Ranking the World

Grading States as a Tool of Global Governance

Edited by ALEXANDER COOLEY Barnard College

and

JACK SNYDER Columbia University



Seva Gunitsky. 2015. "Lost in The Gray Zone: Competing Measures of Democracy in the Former Soviet Republics." In *Ranking the World: Grading States as a Tool of Global Governance*, p.112-50, edited by Alexander Cooley and Jack Snyder. New York: Cambridge University Press.

ABSTRACT

I examine some fundamental problems with measures of democracy in the former Soviet republics. I find that indices often disagree about particular countries, and occasionally draw contradictory conclusions from observing the same event. Measures of hybrid regimes are particularly unreliable, and regional comparisons of democratic quality are also highly sensitive to measure choice. These differences, I argue, reflect inherent tradeoffs in conceptualizing democratic governance. That is, they arise not from incorrect specifications but from fundamental normative disagreements about a highly contested concept.

113

Lost in the gray zone

Competing measures of democracy in the former Soviet republics

SEVA GUNITSKY

Introduction

5

Recent decades have seen the spread of regimes that claim the mantle of democracy even as their true democratic status remains uncertain.¹ These hybrid regimes reside in the gray zone of democracy indices. They resist easy comparisons, both to each other and across different measures – which has made reliable indicators of their democratic quality all the more important for both academics and policymakers.

The story of two post-Soviet states can illustrate the difficulty. One country saw the rise of crony capitalism, growing restrictions on media freedom, and the increasing centralization of power by the president. The other cut down on corruption, established strong partnerships with the West, and developed a parliament that successfully stood up to the president's unpopular choice of a prime minister. By the turn of the century, as Figure 5.1 shows, their trajectories had irrevocably diverged.

At first glance, these cases seem to be just another example of the diversity in post-Soviet institutional development. The only problem is that the two are in fact *the same country*. What the above figure actually shows is the assessment of Russian democracy by Freedom House (dashed line), compared to a rather more optimistic assessment by Polity IV (solid line).² Both are commonly used indices that claim to measure the same phenomenon. So how did they come away with such vastly different portrayals of Russian democracy? In both cases, plausible narratives could be constructed to fit the statistical outcome, but



Figure 5.1 Divergent political trajectories in the post-Soviet space

which one is closer to reality? And what are the implications of such broad disagreement for how scholars examine the causes of democratization, and for how policymakers evaluate the effectiveness of democracy assistance?

This chapter examines the sources of these disagreements by looking at how various measures assess democracy in the former Soviet republics. I begin with some general drawbacks of commonly used indices, and the implications of these drawbacks for public policy and academic research. I find that indices often disagree about particular countries, and occasionally draw contradictory conclusions from observing the same event. Measures of hybrid regimes are particularly unreliable, and regional comparisons of democratic quality are also highly sensitive to measure choice.

The methodological weaknesses of popular measures have been extensively discussed elsewhere,³ and it is not my goal to rehash these criticisms here. Instead I focus on problems of *definition* as the root source of measure divergence. Disagreements such as the example above, I argue, reflect inherent trade-offs in conceptualizing democratic governance. Though at least some of the measures aspire to objectivity,

¹ Replication data and supplementary materials for this chapter are available at individual.utoronto.ca/seva/measures.html.

² Freedom House measures are reversed so that higher scores mean better democratic quality for both indices.

³ See, for example, Munck 2009 or Munck and Verkuilen 2002.

Seva Gunitsky

Table 5.1. Competing definitions of "democracy" in common indices

Measure	Broad conception of democracy
Polity IV	Divided patterns of elite authority
Przeworski et al.	Electoral turnover
Vanhanen	Party fragmentation and high voter turnout
Freedom House	Individual freedom and equality
Treedom Trouse	individual needoni and equanty

they ultimately represent rival notions of the concept of democracy. For example, the Polity IV measure takes democracy to mean "divided patterns of authority," the Przeworski *et al.* measure takes it to mean "electoral turnover," the Freedom House measure takes the view of democracy as "freedom," while the Vanhanen measure takes it to mean "party fragmentation and high voter turnout."⁴ To say that one measure is more accurate than another misses the point, since they represent overlapping but non-identical conceptions of a highly complex phenomenon (see Table 5.1).

At the same time, these disputes are not simply a case of reasonable people disagreeing about a complicated concept. Ideological or political motivations can also shape definitional preferences. Democracy, after all, is undoubtedly a politically loaded term: "It is almost universally felt that when we call a country democratic we are praising it," wrote George Orwell in 1946. Thus "the defenders of every kind of regime claim that it is a democracy, and fear that they might have to stop using that word if it were tied down to any one meaning."⁵ During the Cold War, for example, both the Soviet Union and the United States proclaimed themselves to be democratic societies but emphasized different conceptions of democracy – individual liberty in the case of the United States and social justice in the case of the USSR.⁶

⁴ These measures are discussed in more detail below.

⁵ Orwell 1946: 132–133.

⁶ These conceptions were partly self-serving and partly hypocritical for both states, but also had a profound effect on Cold War politics. As Capoccia (2005: 241) argues, the protection of human rights as enshrined in a number of international treaties had been impossible during the Cold War precisely because the two superpowers could not agree on a common definition of democracy. "However, after the collapse of international Communism," he writes, "a sufficiently general agreement on at least a minimal 'procedural' conception of democracy has become more possible, and international documents providing for democratic rights have suddenly become enforceable, at least in principle."

More recently, Freedom House has been criticized for espousing a pro-American (and anti-Russian) ideology in its rankings. This charge is particularly problematic because Freedom House acts as both an expert on measuring democracy and an advocate for US-led democracy promotion.

The consequences of definitional disputes extend far beyond scholars excavating through footnotes in regression analyses. As the case of Freedom House demonstrates, democracy measures can shape advocacy efforts by non-governmental organizations, and influence popular perceptions of global and regional trends in democratization. Measurements can also have a direct impact on government policies; USAID, for example, chooses its assistance projects based in part on assessments of where they can be most effective, which includes measuring democratic development and judging the potential for further progress. Good measures are particularly important for evaluating mixed regimes, where foreign assistance can make the biggest difference for democratization, and where policymakers need accurate assessments of the impact of aid. As I show below, democracy indices are particularly unreliable for judging mixed regimes, creating the potential for misguided policies and advocacy efforts.

After examining general problems of defining democracy, the second part of the chapter turns to specific measurement problems as they apply to the former Soviet republics. I find that redundant sub-scores can underestimate the level of democracy and exaggerate regime variation across the region. In addition, current measures don't appear to be particularly sensitive to the quality of elections. Fortunately, these problems stem from poor operationalization and can be treated with the careful application of statistics.

More serious problems, however, appear to be intractable. I find that indices often disagree about particular countries, and occasionally draw distinct or even contradictory conclusions (and ratings) from observing the same event. I find that measures of hybrid regimes are particularly capricious. Regional comparisons of democratic quality (in this case, former Soviet republics compared to Eastern Europe) are also highly sensitive to the choice of measure. These problems appear to stem from the way the creators of the measures define and conceptualize democracy.

Since the choice of definition shapes both the choice of variables and how these variables are aggregated, the lack of consensus about measurement is both fundamental and intractable. This conclusion has two implications: first, the inherent limitations of any measure should be made explicit when evaluating democracy's relationship with other often-studied national attributes like economic development or propensity for war. Second, since different indices capture different elements of democracy, the choice of measure should be shaped by whichever element is most salient for the specific research puzzle or policy proposal at hand.⁷

Evaluating indices through a regional lens can serve two purposes. First, it provides a guide for area specialists who wish to know how well a particular region is served by these measures. Bogaards (2007), for example, argues that indices focusing on election outcomes perform poorly at measuring the quality of democracy in Africa.⁸ As a result, measures that rely on election outcomes may be particularly inappropriate for judging democratic quality in that region. Second, focusing on a region can expose the hidden assumptions and potential pitfalls of the measures themselves. For Bogaards, looking at the measures using African cases reveals not only which approaches are more suitable for the region, but also some general limitations in how they measure electoral contestation around the world. As I aim to show in this chapter, the former Soviet republics can serve the same useful purpose.

