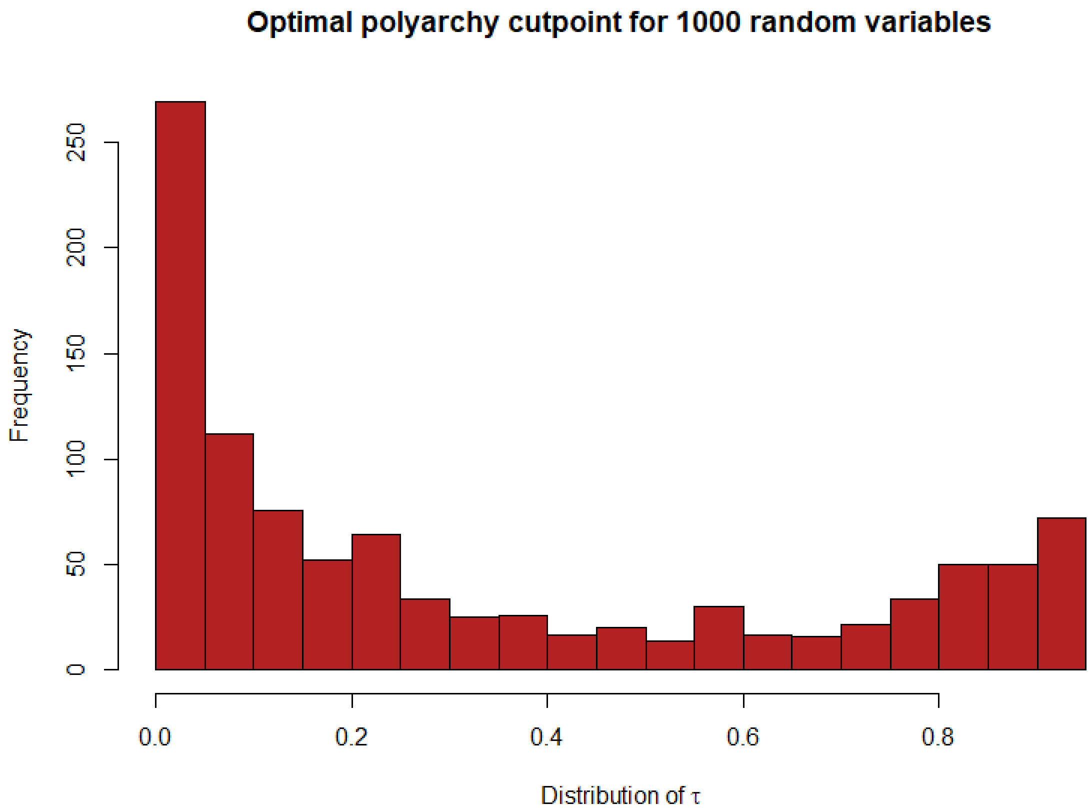

Figure 10: Distribution of optimal cutpoints to match polyarchy to a uniform random variable. Ties between multiple optimal cutpoints are resolved randomly.

