How to Measure Legislative District Compactness If You Only Know it When You See it*

Aaron Kaufman† Gary King‡ Mayya Komisarchik§

July 11, 2017

Abstract

The US Supreme Court, many state constitutions, and numerous judicial opinions require that legislative districts be "compact," a concept assumed so *simple* that the only definition given in the law is "you know it when you see it." Academics, in contrast, have concluded that the concept is so *complex* that it has multiple theoretical dimensions requiring large numbers of conflicting empirical measures. We hypothesize that both are correct — that the concept is complex and multidimensional, but one particular unidimensional ordering represents a common understanding of compactness in the law and across people. We develop a survey method designed to elicit this understanding with high levels of intracoder and intercoder reliability (even though the standard paired comparison approach fails). We then create a statistical model that predicts, with high accuracy and solely from the geometric features of the district, compactness evaluations by judges and other public officials from many jurisdictions, as well as redistricting consultants and expert witnesses, law professors, law students, graduate students, undergraduates, ordinary citizens, and Mechanical Turk workers. As a companion to this paper, we offer data on compactness from our validated measure for 18,215 US state legislative and congressional districts, as well as software to compute this measure from any district shape. We also discuss what may be the wider applicability of our general methodological approach to measuring important concepts that you only know when you see.

*The current version of this paper is available at http://j.mp/Compactness. Our thanks to Steve Ansolabehere, Ryan Enos, Dan Gilbert, Jim Griener, Bernie Grofman, Andrew Ho, James Honaker, Justin Levitt, Luke Miratrix, Max Palmer, Stephen Pettigrew, Larry Tribe, Robert Ward, and participants in "A Causal Lab" for helpful data or suggestions; and to Stacy Bogan, the Center for Geographic Analysis, and the Institute for Quantitative Social Science at Harvard University for research assistance and support.

†PhD Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; aaronrkaufman.com; aaronkaufman@fas.harvard.edu, (818) 263-5583.

‡Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu, (617) 500-7570.

§PhD Candidate, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; scholar.harvard.edu/mkomisarchik; mkomisarchik@fas.harvard.edu, (720) 220-9328.

1 Introduction

Compactness is a widely valued but ill-defined normative criterion in the law for drawing legislative districts. The US Supreme Court is explicit about the ambiguity: "One need not use Justice Stewart's classic definition of obscenity—'I know it when I see it'—as an ultimate standard for judging the constitutionality of a gerrymander to recognize that dramatically irregular shapes may have sufficient probative force to call for an explanation" (*Karcher v. Daggett*, 462 U.S. 725, 755, 1983). Even so, the Court does not go beyond this statement to define the concept.

Requirements for compactness in state constitutions are equally vague. For example, the Constitution of Illinois says "Legislative Districts shall be compact…". The Constitution of Hawaii requires that "Insofar as practicable, districts shall be compact." In Arizona, the Constitution orders that "Districts shall be geographically compact and contiguous to the extent practicable."[1]

The most general justification for compactness in the law (however defined) is as one of the "traditional redistricting principles" which, when followed, can "defeat a claim that a district has been gerrymandered…" on the basis of race (*Shaw v. Reno*, 509 U.S. 630, 647, 1993) or political party (*Davis v. Bandemer*, 478 U.S. 173, 2815, 1986). The empirical claim that compactness requirements constrain racial or partisan gerrymandering is hardly a settled empirical question (Altman and McDonald, 2012; Barabas and Jerit, 2004; Chen, Rodden, et al., 2013), and the role of compactness in ensuring other important normative virtues — such as better knowledge, communication, and trust between representatives and citizens — is also contested (Cain, 1984; Pildes and Niemi, 1993). But regardless of the outcome of these debates, the degree of compactness of legislative districts will remain important because of its essential role in defining the nature of

[1] Some states have passed laws highlighting certain features of compactness that help with intuition if not precision. For example, Virginia passed Senate Joint Resolution 224 (1/14/2015, Article II, Section 6(5)) which reads "Each legislative and congressional district shall be composed of compact territory. Districts shall not be oddly shaped or have irregular or contorted boundaries, unless justified because the district adheres to political subdivision lines. Fingers or tendrils extending from a district core shall be avoided, as shall thin and elongated districts and districts with multiple core populations connected by thin strips of land or water…." Iowa (Iowa Code, Title II §42.4) and Michigan (Congressional Redistricting Act 221 of 1999, Redistricting plan guidelines) discuss precise measures but how to use this information is not specified.