Evaluating statistical measures of democracy

The concept of democracy lacks a consensus definition beyond a core notion of "self-governance."⁹ Some approaches adopt a minimalist,

- ⁸ He concludes that these measures make strong assumptions about the relationship between regime type and the size of electoral victory, "are frequently in disagreement, correspond poorly with Freedom House and Polity classifications and scores, evidently misclassify numerous regimes, and fail to reflect continuity and change" (Bogaards 2007: 1233).
- ⁹ The literature on defining democracy and classifying political regimes is vast. See, for example, Schmitter and Karl 1991; Lawson 1993; O'Donnell 1993; Collier and Levitsky 1997; Mainwaring, Brinks, and Perez-Linan 2001; Altman and Perez-Linan 2002; Reich 2002; Daly 2003; Bennett 2006; and Whitehead 2011.

Lost in the gray zone

procedural view while others aggregate a variety of attributes.¹⁰ The Freedom House index, perhaps the most widely known measure,¹¹ takes a very broad view of the concept.¹² Its civil liberties score, for example, includes measures of media self-censorship, industrial production quotas, government control over sermons, nepotism in university admissions, and socio-economic inequality, among many others (Freedom House 2010). It is thus open to the charge of conflating processes with outcomes. If democracy is measured by positive outcomes like lack of corruption or economic equality, it becomes impossible to measure democracy's causal impact on these outcomes. Countries that promote economic equality become democratic by fiat, and the impact of regime type on economic equality becomes not a subject of study but a tautology. At the same time, economic equality may be necessary for equal political representation if personal wealth determines the extent of political influence. Depending on the electoral rules (such as limits on campaign contributions) and the extent of inequality, socio-economic equality could be seen as either a prerequisite or an outcome of democratic government. A persuasive case can be made for either position, each with its own merits and drawbacks.

A minimalist conception of democracy faces a different set of problems. It may exclude what some might consider crucial elements of democracy. Polity IV, for example, focuses on procedural aspects of democracy at the elite level but ignores mass participation.¹³ Thus the

- ¹⁰ This discussion excludes a number of less common indices, most of which are inapplicable to the former Soviet republics because they either do not cover the region or do not extend beyond 1991. For an overview of other indices, see Munck and Verkuilen 2002.
- ¹¹ Relative popularity can be (unscientifically) gauged by search results in Google Scholar: between 2005 and 2010, Freedom House has approximately 6,700 mentions in academic books and articles, Polity IV has 2,400 mentions, and Vanhanen comes in with approximately a thousand (restricted to searches in social sciences, arts, and humanities).
- ¹² Freedom House has rated every country in the world since 1972 on a scale of 1 (most democratic) to 7 (least democratic). The measure consists of two subcomponents, Political Rights and Civil Liberties, which are themselves aggregates of ten and thirteen subcomponents, respectively, and capture a wide variety of national attributes related to personal freedom. See http://www.freedom house.org for country scores and detailed descriptions of the methodology.
- ¹³ Polity IV is a measure commonly used in the academic literature. It rates most countries in the world (excluding some small states) since 1800, using a scale of -10 to 10, with the latter being the most democratic. Polity measures democracy as a weighted sum of five components the competitiveness and

⁷ A recent step in this direction has been the effort by Coppedge and Gerring (2011) to produce a data set that disaggregates the concept of democracy along its various dimensions. They argue for six separate conceptions of democracy: electoral, liberal, majoritarian, participatory, deliberative, and egalitarian. Many of the conceptual trade-offs involved in emphasizing a particular interpretation of democracy are discussed below.

Seva Gunitsky

United States rates as a perfect democracy after 1871 despite the subsequent expansion of suffrage to women, African-Americans, and citizens over eighteen years of age. If mass participation is a key component of democracy, then Polity IV does not adequately capture the concept.

The problem of definition is further compounded by some inherent trade-offs in democratic governance. Not all aspects of democracy are positively correlated. Majority rule, for example, frequently comes into conflict with individual liberties. If democracy is defined as the will of the majority, then the Bill of Rights and the Supreme Court are inherently undemocratic institutions. This tension between mass participation and individual liberty carries through to modern measures of democracy. Thus, the index of democracy produced by the Economist Intelligence Unit (EIU) regards mandatory voting as decreasing democratic quality because it infringes upon individual liberty.¹⁴ At the same time, mandatory voting clearly increases voter turnout, particularly by the poor, who have few alternative venues for exerting political influence and exhibit lower rates of participation when voting is optional. Mandatory voting therefore improves the quality of mass participation at the expense of individual liberty. The index of democracy produced by Tatu Vanhanen, for example, is based in part on measuring the percentage of the national population that votes in elections.¹⁵ Countries with mandatory voting thus receive a higher score on the Vanhanen index and a lower score on the EIU index. In this case, mandatory voting is punished or rewarded based

openness of executive recruitment, the regulation and competitiveness of participation, and constraints on the chief executive (with the latter weighed most heavily). See Marshall, Gurr, and Jaggers 2010.

- ¹⁴ The Economist Intelligence Unit (EIU) index of democracy has rated most countries in the world (165 states) every two years from 2006 to 2010, and annually since then. The latest index, as of this writing, was released in 2012; this chapter examines the measure up to 2010, although the methodology has remained the same in the latest iterations. The EIU score, which ranges from 0 to 10, is based on sixty indicators combined into five sub-components – free and fair elections, civil liberties, functioning of government, political participation, and political culture. See Kekic 2007.
- ¹⁵ Vanhanen uses a measure that rates most countries in the world between 1810 and 2004. His score is a product of two sub-components: the percentage of votes not cast for the largest party, and the percentage of the population that voted in elections. The two scores are multiplied and divided by a hundred, yielding a scale that could theoretically range between 0 and 100 (although in practice the highest score is 49, achieved by Italy in 1993). See Vanhanen 2000 and 2003.

on whether the measure emphasizes mass participation or individual liberty. The divergence between the two measures reflects a fundamental disagreement about the essential nature of democracy – a normative choice rather than a methodological one.

In addition to these inherent tensions, the way scholars and policymakers view the "appropriate" definition of democracy may be shaped by the historical context of the times and even unrecognized ideological biases. As Ido Oren (2002) has argued, over the twentieth century the definition of democracy by American political scientists has been shaped by the context of America's international rivalries. Conflict with other major powers has caused political scientists to view them as less democratic. Before 1914, for example, prominent contemporaries like John Burgess and Woodrow Wilson portrayed Germany as a progressive state with a liberal constitution and an incorruptible bureaucracy. But conflict with Germany in two world wars caused American scholars to re-label the country as a militaristic and semi-feudal monarchy.¹⁶ The fact that Polity scores for pre-World War I Germany are significantly lower than those of England, France, and the US reflects this ideological shift, according to Oren. The Polity measures illustrate "the role played by data sets in objectifying concepts that were originally rooted in a particular interpretation of political development," he argues, "an interpretation that happens to associate the trappings of democracy with America's enemies and the substance of democracy with America's allies."¹⁷

Freedom House has likewise attracted its share of critics for allegedly espousing a pro-American bias (like Polity, Freedom House has assigned the US a perfect score ever since its inception in 1972). The majority of its funding comes from the US government (its 2007 financial statement puts the share at 66 percent; *The Economist* estimates the average figure at 80 percent).¹⁸ It has formally stated that its diverse

 ¹⁶ For example, Bismarck's introduction of unemployment insurance, old age pensions, and national healthcare in the 1880s set the precedent for the modern European welfare state.

¹⁷ Oren 2002: 25. As Martin Walker notes in his history of the Cold War, American descriptions of the Soviet regime often hinged on its status as an ally. When *Collier's* magazine ran a special issue in December 1943 titled "What Kind of Country is Russia Anyway?", they concluded that it was neither socialist nor communist but "a modified capitalist set-up [moving] toward something resembling our own and Great Britain's democracy." Quoted in Walker (1994: 30).