RETURN TO TEXT

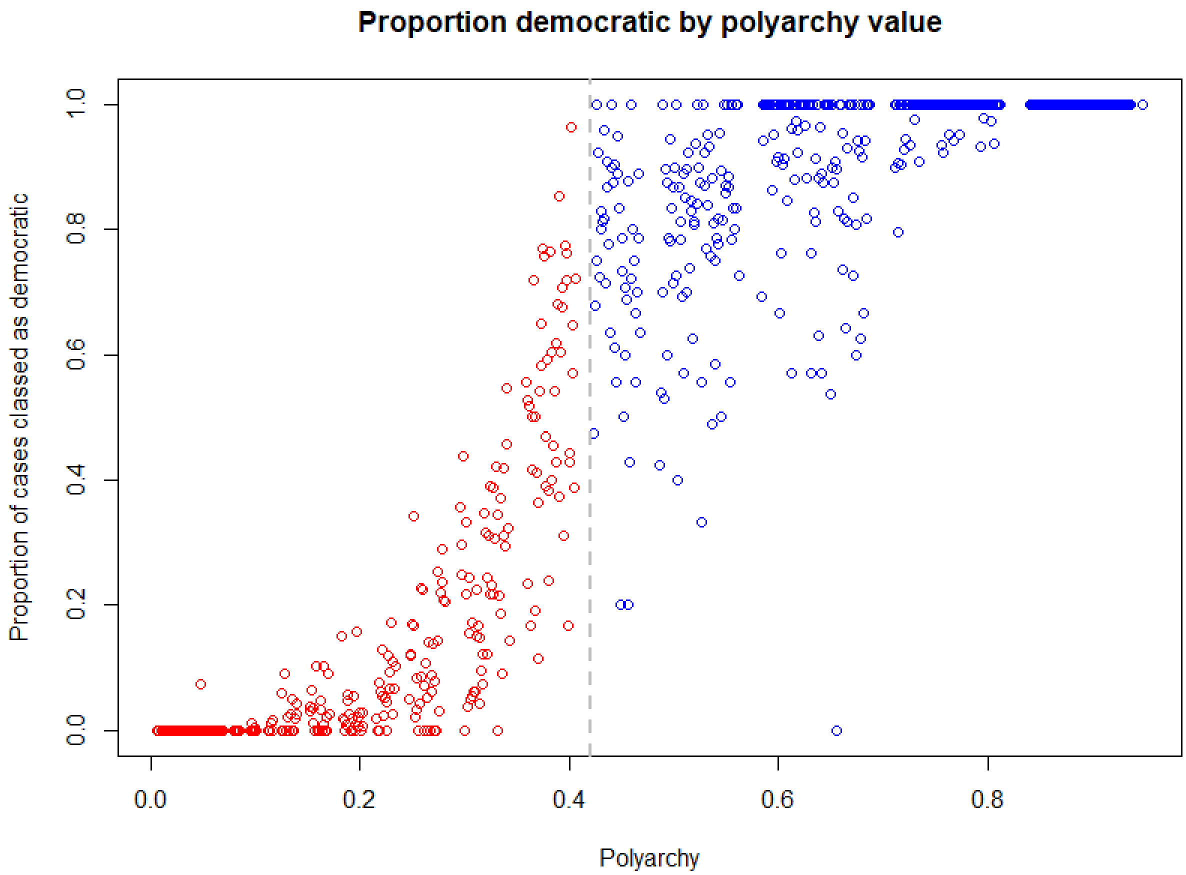

Figure 11: Proportion of cases in all datasets classified as democratic by polyarchy. The cases are coloured according to the cutpoint $\tau = 0.42$, which is marked with a vertical line. To obtain the proportion for each polyarchy value, we count the number of times that any observation at that polyarchy value was classed as a democracy, and divide that by the total number of times that any observation at that polyarchy value was included in a binary dataset. This plot shows the proportions for the entire range of polyarchy (1001 values) for $N = 16,328$ country-years that were contained in Coppedge et al. (2019) and any one of the five dichotomous datasets, double-counting any observations that occured in multiple datasets. So the values that go into the proportions are not distinct country-years, but rather distinct classifications of any country-year, and this plot contains one point for each value of polyarchy that corresponds to a country-year which was classified in at least one dataset.