representation and electoral competition in most modern democracies.[2]

Although some scholars conclude that applying "such a hazy and ill-defined concept" to the law is "impossible" (Young, 1988, p.113), all legal uses of the term assume that compactness is a single coherent concept, easily discernible by viewing a district's shape. Yet, for over a century, insightful scholarly efforts to quantify this concept have produced increasing numbers of conflicting measures by comparing districts to different "theoretically ideal" shapes (e.g., Niemi et al., 1990). An agreed upon quantitative measure would empower a wide range of scholarly research, but it would also be of considerable use in practical redistricting cases as way of reigning in advocates making otherwise unconstrained qualitative arguments that their preferred plans are always the most compact. Although constraining partisans is possible when judges "know it when they see it," an agreed upon quantitative measure would greatly simplify the process and, to some extent, curb this behavior.

We attempt to span this divide, between the simplicity assumed in the law and the theoretical complexity and multidimensionality revealed in social science research, by inferring, measuring, and validating the single underlying dimension of compactness the law seems to require. Since compactness in the law is defined by the qualitative judgment of nonquantitative human observers, such as redistricters, lawyers, and judges, the assumption of a single dimension would only be supported if most educated people evaluate a district's compactness in the same way, and so we show that indeed they do.

We begin in Section 2 by inductively defining this underlying dimension of compactness. We do this by building on the encyclopedia of existing diverse measures, adding new ones that fit with how humans perceive objects like district shapes, and providing intuition about the commonly perceived dimension we seek to measure.

Then, in Section 3, we develop a way to measure this concept by eliciting views of the compactness of specific districts from respondents via a novel survey approach, and rank

[2]Measuring compactness has also long been important to other areas, such in urban politics (e.g., causes of urban sprawl; Tsai 2005), geography (e.g., shapes of Sudanese villages; Lee and Sallee 1970), marine geology (e.g., the shapes of atolls; Stoddart 1965), Paleontology (e.g., the shapes of sand crystals; Cox 1927), geographic information systems (e.g., the accuracy of choroplethic maps; Hsu and Robinson 1970), among others.

ordering districts by levels of compactness. We were forced to develop a new method because the standard approach in the survey literature to a problem like this, Thurstone's venerable paired comparisons, completely fails in our application. The high levels of intercoder and intracoder reliability produced by our alternative approach are consistent with the unidimensionality hypothesis (and suggest that our survey methodology may have other applications). This section then uses these results to build a statistical model that predicts with high accuracy how individuals rank districts, given only the shape of a district.

These results enable us to apply one of the most important principles of statistics — defining the quantity of interest separately from the measure used to estimate it — and, as a result, to provide evaluations that make our approach vulnerable to being proven wrong. We do this in Section 4 with cross-validation and then extensive out-of-sample validations in samples of public officials including judges from many jurisdictions, as well as redistricting consultants and expert witnesses, law professors, law students, graduate students, undergraduates, ordinary citizens, and Mechanical Turk workers. Section 5 concludes.

2 Conceptualizing Compactness

In this section, we attempt to inductively characterize the concept of compactness that most laws, constitutions, and judicial opinions assume human observers intuitively understand.

As districting is "one area in which appearances do matter" (*Shaw v. Reno*, 509 U.S. 630, 647, 1993), our approach is to measure the compactness of the geometric shape of a district, separately from other facts that can impact it. This is the most common basis for a compactness definition, dating well before the famous "Gerry-Mander" cartoon (Tisdale, 1812), but not the only one possible. In other words, our goal is to define and estimate *absolute* compactness based on district shape only. Absolute compactness, in turn, may be constrained or influenced by fixed features of the state geography, such as rivers, coastlines, or highways. Our goal is to measure the quantity that would be influenced by these features, so that it measures the concept in the law and can be useful

for further research. If a researcher had the alternative goal of defining and measuring relative compactness, based on how close it is to a realistic ideal, then our measure should be regarded as a potentially useful first step.