¹⁸ Freedom House 2008; "Measuring liberty: When freedom stumbles," The Economist January 19, 2008.

trustees are "united in the view that American leadership in international affairs is essential to the cause of human rights and freedom."¹⁹ Freedom House has strongly denied charges of partisanship, pointing to its willingness to criticize America's strategic partners like Egypt, Pakistan, and Saudi Arabia.

Beyond ideology, the political zeitgeist can also shape what scholars and policymakers consider to be "essential" elements of democracy. Athens was considered a democracy by contemporaries despite widespread slavery and the exclusion of women and metics from political participation. More recently, foreign visitors to the United States in the early nineteenth century called it a democracy despite slavery and the political exclusion of women. But by the 1970s, notes Markoff, "although South Africa had multiparty competition, very few would have called that country a democracy because only its white minority could vote. A great change in the meaning of democracy had occurred."20 Nor do such shifts require a long time span: attempting to define democracy in 1944, Raymond Aron concluded that "the idea of popular sovereignty is not essential: It can lead as easily to despotism as to liberty. And after all, to a large extent it has been popular majorities that have abused their power."21 Writing scarcely a decade after Hitler's democratic ascent to power, Aron could not conceive of mass participation as a key component of democracy. Today, however, elections are considered a crucial part of democratic governance. Their importance is revealed in the significance assigned to elections as democratic milestones in countries like Iraq or during the Arab Spring, as well as in the efforts devoted to international election monitoring. To take another example, during the Great Depression prominent political scientists defined democracy in economic and substantive terms as much as in the political and procedural terms prevalent today. Thus Merle Fainsod argued in 1934 that the eradication of poverty and unemployment was as significant to democracy as universal suffrage, while Charles Merriam saw economic equality as a key measure of democracy. It was only after World War II that procedural, Schumpeterian conceptions replaced substantive definitions.²²

Conceptions of democracy vary across regions and cultural contexts as well as over time. Arab Barometer surveys from the summer of 2011, Lost in the gray zone

for example, show that 79 percent of Egyptians see democracy as the best form of government, but they perceive it in terms of social justice and economic equality. Sixty-five percent of the survey respondents listed low levels of inequality or government provision of basic services for all citizens as the most essential feature of democracy. Only 6 percent defined its most essential feature as the ability to change governments through elections, and only 4 percent defined it as the right to criticize the government.²³ This definition is much more consistent with substantive rather than procedural conceptions of democracy.

In sum, measures of democracy face inherent trade-offs in emphasizing different aspects of democratic rule, and choosing among these trade-offs can be shaped by both ideological and methodological preferences. Whatever the predispositions of their makers, definitions vary because the meaning of democracy encompasses competing normative elements. The choice of measurement, to which I now turn, is shaped by these definitional considerations.

Problems of measurement

Growing self-consciousness about measure choice has produced a number of debates about transforming vague concepts into precise measures. For example, should democracy be considered a dichotomous or a continuous variable? Przeworski, Alvarez, Cheibub, and Limongi ("PACL" in this chapter) offer a spirited defense of a dichotomous measure, going so far as to call intermediate measures "ludicrous."²⁴ They argue that democracy is a type rather than a degree, a notion perhaps best summarized in the words of writer

 ¹⁹ Freedom House, www.freedomhouse.org, 'About Us' (accessed 2011).
 ²⁰ Markoff 1996: 60.
 ²¹ Aron 1944: 175.
 ²² Oren 2002: 12.

²³ Michael Robbins and Mark Tessler, "What Egyptians mean by democracy," Foreign Policy Middle East Channel, September 20, 2011. Available at: http:// atfp.co/1qkpoAH.

²⁴ Alvarez *et al.* 1996: 21. Their measure tracks most countries in the world (maximum of 149 states) from 1950 to 2008 using a dichotomous measure. Przeworski et al. (2000) originally scored countries for the years 1950-1990; Cheibub et al. (2010) extended the measures to 2008. A country is considered a democracy if it fulfills all four of the following criteria: the chief executive and the legislature must be elected, more than one party must compete in the elections, and an alternation of power under identical election rules must take place. If any of those components are missing, the country is rated a dictatorship. See also Przeworski et al. 1996 and Przeworski and Limongi 1997: 178-179. Samuel Huntington (1991: 11-12) and Juan Linz (1975: 184-185) likewise adopt dichotomous measures.

Seva Gunitsky

Amiri Baraka: "A man is either free or not. There cannot be any apprenticeship for freedom."25 They also claim that dichotomous measures contain less measurement error because they are more transparent and easier to replicate. Thus dichotomous measures are, as Elkins (2000) puts it, theoretically more valid and more reliable.²⁶ But in testing these claims, he finds that graded measures have "superior validity and reliability" and that dichotomous measures seem insensitive "to the incremental, and sometimes partial, process that characterizes many democratic transitions."²⁷ Indeed, although PACL aims to be strict in its criteria, the imprecision of the measure means that countries like Venezuela and Sierra Leone are judged to be as democratic as (say) Norway.

Ultimately, as Collier and Adcock (1999) argue, the choice between binary and continuous measures should depend on the specific goals of the research. Donnelly (2006: 88) suggests a three-point checklist for when a dichotomous variable may be preferable to a continuous one: when the dividing line is sharp and clear, when the gray area between the two cases is small, and when few important cases fall into that gray area. By these criteria, a continuous measure is preferable for most kinds of research.²⁸ Because binary measures are very sensitive to where one makes the cut, they would not work well in examining regimes in the middle of the spectrum, such as competitive autocracies. At the same time, dichotomous measures may be more appropriate for studying major political transformations or tracking the duration of democratic regimes.

Beyond the choice of scale sensitivity, commonly used measures suffer from a number of coding problems.²⁹ For example, Coppedge

²⁶ The claim of reliability is methodological while the claim of validity is theoretical: Elkins is thus incorrect in criticizing the proponents of the dichotomous approach for neglecting "advances in data collection and analysis" (2000: 293) that allow more precise measures of regime change. Their insistence on binary measures stems not from the paucity of data but the theoretical conception of democracy as a non-continuous concept.

- ²⁸ Epstein et al. (2006) adopt a trichotomous measure of democracy on similar grounds, and Kellstedt and Whitten (2009: 96) likewise argue that continuous variables are more appropriate for measuring democratization.
- For an overview, see Munck and Verkuilen 2002 or Munck 2009. Earlier writing on problems of democracy measurement includes Inkeles 1991 and Beetham 1994.

and Gerring (2011: 250) point out that when human judgments are required for coding, indicators with subjective criteria may produce different results from one expert to another. Bollen and Paxton argue that this subjectivity pervades democracy indices: examining measures between 1972 and 1988, they find "unambiguous evidence of judgespecific measurement errors."³⁰ It's not surprising, therefore, that regional experts frequently see a disconnect between the measures and their understanding of the region. For example, Bowman, Lehoucg, and Mahoney (2005) find that large-n measures of Central American states are plagued by errors stemming from inaccurate, partial, or misleading data sources. Comparing scores of some African states with qualitative observations, McHenry (2000) finds discrepancies both within and across data sets: measures do not reflect the observed facts on the ground, and measures of the same state do not reflect the same phenomena.