RETURN TO TEXT

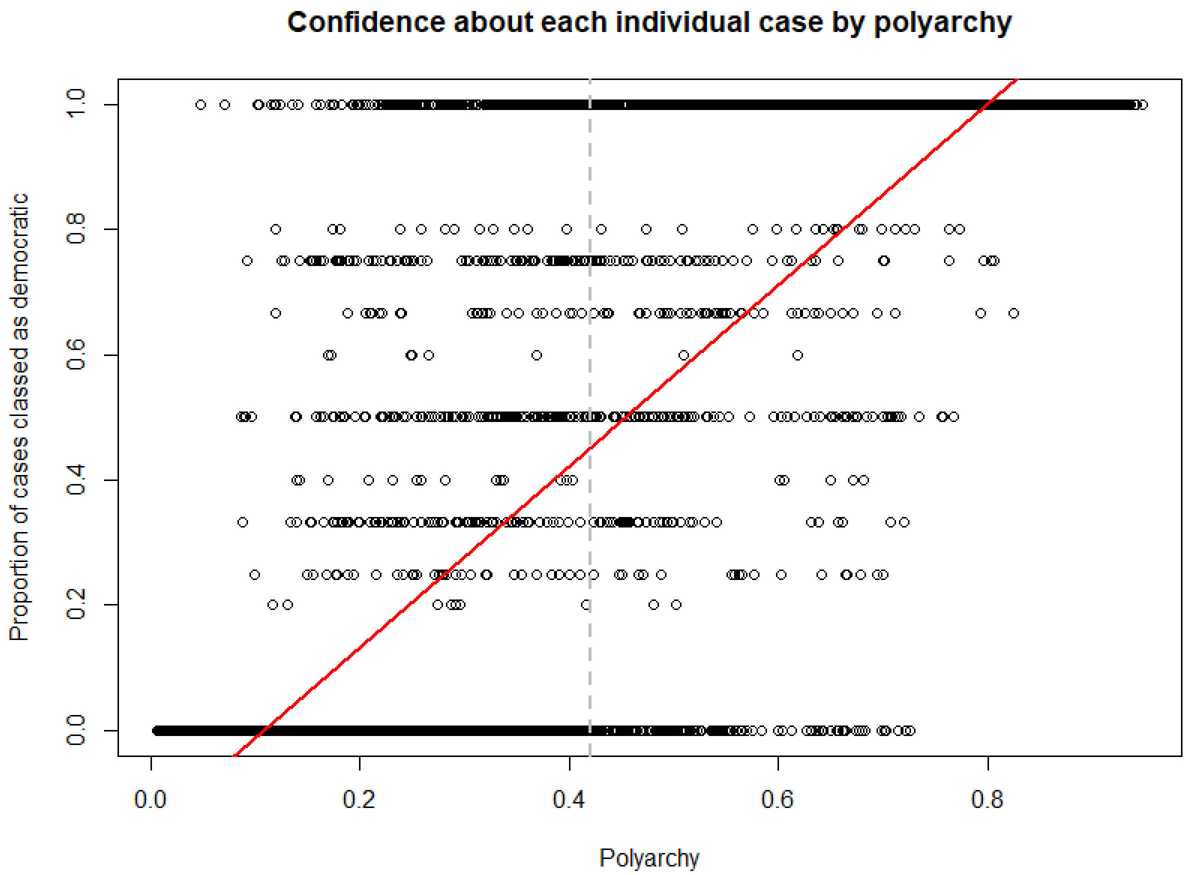

Figure 12: Proportion of datasets which classify each country-year as a democracy is indeed strongly increasing in polyarchy. Because the density of dots is difficult to visually assess, we plot a trend line, which is a simple linear regression where the dependent variable is the proportion of times that each country was classified as a democracy and the single independent variable is polyarchy. The trend is sharply increasing. A vertical line marks the optimal cutpoint $\tau = 0.42$.