The approach we develop here can also be applied to some other redistricting criteria if additional data are available (or to other unrelated concepts that you only know when you see). These may include other characteristics of districts such as size; population equality across districts; where people live within a district (Fryer Jr and Holden, 2011); whether the district divides communities of interest or local political subdivisions; whether incumbents are paired or grouped in the same district and so have to run against each other to keep their jobs; what types of people are included in or excluded from a district; and, as a result, partisan fairness, electoral responsiveness (Gelman and King, 1994b; Grofman and King, 2007), and racial fairness (King, Bruce, and Gelman, 1996). Redistricting also influences more personalistic factors (that we have seen debated in real redistricting cases), such as whether a specific district includes features like a military base (which can influence a candidate's policy preferences) or a prison (which counts under "equal population" requirements but not votes), or even whether a candidate's parents homes or children's schools are drawn out of his or her district.

Section 2.1 highlights empirical inconsistencies in existing shape-based measures to convey that the possible conceptual definitions of compactness, underlying these measures, are multidimensional. Then Section 2.2 provides intuition and tools to build toward a single concept of compactness.

2.1 Multiple Dimensions Underlying Existing Measures

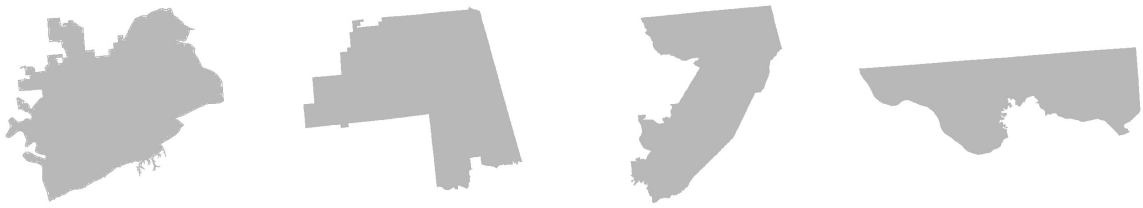
Numerous specific compactness measures have been proposed in the academic literature, each one fitting different qualitative conceptual definitions and intuitions for certain geographical configurations and violating it for others (Altman, 1998; Niemi et al., 1990; Stoddart, 1965; Young, 1988). These measures are based on geometric concepts such as perimeters, areas, vertices, and centroids, often in comparison with some pure form geometric object such as a circle, rectangle, polygon, or convex hull. Each, however, focuses on a different dimension of what might be called compactness. Consider, for example,

the five most frequently used measures by academic researchers, and also by experts in redistricting litigation: *Length-Width Ratio*, the ratio of the length to the width of the minimum bounding rectangle (C. C. Harris 1964; Timmerman, 100 N.Y.S. 57, 51 Misc. Rep. 192 (N.Y. Sup. 1906)); *Convex Hull*, the ratio of the area of the district to the area of the minimum bounding convex hull; *Reock*, the ratio of the area of the district to the area of a minimum bounding circle (Reock, 1961); *Polsby-Popper*, the ratio of the area of the district to the area of the circle with the same perimeter as the district (Polsby and Popper, 1991; Schwartzberg, 1965); and (modified) *Boyce-Clark*, the (normalized) mean absolute deviation in the radial lines from the centroid of the district to its vertices (Boyce and Clark, 1964; Kaiser, 1966; MacEachren, 1985). For details on these and others, see Appendix A.

Without a gold standard, we cannot determine any measure's formal statistical properties, its error rates, or any hint of when it might fail. Although different measures are sometimes correlated, choices among these are presently made by qualitative judgment. Creative scholars have managed to use existing measures productively in research by combining multiple measures, adjusting or weighting each for specific purposes, or making careful qualitative decisions in specific cases (Ansolabehere and Palmer, 2016; Niemi et al., 1990).

We illustrate the issues with measuring compactness by presenting a set of four districts in Figure 1. This includes four state house districts from Alabama in 2000. Readers may wish to draw their own conclusions about the relative compactness of these districts, but we now provide in Table 1 an indication of how the most popular five measures rank them. As can be seen from the first five rows of Table 1, every one of these measures gives a different rank order for the four districts. We introduce two new compactness measures in Section 2.2 for a different purpose; these are given at the bottom of Table 1 and also give unique rankings of the same districts. This example is merely a proof of concept, but finding such examples is easy: By random sampling, we estimate that in our collection of 18,215 state legislative and congressional districts (see Appendix B), there exist 162 trillion sets of four districts such that every one of the seven measures provides

a unique rank order. Of course, there is a large number from which to choose (this large number being about 0.15% of the total), but inconsistencies among in rankings on fewer than seven measures is both commonplace and is congruent with the long literature on this subject.