These problems extend to specific measures. Vanhanen, for example, measures turnout as a percentage of the national population, in effect punishing countries that have a relatively large number of people below voting age. Since fertility rates correlate negatively with income, this has the effect of penalizing poorer states and introducing a systemic measurement bias. His measure of competition also favors highly fragmented legislatures, and may thus be capturing the nature of electoral systems rather than democratic quality.³¹ Other measures fare no better: Gleditsch and Ward argue that the recruitment and participation dimensions of Polity IV are "empirically extraneous despite their centrality in political theory."32 They suggest that Polity ought to be treated as categorical rather than continuous: it clusters at the +10 category, for instance, suggesting insensitivity to gradations at higher levels of democracy (see Figure 5.2). Treier and Jackman demonstrate "considerable error in the latent levels of democracy underlying the Polity scores."³³ And Munck points out that while the methods used to compute the Freedom House index have changed over time, scores from previous years have not been adjusted to reflect the new methodology. Freedom House scores thus suffer from the additional problem of internal inconsistency, so that the same score reflects different judgments of democratic quality in different years.³⁴

- ³⁰ Bollen and Paxton 2000: 58.
 ³¹ Berg-Schlosser 2004: 252.
 ³² Gleditsch and Ward 1997: 361.
 ³³ Treier and Jackman 2008: 213.
- 34 Munck 2009: 10. Polity IV also appears to re-evaluate its criteria, but recodes the scores retroactively. See footnote 57.

²⁵ Baraka 1962.

²⁷ Elkins 2000: 293.



Figure 5.2 The uneven distribution of Polity scores suggests the measure may be categorical rather than continuous

The fact that many indices are highly correlated may suggest that they are measuring roughly the same phenomenon. Polity IV and Freedom House, for example, have a Pearson's R of 0.88. But as Casper and Tufis (2003) point out, highly correlated measures can produce very different regression results. Using three measures of democracy (Polyarchy, Freedom House, and Polity IV), they show that regressions of these highly correlated measures on variables like education and economic development often produce inconsistent results.³⁵ Moreover, the correlation among measures is highest for clear-cut cases. The correlation between Polity and Freedom House drops to 0.64 when democracies are excluded from the analysis.³⁶ And when both clear-cut autocracies and democracies are excluded, the correlation drops further to 0.50.³⁷ This is especially problematic Lost in the gray zone

since democracy promotion and foreign aid is often directed at countries in the middle of the range. In these cases, the substantive results of any analysis may be greatly affected by the choice of measure. As I discuss below, this uncertainty remains a problem when comparing measures of democracy in the former Soviet republics.

Problems of aggregation

Aggregation involves combining sub-scores to produce a single measure of democratic quality. It faces two common problems: the choice of sub-scores (which is related to the problem of definition), and how to combine these sub-scores into a single measure. Aggregation can conceal differences among sub-components; Armstrong argues, for example, that the disaggregation of Freedom House scores reveals that many countries receiving the same overall score are "interestingly different from each other" while some countries with different overall scores appear fundamentally the same.³⁸ As Coppedge and Gerring point out, multicollinearity among sub-components suggests that measures are not independent of each other.³⁹ They note that the seven subcomponents used to create measures of political rights in the Freedom House index have a Pearson's R of 0.86 or higher. While it's possible that all these components move together, their high correlation and ambiguous coding procedures suggest that "country coders have a general idea of how democratic each country is, and ... this idea is reflected in consistent scores across the multiple indicators." As I show below, high sub-score correlations can push the measure distribution toward the tails, making relatively high-scoring countries appear even more democratic and relatively low-scoring countries even more autocratic than they are. The choice of appropriate sub-components is therefore a consequential one.

The choice of the aggregation procedure itself is also crucial, since different methods contain distinct assumptions and produce divergent measures.⁴⁰ As a simple example, consider the aggregation of two subscores that measure two aspects of democratization. Using factor analysis, Coppedge, Alvarez, and Maldonado have created two such

³⁸ Armstrong 2011: 653. ³⁹ Coppedge and Gerring 2011: 251.

⁴⁰ Sanin, Buitrago, and Gonzales (2013), for example, argue that the use of weighted averages is based on "untenable" assumptions like strict monotonicity.

³⁵ On this point see also Munck and Verkuilen 2002: 29.

³⁶ Freedom House defines a country as "free" if their score is 2.5 or lower; Polity defines coherent democracies as those that score 7 or higher.

³⁷ Coppedge and Gerring (2011: 252) note similar trends in correlations of Polity IV and the political rights component of Freedom House. For detailed correlation tests across regime types, see Hadenius and Teorell 2005.



Figure 5.3 Two methods of aggregating contestation and inclusiveness into a single measure (multiplication or averaging) produce dramatically different results.

sub-scores based on what they find to be "two persistent dimensions of democracy" – democratic contestation and inclusiveness.⁴¹ In order to derive a single measure and compare their results to other indices, we can aggregate the two scores in two ways – by averaging them (which, in terms of measure distribution, is equivalent to adding them) or by multiplying them together. Indices have employed both methods – while Freedom House averages scores and Polity uses a weighted addition scheme, Vanhanen and PACL use multiplicative aggregation.⁴²

Figure 5.3 compares the measure distributions that result from two different methods of aggregation of their sub-scores. The chart on the left shows the measure distribution when the sub-scores are multiplied, while the chart on the right shows the measure distribution when the same two sub-scores are averaged. The two methods produce clearly different outcomes – the world of sub-component averaging is far more democratic than the world of sub-component multiplication.⁴³

Despite the dramatic effects of aggregation choice, Munck points out that "with a few notable exceptions, existing democracy indices have displayed a fairly low level of sophistication concerning the process of

- ⁴¹ Coppedge, Alvarez, and Maldonado 2008. More precisely, each is a principal component factor index of a number of indicators associated with contestation and participation.
- ⁴² PACL is generally not considered to be an aggregated measure, but as I discuss shortly, this is incorrect.
- ⁴³ Coppedge *et al.* (2008: 633) state that "the truth of one [dimension] does not imply the truth of the other" and argue that the two dimensions are independent of each other.

Lost in the gray zone

aggregation."44 He and Verkuilen argue that the aggregation method must be explicitly justified and linked to the measure's theoretical conception of democracy rather than constructed ad hoc.⁴⁵ The weighing of sub-components, for example, reflects choices about which elements of democracy are more important, which is tied to how democracy is defined in the first place. Moreover, the aggregation procedure contains certain assumptions about the nature of democracy. PACL, for example, examines the presence or absence of four factors: the election of an executive and of a legislature, party competition, and electoral turnover. If any of those components are missing, the country is rated a dictatorship. In effect, this scheme uses a multiplicative aggregation method with four dichotomous variables. It follows the logic of individually necessary and jointly sufficient conditions - each component is required for a country to be considered a democracy, and the presence of all four is enough to label a country democratic. Vanhanen also uses a multiplicative aggregation method, deriving his measure from the product of two scores: the percentage of votes not cast for the larger party, and the percentage of the population that voted in the elections. Thus a high turnout in which no other parties competed for office (as in the case of the Soviet Union, for instance) would result in a score of zero. Generally, the use of multiplicative aggregation with sub-components that could in theory equal zero imply strict criteria for inclusion - if any of the elements is zero, the entire score collapses to zero as well.⁴⁶ The same is true *a fortiori* if the subcomponents are dichotomous, as in the case of PACL. In effect, multiplicative aggregation in which sub-scores can equal zero takes the view that no country can be considered a democracy if it lacks particular elements or institutions, regardless of other factors. That is, no single factor is enough to label a country democratic.

In short, because the aggregation method contains inherent theoretical assumptions, it should be explicitly linked to the particular conceptualization of democracy. As Goertz (2006) points out, the disconnect between theory and aggregation is a common problem for democracy measures: while most theories of democracy imply necessary and sufficient conditions, most measures of these theories (such as Polity IV) use the "family resemblance" strategy typified by addition or averaging.

⁴⁴ Munck 2009: 35. ⁴⁵ Munck and Verkuilen 2002: 22–27.

⁴⁶ In Vanhanen's measure, roughly 35 percent of all cases receive a score of zero.

Evaluating democracy in the former Soviet republics

This section examines how specific measurement problems can distort the evaluation of democracy in the former Soviet republics. I focus on five mini "case studies" that examine redundant sub-scores, the relationship between election outcomes and election quality, measurements of hybrid regimes, regional comparisons, and major disagreements among measures (using Russia and Armenia as two examples). In each instance, analysis reveals potentially serious shortcomings in how various measures evaluate democracy in the region. While some of these problems can be minimized with the careful use of statistics, most are fundamental to the measures themselves.