RETURN TO TEXT

Country	False negatives	False positives
United States of America	96	
Greece	78	
Guatemala	37	
Chile	34	
Colombia	29	
Lebanon	25	
Pakistan	20	
Norway	19	
Panama	19	
Ecuador	16	
Namibia		20
Senegal		20
Burkina Faso		19
Mozambique		17
Tanzania		17
Cyprus		16
Zambia		14
Bosnia and Herzegovina		13
Sri Lanka		12
Côte d'Ivoire		11

Table 5: The 10 countries with the most false negatives (years for which a majority of binary datasets considered them to be democracies but the polyarchy cutpoint classified them as autocracies), and the 10 countries with the most false positives (years for which a majority of binary datasets considered them to be autocracies but the polyarchy cutpoint classified them as democracies). The optimal cutpoint is seen to more frequently classify binary democracies as autocracies than to classify binary autocracies as democracies.

RETURN TO TEXT

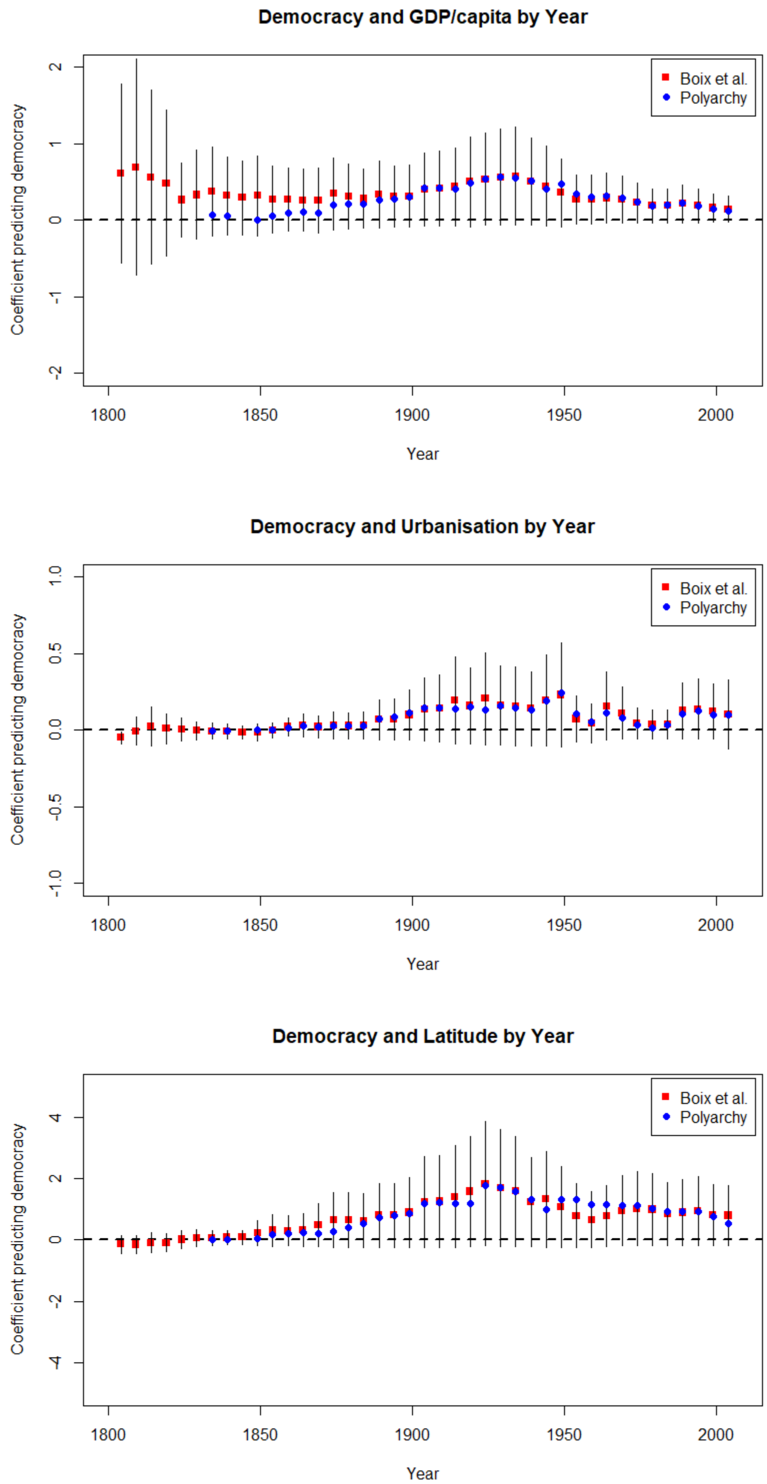

Figure 13: A replication of Boix et al. (2012) always falls within its confidence intervals. The years that do not appear have too few observations.