(a) AL 1 (b) AL 37 (c) AL 23 (d) AL 2

Figure 1: Four Districts from the Alabama State House in 2000.

	Legislative Districts			
	(a) AL 1	(b) AL 37	(c) AL 23	(d) AL 2
Convex Hull	4	3	2	1
Reock	1	2	3	4
Polsby-Popper	4	1	2	3
Boyce-Clark	2	3	1	4
Length/Width	3	2	1	4
X-Symmetry	1	4	3	2
Significant Corners	4	1	3	2

Table 1: Seven Unique Compactness Rankings of the Same Four Districts: Five Existing and Two New Metrics

2.2 Toward a Single Compactness Dimension

We now provide intuition helpful in turning the multiple types and dimensions of compactness illustrated in Section 2.1 into a single unidimensional concept underlying common conceptions. We continue to proceed inductively, with Section 3 devoted to measuring this concept.

Since we are attempting to quantify human perception, we try to avoid imposing theoretical notions of what compactness should be, what might be rational, or what meets various mathematically pure standards. Human perception has been described as suffering from various psychological biases, illusions, and frailties (Kahneman, 2011), but for

our purposes we treat these as features to measure rather than problems to avoid. We do this in two ways, followed by a characterization of the entire dimension of interest.

First, given the absence of some gold standard or legal definition, researchers have developed compactness measures by ensuring that they meet aesthetically pleasing theoretical or mathematical standards. One key issue is that all existing compactness measures are *rotationally invariant*, meaning that if we rotate a district, say 63 degrees, a compactness metric should remain the same. This seems like a perfectly reasonable theoretical standard, but it turns out to also be a deeply normative choice, which means it is a preference scholars have asserted rather than an objectively correct standard.

In fact, if we do "know it when we see it" as required in the law, measures that are rotationally invariant are blind to certain human perceptions. The reason is that human perception is famously sensitive to the rotation of objects: Even familiar faces can become unrecognizable when viewed upside down (e.g., Maurer, Le Grand, and Mondloch, 2002). Our own experimentation suggests that people sometimes view long thin district shapes located on a diagonal (such as left to bottom right; ) as less compact than the same shape located along the horizontal or vertical axis (). In contrast, legislative districts always have a well defined up (north) and down (south), as displayed on every commonly used map. Indeed, courts, redistricters, and judges do not rotate districts when judging compactness; they merely look at them on the map in the usual orientation. In other words, since the usual orientation of a district has precedence in how humans interpret it, some of our measures should also be sensitive to this orientation instead of being rotationally invariant.

Thus, primarily for illustration in this section, and later as a measurable feature of district shape we include in our statistical model, we define here a new compactness measure that is not rotationally invariant. We do not intend this measure to substitute for other measures or to even be especially important on its own, but it will be useful to convey our point and tap into this aspect of compactness. Thus, we define *X-Symmetry* by dividing the overlapping area, between a district and its reflection across the horizontal axis, by the area of the original district. Shapes like circles and rectangles have overlap regions equal

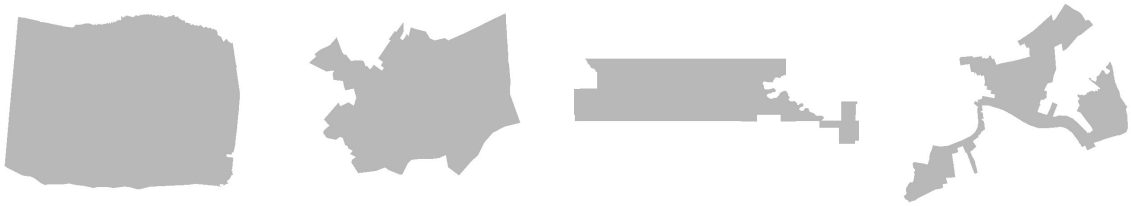
to that of the original district and so have X-Symmetry values of 1. A long thin district stretched out from top left to bottom right, or one like  , have X-Symmetry values close to zero. This measure, applied to the four districts in Figure 1, gives unique rankings for each; see the sixth row of Table 1.