Sub-score choice and the EIU index

The choice of sub-scores reflects how the measure conceptualizes distinct elements of democracy. As mentioned above, high correlation among subscores (as in the case of Freedom House, for example) suggests that these measures are not independent of each other. The EIU index, for example, includes sub-scores that measure the government's ability to function (i.e. the rule of law) and the extent of civil liberties, but also sub-scores that measure the extent of free elections and public participation. It could be argued, however, that free and fair elections and high levels of public participation *imply* and in fact *require* rule of law and civil liberties, so that the latter two measures are redundant, and including all four can produce multicollinearity in regression models. In fact, government functioning is highly correlated with both free elections (r = 0.85) and participation (0.84). Likewise, the measure of civil liberties is also highly correlated with free elections (0.93) and participation (0.89).

The statistical effect of these high correlations is to increase measure variance by pulling the distribution of scores toward the tails – the level of democracy in relatively democratic countries will be exaggerated, as will the level of autocracy in relatively autocratic countries. Is this effect indeed present in the EIU scores? Figure 5.4 compares the original EIU score with an adjusted score in which the sub-components of government functioning and civil liberties have been removed. The results show that the difference between the modified and the original score is indeed negatively correlated with the level of democracy. In the revised measure, autocratic countries like Uzbekistan, Turkmenistan, and Tajikistan



Figure 5.4 Adjusting the EIU score for redundant sub-components. The y-axis represents the effect of the adjustment on a country's score while the x-axis represents the level of democracy in the original score (countries are more democratic as one moves toward the right end of the spectrum).

receive a higher score while countries toward the more democratic end of the spectrum like Latvia and Estonia receive a lower score.⁴⁷

Adjustment has the overall effect of increasing democracy scores in autocratic countries and decreasing democracy scores in democratic countries. The effect on hybrid regimes is mixed. These results suggest that the EIU measure appears to exaggerate regime type at the tails of the distribution, particularly for low-ranked countries (after adjustment, the regional average increases from 4.6 to 4.75). Figure 5.5 compares the distribution of the original and the modified scores for the region. For modified scores, the low end of the spectrum has been "pulled in" (this does not affect relative country rankings, since the omitted measures are highly correlated with the included measures).

⁴⁷ The standard deviation of the scores decreases from 1.99 to 1.82, as expected. Note also the distinct clustering of measurement effects for three categories of countries – the obvious autocracies (Uzbekistan, Turkmenistan, Tajikistan), the hybrid regimes (Azerbaijan, Belarus, Kazakhstan, Kyrgyzstan, Russia, Georgia, Armenia) and the regional democracies (Moldova, Ukraine, Latvia, Lithuania, Estonia).



Figure 5.5 Distribution of original (a) and modified (b) EIU scores

In short, the presence of redundant sub-scores in the EIU index introduces two potential errors. First, it exaggerates the range of regime types in the region as a whole. Second, it inflates measures at the tails of the distribution, particularly among autocracies, making them appear even worse than they actually are; the overall effect is to underestimate the regional level of democracy. These problems demonstrate that the choice of sub-components (and their aggregation) requires careful theoretical justification.

Measuring election quality

Traditional measures of electoral participation such as voter turnout may be unreliable indicators of democratic quality, because they ignore differences in the quality of elections. As the playwright Tom Stoppard put it, "it's not the voting that's democracy, it's the counting." This problem is especially acute for competitive autocracies in which incumbents allow popular elections but "routinely abuse state resources, deny the opposition adequate media coverage, harass opposition candidates and their supporters, and in some cases manipulate electoral results."⁴⁸ As mentioned in the introduction, Bogaards (2007) has criticized democracy measures in Africa for ignoring election quality, showing that election outcomes in the region are not consistently related to levels of democracy.

I use Judith Kelley's (2012) data set of election quality to see if existing measures of democracy adequately capture the quality of

⁴⁸ Levitsky and Way 2002: 53. See also Levitsky and Way 2010.



Figure 5.6 Relationship between election quality and overall democratic quality (taken as the average of other measures rescaled from zero to one). Each data point represents an election; selected outliers are labeled with country and year information. The two measures have an overall correlation coefficient of 0.57.

elections in the region. In general, the relationship between election quality and overall democratic quality (measured as the average of other indices) is present though not particularly strong, with a correlation coefficient of 0.57 (see Figure 5.6).

Individual measures do not fare particularly well when compared with a measure of election quality. The correlations between election quality and other measures of democracy are lower than correlations among the measures themselves, and election quality has the lowest overall correlation with other measures (see Table 5.2).

The Vanhanen measure has the lowest correlation with election quality (0.58). This is especially problematic since this measure takes care to incorporate election outcomes (in the form of turnout) into its score.

Comparing election turnout with election quality reveals no relationship between the two, as shown in Figure 5.7. Turnout is in fact a very poor predictor of election quality. Overall, the results are mixed. There is indeed some relationship between existing measures of democracy

Table 5.2. Correlations among measures of democracy and election quality in the former Soviet republics (n = 40)

	FH	POL	VAN	UDS	CAM	OVERALL
Polity IV (POL)	0.83					0.80
Vanhanen (VAN)	0.73	0.71				0.75
Universal	0.94	0.90	0.86			0.86
Democracy Score (UDS) ^a				¢		
Coppedge et al.	0.91	0.89	0.89	0.93		0.87
(CAM) Election Quality	0.72	0.68	0.58	0.68	0.72	0.68

^aUDS, or Universal Democracy Score, is a measure developed by Pemstein, Meserve, and Melton 2010. It is synthesized using a Bayesian latent variable approach from ten other measures. According to the authors, this approach allows the simultaneous leveraging of a number of indices.



Figure 5.7 The relationship between election quality (from Kelley 2012) and percentage turnout (from Vanhanen 2000). Each dot represents a national election.

Lost in the gray zone

and election quality (it would be surprising if there was not), but measures of democratic participation appear insufficiently sensitive to election quality. Research or policy that focuses on the quality of elections should keep this limitation of current measures in mind.

Measure disagreement and regime type

How do the different measures compare in evaluating particular countries? Do certain indices consistently rank countries higher than others? And if there are disagreements among indices, are there any patterns to that disagreement? In line with previous findings, the overall correlations among different measures for the former Soviet republics are relatively high (although the Vanhanen and PACL measures have consistently lower correlation rates; see Table 5.3 below).

As mentioned above, high correlations do not necessarily mean the measures are interchangeable in regression analysis, and could conceal disagreement about hybrid regimes. To gauge their consistency, we need to compare the measures directly. To do so, I rescaled the scores of seven indices – Freedom House; Polity IV; EIU; Coppedge, Alvarez, and Maldonado 2008 (henceforth CAM); Vanhanen; PACL; and UDS – on a scale of zero to one, with higher numbers representing greater levels of democracy.⁴⁹ The summary statistics for the modified measures are presented in Table 5.4.

Table 5.3. Correlations among measures of democracy in the former
Soviet republics

	POL	FH	UDS	VAN	CAM
Freedom House (FH)	0.86				
Universal Democracy Score (UDS)	0.93	0.94			
Vanhanen (VAN)	0.79	0.77	0.84		
Coppedge <i>et al</i> . (CAM)	0.92	0.95	0.95	0.87	
Przeworski <i>et al.</i> (PACL)	0.75	0.73	0.77	0.65	0.73

⁴⁹ This required rescaling the indices so that all measures were non-negative, then dividing them by the highest possible global measure for that index. The latter operation provides a sense of where each index places the country compared to the rest of the world (the global maxima for each rescaled measure are as

Lost in the gray zone

Table 5.4. Summary statistics for the rescaled measures of democracy inthe former Soviet republics

Measure	Ν	Years	Minimum	Maximum	Mean
Freedom House	300	1991–2010	0	0.86	0.38
Polity IV	300	1991–2010	0.05	1	0.55
EIU	45	2006, 2008,	0.17	0.78	0.47
		2010	,		
UDS	270	1991–2008	0.15	0.83	0.49
CAM	149	1991–2000	0.13	0.78	0.44
Vanhanen	209	1991–2004	0	0.86	0.30
PACL	270	1991–2008	0	1	0.43

Before exploring the differences among the measures, we can compare their overall sensitivity to shifts in the levels of democracy. That is, how often do these indices change, and when they do, what is the average size of that change? Figure 5.8 charts the relationship between the proportion of years in which a measure indicates a change and the average magnitude of that change for each measure.