RETURN TO TEXT

	Fearon and Laitin (2003)	Cheibub et al. (2010)	Polyarchy replication
Anocracy	**0.54**		
	(0.24)		
Democracy	0.11		
	(0.31)		
Dictatorship with legislature		-0.36	-0.40
		(0.21)	(0.21)
Instability	**0.53**		
	(0.24)		
Prior war	**-0.84**	**-0.85**	**-0.87**
	(0.31)	(0.32)	(0.32)
GDP/capita	**-0.31**	**-0.34**	**-0.35**
	(0.07)	(0.07)	(0.07)
Log(population)	**0.27**	**0.25**	**0.26**
	(0.07)	(0.07)	(0.07)
Log(% mountainous)	**0.20**	**0.24**	**0.25**
	(0.09)	(0.09)	(0.09)
Noncontinuous state	0.33	0.37	0.38
	(0.28)	(0.28)	(0.28)
Oil exporter	**0.79**	**0.93**	**0.92**
	(0.28)	(0.27)	(0.27)
New state	**1.63**	**1.52**	**1.59**
	(0.35)	(0.34)	(0.34)
Ethnic fractionalization	0.15	0.10	0.11
	(0.37)	(0.37)	(0.37)
Religious fractionalization	0.43	0.52	0.48
	(0.51)	(0.51)	(0.51)
Constant	**-7.09**	**-6.48**	**-6.48**
	(0.76)	(0.73)	(0.73)
N	6,217	6,255	6,255

Table 6: Cheibub (2010) Table 2 replication. Bolded numbers are significant at $p < 0.05$.

RETURN TO TEXT

	Split L-l	Simple L-l	Split Weibull	Split & Frailty L-l
GDP/capita	0.09 → **0.17**	**0.34** → **0.35**	0.12 → **0.17**	0.09 → **0.17**
	(0.08) → (0.07)	(0.06) → (0.06)	(0.07) → (0.07)	(0.08) → (0.07)
GDP growth	**0.05** → 0.03	**0.03** → 0.02	**0.04** → 0.02	**0.05** → 0.03
	(0.02) → (0.01)	(0.01) → (0.02)	(0.01) → (0.13)	(0.02) → (0.01)
Military	-0.30 → -0.45	**-0.88** → **-1.16**	-0.31 → -0.33	-0.287 → -0.45
	(0.32) → (0.31)	(0.32) → (0.34)	(0.27) → (0.29)	(0.32) → (0.31)
Civilian	0.14 → 0.09	-0.08 → -0.29	0.07 → 0.18	0.14 → 0.09
	(0.34) → (0.33)	(0.33) → (0.35)	(0.31) → (0.32)	(0.34) → (0.33)
Monarchy	0.93 → 0.59	-0.10 → -0.49	**0.98** → **0.83**	0.93 → 0.59
	(0.53) → (0.48)	(0.56) → (0.53)	(0.45) → (0.41)	(0.53) → (0.48)
Parliamentary	-0.29 → -0.27	-0.02 → 0.21	-0.25 → -0.17	-0.29 → -0.27
	(0.31) → (0.28)	(0.33) → (0.31)	(0.32) → (0.27)	(0.32) → (0.28)
Presidential	0.39 → 0.32	-0.01 → 0.17	0.37 → 0.38	0.39 → 0.32
	(0.29) → (0.25)	(0.33) → (0.31)	(0.31) → (0.26)	(0.29) → (0.25)
Intercept	**2.30** → **2.18**	**2.56** → **2.54**	**2.59** → **2.30**	**2.30** → **2.18**
	(0.40) → (0.37)	(0.40) → (0.41)	(0.38) → (0.36)	(0.40) → (0.37)
Shape α	**1.94** → **2.03**	**1.60** → **1.62**	**-0.38** → **-0.42**	**1.94** → **2.03**
	(0.21) → (0.22)	(0.17) → (0.17)	(0.10) → (0.11)	(0.21) → (0.22)