Second, another feature of human perception is how we define what constitutes a "significant" feature of a district. If a roughly circular district has a ragged boarder, which of the small border inlets and peninsulas count as deviations from the circular shape? For example, suppose we give a large number of people the task of drawing from memory the shape of the continental United States. These drawings will all differ, but they will likely all include some of the same features — a roughly rectangular shape, a peninsula for Florida, a larger one for New England, and perhaps a somewhat rounded western ocean boarder. In other words, despite the enormous number of specific small features and vertices along the boarder to choose from, virtually all Americans are likely to recall, thus judging as significant, a small number of the same features.

To include this highly qualitative feature of human perception, we consider algorithms computer scientists design to list all of the "objects" in an image. There is no correct answer, but it turns out that different people are likely to give similar answers, and the automation goal is to list the objects a human would identify. As we did with X-Symmetry, we illustrate this idea quantitatively, and give an example that will later become part of our model. To do this, we turn the geometric district shape into a set of pixels (i.e., changing from vector to raster representation), apply the Harris corner detection algorithm (C. Harris and Stephens, 1988), and count the number of "significant" corners. The more significant corners, the less compact the district by this metric. The last row of Table 1 gives the rankings of the four districts in Figure 1 according to the number of significant corners. This measure also gives the four districts a unique ordering.

Finally, we try to convey intuition about the underlying dimension of compactness we will quantify in the next section. We do this visually, by presenting in Figure 2 a set of districts that range from most (panel a) to least (panel d) compact. We find that almost anyone familiar with the district-based nature of modern democracy, and some sense of

the word compactness, finds that district (a) is more compact than (b), which is more compact than (c), which is more compact than (d). The question is how to quantify this notion, so that it works for these four districts and all other geometric shapes, a topic to which we now turn.



(a) (b) (c) (d)

Figure 2: The Underlying Compactness Dimension, from most compact (a) to least compact (d) (all five of the most common compactness measures in agree with this ordering). (Districts include, (a) Wyoming State House District 42, 2010; (b) Pennsylvania State House District 185, 2010; (c) Oklahoma Congressional District 1, 1950; (d) Louisiana State Senate District 3, 2010.)

3 Measuring Compactness

We now develop an explicit measure of the concept of compactness inductively defined in Section 2. The result is a method of rank ordering any set of n districts given only their geometric shapes. To do this, we first develop a method of eliciting views about compactness directly from survey respondents, something generally recognized as important but rarely done in this literature (Angel and Parent, 2011; Chou et al., 2014). Section 3.1 attempts this by applying best current practices in survey research — using a modern version (David, 1988) of Thurstone's venerable paired comparisons approach (Thurstone 1927, a method that dates at least to 1860; see Fechner 1966). Under this approach, we pose a set of simple survey questions, each asking the respondent to decide which of two districts is more compact and, from the many answers, we construct the full ranking. We explain the motivation behind this approach and then demonstrate empirically that it utterly fails to accomplish its goal for this application. Given this result, we have no choice but to develop a new approach. Thus, in Section 3.2, we turn to the method that paired comparisons was originally designed to supplant — asking respondents to rank

many districts all at once. We show that, as we apply it, this approach turns out to work extremely well in our application (and may also work for many others too). As we describe, the supposed advantages of paired comparisons turn out to be disadvantages and the disadvantages of ranking turn out to be advantages. Section 3.3 takes the resulting survey elicitation method as our outcome variable, and new gold standard, and builds a statistical model to predict it from geometric features of the districts. Details about data used appear in Appendix B.

3.1 How Paired Comparisons Fails

The method of paired comparisons has been touted for more than a century and a half for its two key advantages. First, this approach puts fewer demands on survey respondents than asking respondents to do a full ranking. That is, to produce a ranking of n items requires the choice among $n!$ possible rankings, whereas the same information can be elicited with only $\binom{n}{2}$ paired comparisons. This is not trivial since $n! \gg \binom{n}{2}$; for example, with $n = 20$, we have $20! = 2.4 \times 10^{18}$, or 2 quintillion possible rankings, whereas $\binom{20}{2} = 190$ paired comparisons is large but still manageable in a single survey (and may even be reduced; see Mitliagkas et al. 2011). For these reasons, Converse and Presser (1986, p.28) comment on a historical example with only 13 items: "Tasks of this scope were soon seen as much too difficult…, and in our own time, rank orders of this size are all but invisible in the literature". Thus, if full ranking is used, the best practice has been "not to use lists longer than three or four items" (Gideon, 2012).