The analysis reveals some fundamental differences in how measures track changes. For example, Freedom House records more than twice as many changes in scores as Polity IV for the same time period. The two most sensitive measures are UDS and CAM, which change every country-year (in the case of UDS, this is not surprising, since it's derived from a number of other measures). At the same time, there is a strong linear relationship between the number of changes and the magnitude of the change. Stable indices are more likely to record large changes; the most extreme case is PACL, which switches classifications only twice (Georgia and Kyrgyzstan are both reclassified as democracies after their color revolutions) but the magnitude of change is great. As we will see later, measure sensitivity may affect the plausibility of regional comparisons – to a non-area specialist, judging the region solely by the PACL measure would suggest that democracy is on the rise.

follows: Polity IV – 20; Freedom House – 14; EIU – 9.88; UDS – 4.22; CAM – 17.35; Vanhanen – 49; PACL – 1). The CAM measure was obtained by multiplying the two sub-scores (the formula for the FH ratio was one minus the ratio to make higher numbers represent more democracy). Since PACL was already scored as 0 or 1, it did not need to be modified.

Table 5.5. Average differences between measures

E23(5)	FH	POL	UDS	VAN	CAM	PACL
FH		-0.177	-0.111	0.072	-0.063	-0.055
POL	0.177	-	0.063	0.246	0.107	0.120
UDS	0.111	-0.063	-	0.181	0.039	0.057
VAN	-0.072	-0.246	-0.181	_	-0.146	-0.103
CAM	0.063	-0.107	-0.039	0.146	_	0.040
PACL	0.055	-0.120	-0.057	0.103	-0.040	-



Figure 5.8 Measure sensitivity. The x-axis shows the average magnitude of annual changes in scores for each measure, compared to the proportion of country-years in which a change was recorded (y-axis).

To go deeper into disagreements among measures, Table 5.5 reveals the average difference between each pair of measures.

The number in each cell represents the average difference between the row measure and the column measure. For example, the average difference between Polity IV and Freedom House measures for the same country-years is 0.177, which suggests that Polity scores are on average 17 percent higher than Freedom House scores. On average,



Figure 5.9 Comparisons of Freedom House and Polity IV evaluations of democracy in the former Soviet republics, 1991-2010

Polity IV appears to score countries higher than other measures, while Vanhanen scores them lower.

Of course, such averages may conceal significant variation across countries. Do measures disagree significantly about particular countries? Figure 5.9 compares Freedom House and Polity IV scores for each of the fifteen post-Soviet states from 1991 to 2010.

As the graphs show, there are a number of countries where the measures are basically in agreement. These tend to be clear-cut cases

Lost in the gray zone

such as Turkmenistan on the autocratic end of the spectrum or Estonia on the democratic end (although even in the latter case there are some minor variations in trends). In other countries, the overall assessment is similar but there are noticeable differences in trends. In Tajikistan, for example, Polity records a decrease in democratic quality in the early 2000s while Freedom House records an *increase* during approximately the same period. Azerbaijan is coded by Polity as becoming less democratic in its early history, while Freedom House records a slight increase during the same period.

For a few countries, the two indices paint very different pictures. Armenia experiences a large but temporary dip in its Polity score in the mid-1990s, while the corresponding decrease in the Freedom House rating is much less noticeable. Georgia, Armenia, Moldova, and Ukraine are consistently ranked lower by Freedom House. The levels might not matter if the two measures had different global distributions - it might be the case, for example, that Freedom House ranks countries lower in general, which would mean that the relative rankings for these countries are the same for both indices. This is not the case, however. In 2008, for example, Moldova received a 4 from Freedom House and an 18 from Polity IV. By comparison, the average Polity score of countries that received a 4 from Freedom House in that year was less than 13. Freedom House ranked Moldova on par with Kuwait and Bangladesh, while Polity ranked it on par with Belgium and South Korea. In short, the two indices do rank Moldova differently. Such discrepancy is greatest in Russia, where the two scores diverge widely. Starting in the mid-1990s, Freedom House records a steep decline in democratic quality, while the Polity index records a slight rise followed by a leveling off.

Below, I will examine the sources of disagreement between Polity and Freedom House more closely, focusing on the cases of Armenia and Russia. But such discrepancies are not limited to those two measures. Figure 5.10 shows how seven different measures assess the quality of democracy in Kazakhstan.

In the case of Kazakhstan, Polity IV and Freedom House are basically in agreement for the past ten years, although the decline in Polity scores is much more gradual (Freedom House sees the country as stabilizing its regime by 1994, whereas for Polity the process takes a decade longer). Vanhanen, on the other hand, records a gradual improvement in democracy during the late 1990s. Depending on which measure is



138



used, Kazakhstan can be judged as becoming less democratic (CAM, Polity, Freedom House), remaining a stable autocracy (UDS, PACL) or undergoing partial democratization (Vanhanen). At the same time, the measures do converge over time, which may indicate either better assessment as time goes on or the rise of conventional wisdom among some coders about Kazakhstan's rough position on the democracy continuum.

Nor is Kazakhstan particularly exceptional in the level of variation among measures – Appendix 2 includes similar charts for each of the fifteen countries, which reveal that disagreement among measures is common though not universal. The measures do seem to be in rough agreement about clear-cut autocracies and democracies, which brings us back to the question of greater disagreement about hybrid regimes. To find out, I compared the average measure variance for each countryyear with its level of democracy (measured as the average of the indices rescaled from zero to one).⁵⁰ Figure 5.11 shows the results of this comparison, revealing a clear curvilinear relationship between the Lost in the gray zone



Figure 5.11 Disagreement among measures by regime type

degree of democracy and measure variance. Countries in the middle (0.4 to 0.6 range) have consistently higher measure disagreement. A lowess plot (locally weighted scatterplot smoothing, designed to look for non-linear relationships) shows an inverse-U relationship between democracy and measure variance.

To see if measure variance is consistent across countries, Figure 5.12 compares the average disagreement for each country with its overall democracy score. We would expect middle-ranking countries to have higher levels of variance.

Once again, the results show agreement among measures on clearcut autocracies like Turkmenistan and clear-cut democracies like the Baltic states. Disagreement increases for more democratic countries and peaks for countries in the middle of the range before declining once more. There is a significant disagreement among measures about hybrid regimes such as Moldova, Armenia, Georgia, and Russia. Overall, these results suggest that measures of democracy for these countries may be particularly unreliable, and that conclusions about the causes and effects of democratic development in hybrid regimes are particularly sensitive to the choice of measure.

⁵⁰ The following calculations were performed without PACL due to its dichotomous distribution.



Figure 5.12 Disagreement among measures by country

Measure variance and regional comparisons

How much do the indices agree with each other in terms of capturing overall regional trends? And how do the differences in these measures affect the comparison of the former Soviet republics with other regions in the world? Figure 5.13 shows the regional average between 1991 and 2010 according to five different indices.