Table 7: Svolik (2008) reversal timing model: non-immune results with original and polyarchy-coded consolidation. Replicating Svolik (2008) by replacing the c-value of every country-year with polyarchy ≤ 0.42 by $c = 1$ and every country-year with polyarchy > 0.42 by $c = 0$. Svolik's value → replicated value. "L-l" stands for "Log-logistic". Bolded numbers are significant at $p < 0.05$

RETURN TO TEXT

	Split L-l	Split Weibull	Split & Frailty L-l
GDP/capita	**2.12** → **1.53**	**2.04** → **1.46**	**2.12** → **1.53**
	(0.59) → (0.55)	(0.55) → (0.57)	(0.59) → (0.55)
GDP growth	-0.01 → 0.21	-0.05 → 0.18	-0.01 → 0.21
	(0.23) → (0.25)	(0.25) → (0.26)	(0.22) → (0.25)
Military	**-4.06** → -4.49	**-3.99** → -4.36	**-4.06** → -4.49
	(1.89) → (2.30)	(1.86) → (3.15)	(1.90) → (2.30)
Civilian	-0.42 → -1.03	-0.55 → -1.07	-0.42 → -1.03
	(1.10) → (1.05)	(1.07) → (1.04)	(1.09) → (1.05)
Monarchy	-15.8 → -14.8	-15.0 → -14.6	-28.0 → -14.8
	(897) → (836)	(896) → (726)	(2832) → (836)
Parliamentary	2.23 → 4.40	2.29 → 4.36	2.23 → 4.40
	(2.23) → (3.07)	(2.33) → (3.15)	(2.22) → (3.07)
Presidential	**-8.31** → -2.35	**-8.19** → -2.21	**-8.31** → -2.35
	(3.96) → (3.33)	(4.03) → (3.51)	(3.95) → (3.33)
Intercept	**-6.14** → -6.90	**-5.92** → -6.63	**-6.14** → -6.90
	(2.65) → (3.65)	(2.64) → (3.68)	(2.64) → (3.65)

Table 8: Svolik (2008) consolidation status model: immune results with original and polyarchy-coded consolidation. Replicating Svolik (2008) by replacing the c-value of every country-year with polyarchy ≤ 0.42 by $c = 1$ and every country-year with polyarchy > 0.42 by $c = 0$. Svolik's value → replicated value. "L-l" stands for "Log-logistic". Bolded numbers are significant at $p < 0.05$

RETURN TO TEXT

References

Alvarez, M., Cheibub, J. A., Limongi, F., and Przeworski, A. (1996). Classifying political regimes. *Studies in Comparative International Development*, 31:3–36.

Anckar, C. and Fredriksson, C. (2019). Classifying political regimes 1800–2016: a typology and a new dataset. *European Political Science*, 18(1):84–96.

Bernhard, M., Nordstrom, T., and Reenock, C. (2001). Economic performance, institutional intermediation, and democratic survival. *Journal of Politics*, 63(3):775–803.

Boese, V. A. (2019). How (not) to measure democracy. *International Area Studies Review*, 22(2).

Bogaards, M. (2010). Measures of democratization: From degree to type to war. *Political Research Quarterly*, 63(2):475–488.

Boix, C., Miller, M., and Rosato, S. (2012). A complete data set of political regimes, 1800-2007. *Comparative Political Studies*, XX(X):1–32.