Second, Thurstone's approach only requires simple questions that are easy to understand, concrete, and specific. With it, we ask a respondent which among a pair of legislative districts is more compact, and then repeat this simple question multiple times with different pairs of districts. Then, after eliciting information in this manner, the researchers combine these binary decisions into a ranked scale (using Guttman scaling or a more sophisticated approach accounting for measurement error; e.g., Mitliagkas et al. 2011). The method assumes all respondents will use the same unidimensional scale to make their choices for all their paired comparisons (an issue we return to). The supposed advantage of this approach is that respondents are asked only what they know (a paired comparison)

and researchers do what they are better at, which is taking on the complicated task of inferring the underlying full ranking from all the elicited information.

To apply this method, we conducted multiple iterated rounds of pre-testing and cognitive debriefing while adjusting question wording and how the districts appeared. But despite dozens of trials over many months, testing numerous variations, and with a wide range of research subjects, online and in person, our inter- and intracoder reliability statistics were rarely much above random chance. To see what we found, consider a simple experiment with 40 respondents (in this case on Amazon's Mechanical Turk), each asked to choose the more compact district from each of twenty pairs, producing a 20-length binary decision vector. This survey enabled us to compare the percent agreement among the 20 decisions for each of $\binom{40}{2} = 780$ pairs of respondents. Figure 3 gives a histogram of these percent agreements (in blue, marked "paired", computed as a density estimate). For comparison, we also generate a placebo test, under the null hypothesis of no agreement, by randomly generating 780 pairs of 20-length vectors and computing from them the percent agreement and plotting its histogram (white with a black outline, marked "Random"). (We discuss the "Ranking" figure in the next section.)

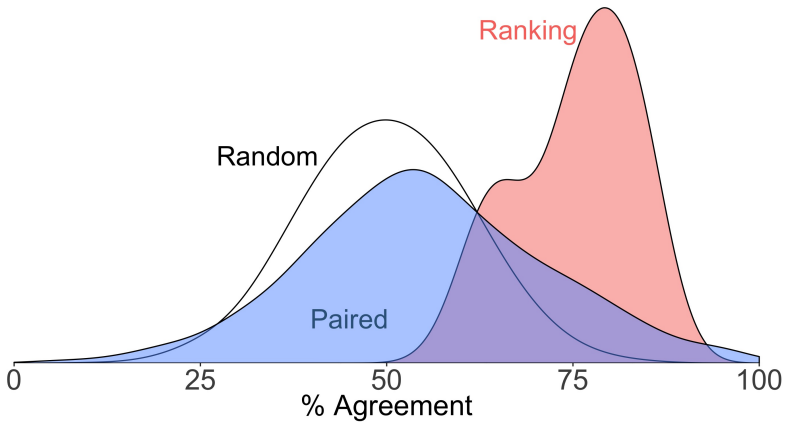


Figure 3: Intercoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

As expected when comparing coin flips, the random placebo percent agreement is centered at 50%. In contrast, the paired comparison percent agreement histogram is shifted

farther to the right than the placebo histogram, but the mean only moves to 54%, leaving the two distributions with considerable overlap. Put differently, the best we could do with the method of paired comparisons, even before the step of turning paired decisions into rank orders, is results with unacceptably low levels of intercoder reliability.

We now rule out the possibility that these results are due to different people having incompatible notions of compactness by studying intracoder reliability. To do this, we waited two weeks, randomly shuffled the order of the 20 paired comparison questions, and administered the survey to the same people. (Of the 40 people, only one mentioned, on post-survey cognitive debriefing, that "some" of the districts may have been the same as the first week.)

These results appear in Figure 4 (also as a blue histogram marked "Paired") and are more distinct from the random placebo test (in white with a black outline marked "Random") than with intercoder reliability in Figure 3, as would be expected. The mean of the paired comparison histogram is now at 65% agreement, although the overlap with the random distribution is still large. (We discuss the third histogram in the next section.)