Once again, there are significant variations among measures. Polity generally ranks the region as more democratic than other indices, while Vanhanen ranks it as less democratic. Of course, looking at levels alone may not tell us much about how the region compares to other countries in the world. For example, if Vanhanen has a distribution that clusters toward the bottom, a lower regional average would not translate into a lower ranking for the former Soviet republics compared to other regions. To see if this is the case, Figure 5.14 compares the average level of democracy in the former Soviet republics as a proportion of the average level of democracy in Eastern Europe.⁵¹

⁵¹ Eastern Europe is defined as countries in existence between 1991 and 2010 in eastern and southeast Europe: Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Czechoslovakia, Czech Republic, Hungary, Kosovo, Macedonia, Montenegro, Poland, Romania, Slovakia, Slovenia, and Yugoslavia. Lost in the gray zone



Figure 5.13 Average regional measures of democracy in former Soviet republics for five different indices, 1991–2010



Figure 5.14 Average level of democracy in the former Soviet republics, as a proportion of the average level of democracy in Eastern Europe

Seva Gunitsky

There is broad overall agreement among the indices that the former Soviet republics are on average less democratic than countries in Eastern Europe. Yet there are disagreements about the size of that difference, as well as about trends over time. According to Polity, the region has become steadily less democratic compared to Eastern Europe, leveling off in the early 2000s. Freedom House also records a sharp relative decline, albeit one that shows no signs of leveling off. Vanhanen, on the other hand, records no such trend – Eastern Europe is generally more democratic, but this relationship is stable over time. Finally, PACL rates the former Soviet republics as much less democratic, but records a relative improvement after 2004. This change is driven by just two measures, with Georgia and Kyrgyzstan re-coded as democracies following their color revolutions.

A scholar comparing democratic development in the two regions using Polity or (especially) Freedom House would note the steady deterioration of democracy compared to Eastern Europe, and wonder why post-Communist transitions led to such different trajectories in the two regions. A scholar using Vanhanen would note that the region has historically been less democratic than Eastern Europe, but has not become more so since the Soviet collapse. A scholar using PACL would conclude that the former Soviet republics have historically been, on average, not nearly as democratic as Eastern Europe, but recently experienced a large relative gain. Avoiding this spurious conclusion would require some knowledge of the region as well as the coding particularities of that measure.

Given this confusion, it's not surprising that area specialists are critical of large-n measures of democracy. Regional comparisons that use large-n statistics must pay careful attention to both the choice of measure and the state of facts on the ground.

Sources of disagreement: Freedom House and Polity IV in Armenia and Russia

Turning back to disagreements about specific countries: why do measures sometimes present very different pictures of the same event? To explore this in more detail, I examine two examples of disagreement between Freedom House and Polity IV, since these are both very commonly used measures of democratization. First, what causes the



Figure 5.15 Evaluations of democratic quality in Armenia by Freedom House and Polity IV

large dip (and recovery) in Armenia's Polity score in the mid-1990s, a trend not captured or much less pronounced in any other index? Second, why do the Freedom House and Polity measures for Russia diverge so widely in the mid-1990s? Russia is a particularly important case since it forms the background for much analysis of both foreign policy and domestic developments in the post-Soviet region.

In the winter of 1992–1993, the Armenian Revolutionary Federation (ARF) party united with other opposition parties to stage large demonstrations against the government of Levon Ter-Petrossian, who had been elected in a landslide in 1991. ARF is a nationalist party with strong socialist roots that advocates the creation of a Greater Armenia. It enjoys support among segments of the Armenian population and especially among the country's diaspora. In 1994, Ter-Petrossian banned ARF and arrested thirty-one of its leaders. He accused the party of plotting to overthrow the government and of having foreignbased leadership. The move decreased his popularity both with domestic groups and among the diaspora in particular. In the 1996 presidential election, his share of votes fell to 52 percent (compared to 83 percent in the previous election in 1991). The main opposition

candidate, a former prime minister, received 42 percent of the vote, in contrast to the 7.2 percent received by the runner-up in the 1991 election. The vote brought accusations of voting irregularities, leading to large public protests, suppressed when Ter-Petrossian sent tanks into the nation's capital.

What happens with Armenia's Polity sub-scores between 1994 and 1996? Openness of executive recruitment remained the same, while the regulation of participation actually improved. The big drops were in measures of executive constraints (5 to 3), competitiveness of executive recruitment (3 to 1), and the competitiveness of participation (4 to 2). Since executive constraint is the most heavily weighted sub-component, a decline in this indicator is particularly influential for the overall score. In Freedom House measures, the civil liberties indicator moves from 3 to 4 in 1993 and stays at that level through the present day. The political rights indicator shows worsening from 3 (1993-1994) to 4 (1995) to 5 (1996-1997) before returning to 4 (1998–2003). The stability of civil liberties during this period is puzzling, because the Freedom House media freedom index for Armenia increases from low to high 50s (a higher number equals less media freedom). In the Freedom House rankings, Armenia declines from 87th place in 1994 to 122nd place in 1996 (out of 191), moving from the 55th to the 69th percentile. The decline is much more precipitous for Polity IV, where Armenia moves from 54th to 119th place (out of 159), or from the 33rd to the 75th percentile. An observer of Armenian democracy who used Polity scores would conclude that the democratic crisis of 1994–1996 was much more severe than would an observer using Freedom House scores.

It is difficult to determine precisely how Freedom House gauges democracy in Armenia during this period, because disaggregated scores are not available until 2006 (this once again demonstrates the need for transparency and disaggregation in democracy measures). A reasonable guess is that while Polity IV focused on elite behavior, Freedom House focused on the democratic situation as a whole. Both measures recorded a decline, but the Polity IV decline was particularly large because the banning of opposition parties signaled a loss of executive constraint, which Polity considers to be the single greatest indicator of democratic quality. Freedom House took the same development as negatively affecting political freedom at the elite level but not the individual rights of Armenia's citizens. In short, the disagreement



Figure 5.16 Evaluations of democratic quality in Russia by Freedom House and Polity IV

among measures appears to stem from the different ways they conceptualize democracy and the different emphasis they place on the relative importance of elite constraints and individual liberty.

Turning to Russia, the discrepancy between the two measures is striking, as illustrated in the opening of the chapter (see Figure 5.16). The divergence in trends in the late 1990s is particularly puzzling. In this period, Russia's Polity IV measure improves from 13 to 16 (out of 20), while the country's Freedom House score worsens from 3.5 to 5 (out of 7).

Looking at Polity IV sub-scores during this period shows that the increase comes from a change in a single measure – executive constraint, which increases from 3 to 5 between 1999 and 2000 and stays at this level until 2007, when it declines to 4. In fact, changes in the measure of executive constraint accounts for *all* of the variation in Russia's score. For Polity, variation in media freedom or individual liberties was simply insignificant in shaping the measure's assessment of Russian democracy. It seems likely that the increase in executive constraint came from the Duma's unexpected rejection of Yeltsin's nomination of Viktor Chernomyrdin as prime minister in

Seva Gunitsky

August of 1998 (the unpopular Chernomyrdin received only 94 votes out of a possible 450). The rejection signaled legislative independence, and since executive constraint forms the most heavily weighted sub-component of Polity, the perceived increase in the power of the Duma led to a significant increase in the country's overall score.

Freedom House records Russia's level of democracy as stable (at 3.5) between 1992 and 1997. In 1998, political rights worsen from 3 to 4; in 1999, civil liberties worsen from 4 to 5; and in 1999, political rights worsen once more from 4 to 5. Both sub-components remain at 5 until 2004. Once again, it's hard to determine what drives the decline in the absence of disaggregated measures. Russia's press freedom index declined during this period, and the financial crisis of 1998 almost certainly undermined socio-economic equality, as did the increasing political clout of the oligarchs and perceptions of corruption among Yeltsin's Kremlin "family" during this period. In short, as Pemstein *et al.* conclude:

Russia seems to be a case where two measures looked at the same information and came to two different conclusions. Freedom House thought the events from 1996 on indicated a weakening of democracy,⁵² while Polity thought these same events led to a strengthening of democracy ... The raters make judgments that are sensible, but incomplete. Both measures cannot be correct but neither have both judges completely missed the mark.⁵³