Boix, C., Miller, M., and Rosato, S. (2018). Boix-Miller-Rosato Dichotomous Coding of Democracy, 1800-2015.

Boix, C. and Rosato, S. (2001). A complete dataset of political regimes, 1800-1999.

Bollen, K. and Jackman, R. (1989). Democracy, stability and dichotomies. *American Sociological Review*, 54(4).

Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, pages 370–390.

Bormann, N.-C. and Golder, M. (2013). Democratic electoral systems around the world, 1946–2011. *Electoral Studies*, 32(2):360–369.

Bowman, K., Lehoucq, F., and Mahoney, J. (2005). Measuring political democracy: Case expertise, data adequacy, and central america. *Comparative Political Studies*, 38(8):939–970.

Cheibub, J. A., Gandhi, J., and Vreeland, J. R. (2010). Democracy and dictatorship revisited. *Public Choice*, 143:67–101.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Bernhard, M., Fish, M. S., Glynn, A., Hicken, A., Lhrmann, A., Marquardt, K. L., McMann, K., Paxton, P., Pemstein, D., Seim, B., Sigman, R., Skaaning, S.-E., Staton, J., Cornell, A., Gastaldi, L., Gjerlw, H., Mechkova, V., von Rmer, J., Sundtrm, A., Tzelgov, E., Uberti, L., ting Wang, Y., Wig, T., and Ziblatt, D. (2019). V-dem dataset.

Cutright, P. (1963). National political development: Measurement and analysis. *American Sociological Review*, pages 253–264.

Dahl, R. A. (1973). *Polyarchy: Participation and opposition*. Yale University Press.

Fearon, J. D. and Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, 97(1):7590.

Gasiorowski, M. J. (1996). An overview of the political regime change dataset. *Comparative Political Studies*, 29(4):469–483.

Geddes, B., Wright, J., and Frantz, E. (2014). Autocratic breakdown and regime transitions. *Perspectives on Politics*, 12(2).

Gleditsch, K. S. and Ward, M. D. (2006). Diffusion and the international context of democratization. *International organization*, pages 911–933.

Kasuya, Y. and Mori, K. (2019). Better regime cutoffs for continuous democracy measures. *V-Dem working paper*.

Lerner, D. (1958). *Democracy and development: political institutions and well-being in the world, 1950-1990*. Free Press.

Lipset, S. M. (1960). *Political Man: The Social Bases of Politics*. Anchor.

Mainwaring, S., Brinks, D., and Pérez-Liñán, A. (2007). Classifying political regimes in latin america, 1945-2004. *Regimes and democracy in Latin America: Theories and methods*, pages 123–160.

Marshall, M. G., Gurr, T. R., and Jaggers, K. (2017). Polity iv project.

Mickey, R. (2015). *Paths Out of Dixie: The Democratization of Authoritarian Enclaves in America's Deep South, 1944-1972*. Princeton University Press.

Pemstein, D., Meserve, S. A., and Melton, J. (2010). Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis*, 18.

Popper, K. R. (2020). *The open society and its enemies*. Princeton University Press.

Przeworski, A., Alvarez, M., Cheibub, J. A., and Limongi, F. (2000). *Democracy and development: political institutions and well-being in the world, 1950-1990*. Cambridge University Press.

Puddington, A., Dunham, J., Aghekyan, E., Brandt, C., Bhatia, R., Linzer, I., O'Toole, S., Repucci, S., and Roylance, T. (2018). Freedom in the world.

Schumpeter, J. A. (1942). Capitalism, socialism, and democracy. *New York*.

Skaaning, S.-E., Gerring, J., and Bartusevičius, H. (2015). A lexical index of electoral democracy. *Comparative Political Studies*, 48(12):1491–1525.

Svolik, M. (2008). Authoritarian reversals and democratic consolidation. *American Political Science Review*, 102(2):153–168.