Curiously, the 1996 presidential election, which had been marred by accusations of fraud and American interference, does not lead to a decrease in Freedom House scores. And if the rise of the oligarchs caused a decline in Freedom House scores, their decline under Putin has not brought a corresponding increase. The measure provides ammunition for critics who argue that Freedom House promotes a pro-American agenda. The organization's treatment of Russia has in fact been criticized both by Russian officials and Western academics. Treisman (2011: 341–342) notes that Russian civil liberties are ranked on par with Yemen's, where legislation grants capital punishment for women who commit adultery. Russia's political rights, meanwhile, are ranked on par with the United Arab Emirates, a collection of monarchies that did not hold any elections until 2006.⁵⁴ According to Freedom House, UAE lacks political parties or an independent judiciary, and severely restricts freedom of expression.⁵⁵ In 1998, Polity IV gave Russia a score of 14 (out of 20), while Saudi Arabia and UAE received scores of 0 and 2, respectively. In that year, countries that shared Russia's Freedom House evaluation averaged a score of only 6 in Polity IV. As Treisman concludes:

At a minimum, an acceptable cross-national rating of democracies should be able to distinguish between the kind of system in Russia and government by a federation of dynastic monarchies free from any checks whatsoever, as in the United Arab Emirates. This rules out the Freedom House index.⁵⁶

Turning instead to Polity IV, Treisman finds that by the latter measure Russia's degree of political freedom is unexceptional when compared to countries with similar incomes.⁵⁷

How do other indices measure Russia? As Figure 5.17 shows, most of the other measures fall somewhere between Polity and Freedom House (the exception is PACL, who consistently code Russia as an autocracy). Vanhanen records a sharp spike in 1993, followed by an equally sharp decline after 2002. CAM, UDS, and EIU grade Russia as more democratic than Freedom House but less

⁵² The authors make a minor mistake here: the decline in Freedom House subscores does not begin until 1998.

⁵³ Pemstein, Meserve, and Melton 2010: 429.

⁵⁴ And even then, the elections were very limited: an electoral college of less than seven thousand people selected by the monarchs was allowed to elect half of the members of the Federal National Council. The other half was directly appointed by the government.

⁵⁵ Freedom House 2011 Country Report: United Arab Emirates (http://www.free domhouse.org).

⁵⁶ Treisman 2011: 352.

⁵⁷ As measured by per capita GDP adjusted for purchasing power parity (Treisman 2011: 352–353). Treisman (2011: 481) notes that even Polity "may not be immune to outside pressure"; in the 2009 update to the index, Russia's score was retroactively lowered by a point for all years after 1992, while Estonia's score was retroactively raised from 16 to 19 during the same period. Since the adjustments were made without substantive justification, Treisman contacted Polity's project director to inquire if the revision was related to the fact that since the late 1990s Polity has been funded by the US government's Political Instability Task Force, which is itself funded by the CIA. The director replied that the perspectives of government supporters were considered as seriously as those of academics and other experts, and that "there has never been any serious arm-twisting from either side."



Figure 5.17 Assessment of democracy in Russia using six different measures

democratic than Polity IV, with UDS and EIU recording a slight decline in recent years.

In short, there is wide variation in how the measures view the quality of democracy in Russia. Whether the source is ideological predispositions or the conceptual emphasis on certain elements of democracy over others, this disagreement suggests that the analysis of political and economic development in Russia will be particularly sensitive to the choice of measure.

Conclusion

148

As this chapter has sought to demonstrate, measures of democracy in the former Soviet republics are fraught with a number of weaknesses. They disagree about particular countries, and about mixed regimes in particular. Because their definitions focus on different aspects of democracy, they sometimes draw different conclusions from the same political outcome. Moreover, since each measure captures different elements of democratic governance, the quest for a "perfect" measure is a futile one. As Munck and Verkuilen conclude, "no single index offers a satisfactory response" to the challenges of definition, Lost in the gray zone

measurement, and aggregation, and "even the strongest indices suffer from weaknesses of some importance."⁵⁸

The historical scope of the measure also affects which theories it can test. Arguments about the historical causes and correlates of war may require a longer reach into the past; as Coppedge and Gerring put it, "we suspect that the enduring value of Polity stems partly from its comprehensive historical coverage."⁵⁹ The scope of the data can also affect substantive conclusions. The "common restriction of data sets to the post-World War II period," argue Munck and Verkuilen, limits the types of theories they can test.⁶⁰ For example, examining the evidence for democratic waves, Przeworski et al. criticize Huntington for using a measure based on the percentage of democratic states, and find no evidence for waves when using the criterion of transitions rather than institutional changes.⁶¹ But since two of the three democratic waves of the twentieth century occur before the period in their data set (1950-1990), while the third takes off just as their analysis is truncated, the failure to find evidence of waves is understandable but hardly convincing. Research that examines cross-border diffusion of democracy, particularly before World War II, would thus be better off using Polity IV or Vanhanen for their empirical tests.

The various problems of large-n democracy indices don't mean that statistical measures ought to be abandoned entirely, or even that the application of these measures will inevitably be flawed. However, two crucial caveats are in order. First, once the choice of measure is made, its inherent biases and limitations must be made clear when evaluating the causes and consequences of democratization. This is particularly important for policymakers assessing the impact of external assistance on hybrid regimes, where rankings are especially problematic.

Second, the choice of measure must be justified in relation to what is actually being examined. Since each measure has its own strengths and weaknesses, the nature of the research shapes which strengths are most suitable for the issue being examined, and which weaknesses can most plausibly be overlooked. Binary measures may work well for examining regime duration or when looking at cases of profound political transformations, but less so for examining more modest changes within regimes.⁶²

⁵⁸ Munck and Verkuilen 2002: 28.

Munck and Verkullen 2002: 6. ⁶¹ Przewor

⁶² Collier and Adcock 1999.

⁵⁹ Coppedge and Gerring 2011: 249.

⁶¹ Przeworski *et al.* 2000: 40–45.

Freedom House should not be used to evaluate democracy's relationship with corruption or economic equality, because these substantive elements are already built into the measure itself. Polity IV measures might be appropriate for research that examines constraints on governing elites, but not for studying the expansion of suffrage over the nineteenth century.

Measures of democracy can mislead as much as they clarify, as the first figure in this chapter demonstrates all too clearly. Moreover, the consequences of poor measurement extend beyond scholarly analysis and affect advocacy efforts and policy choices. While a more careful use of these measures is hardly a panacea, self-consciousness about the drawbacks and underlying assumptions of indices can only serve to improve our understanding of democratic development. Highlighting the limitations of a measure can also highlight its strengths.

6 Winning the rankings game The Republic of Georgia, USAID, and the Doing Business Project

SAM SCHUETH

Although, or perhaps because, there is no consensus about what constitutes place competitiveness (Bristow 2005), recent decades have witnessed a proliferation of regimes of measuring international competitiveness, variously defined, and applied to cities, regions, and states (Malecki 2002; 2004). However, the rise of international rankings appears to have crystallized more specific understandings of competitiveness and judgments of what constitutes a business-friendly state. For many years, the World Economic Forum's annual Global Competitiveness Index (GCI) was the best-known regime of measuring and ranking the economic competitiveness of states (Schwab and Porter 2008). More recently, however, the GCI has been displaced by the World Bank's Ease of Doing Business Index (DBI), which enrolls a global network of analysts to rank state competitiveness using indicators of regulatory quality (World Bank and IFC 2009). Published annually since 2003, DBI is designed to further the World Bank's long-term effort to transform "second" and "third" world countries into "emerging" and "frontier" markets, where economic development is financed by foreign investment (Ó Tuathail 1997; Sidaway and Pryke 2000; Lavelle 2000; Lee 2003).

In terms of its role as an international ranking in this volume's framework, DBI can be viewed as both a source of regulatory power and of expert judgment. More than a schema for measuring competitiveness, DBI is also a tool of what Peck (2002; 2004) calls "neoliberal fast-policy transfer" and "reform at a distance." It works by fostering international competition between policymakers to raise their state's ranking, thereby demonstrating to international investors a commitment to business-friendly policies. Indeed, the considerable authority carried by DBI's institutional sponsor, the World Bank, may explain why the ranking has so quickly become a widely cited assessment of a