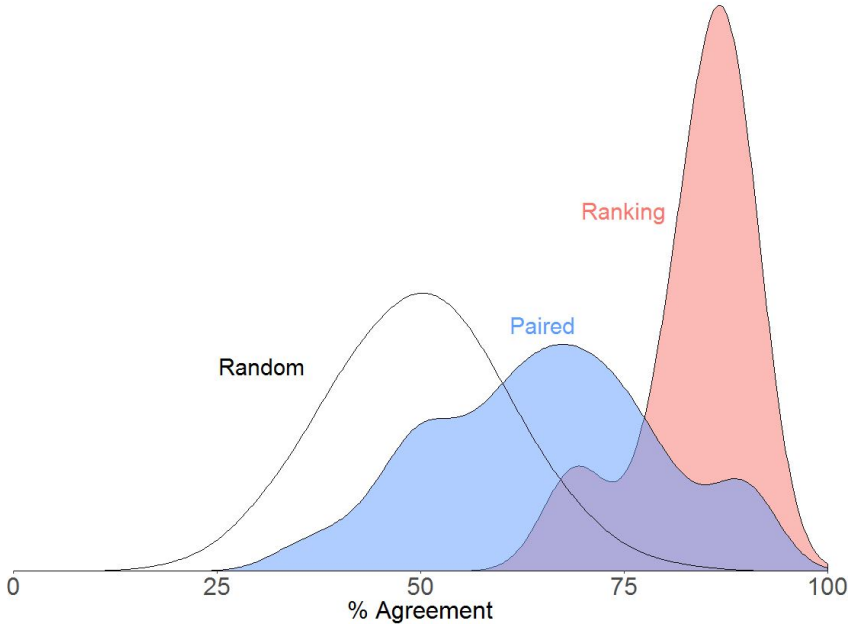


Figure 4: Intracoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

We thus conclude that these standard, best practice approaches are inadequate, at least

for our application, and turn to an alternative.

3.2 How Ranking Outranks Paired Comparisons

Why does the method of paired comparisons perform so poorly? We propose four reasons, which together leads us to a workable approach for our application, full ranking — the method which paired comparisons originally supplanted.

First, even given the math at the start of Section 3.1, the apparently obvious intuition may not necessarily follow. After all, how long would it take a person to carefully and accurately rank 20 district shapes by their degree of compactness (or 20 friends by their heights or 20 animals by their friendliness)? A lot less than 2 quintillion seconds. What the idea behind paired comparisons seems to miss is that humans are excellent at pattern recognition and seeing the big picture. Humans also intuitively apply time saving heuristics that reduce the complexity of tasks, such as in our application by grouping districts into distinct types, and considering all members of the group at once before analyzing members within the group.

Thus, in practice with full ranking, we have tried to ensure that respondents are using these skills, such as by suggesting to them that they simplify the task by working hierarchically, first grouping districts into three coarse groups, and then producing groupings within each group, and finally starting from the top and checking and adjusting each district's position within the ranking; however, we found that heuristics and intuitions are strong enough that dropping these instructions did not degrade our full ranking approach. We have also tried full ranking with districts printed on paper and arrayed on a long table, as well as via an online system where districts are dragged and dropped to their chosen location; we find no evidence that the mode of administration matters either.

Second, human respondents work better when motivated and engaged. While paired comparisons successfully avoid the risk of asking respondents questions they do not understand, it is also an unavoidably boring and tedious task, especially after the first few questions. In contrast, ranking a large set of districts is more intellectually challenging and engaging. Our own cognitive debriefing strongly supports the advantages of ranking

in this regard.[3]

Third, if it is possible for a survey respondent to rank (say) 20 districts without much trouble, then we can save considerable time by administering this one engaging survey task rather than having to ask 190 tedious paired comparisons for each respondent. Ranking would then save considerable time, expense, and respondent fatigue (Ip, Kwan, and Chiu, 2007). As a hint that this might work, Krosnick (1999) (studying rating rather than paired comparisons) finds that often "rankings give higher quality data than ratings".

And finally, the literature makes clear that compactness is a multidimensional concept (Niemi et al., 1990, p.1159). Yet, we are trying to tap into a single unidimensional concept of compactness that we hypothesize respondents, if given the choice, would select and use. In this light, the fact that Thurstone's approach enables respondents to make each paired comparison *independently* of the others allows, and may even encourage, them to use different dimensions for different comparisons. This may then result in the low levels of intercoder and intracoder reliability we have documented. In contrast, ranking has the advantage of encouraging respondents to *choose* a single dimension of compactness and to use it for all their decisions. With paired comparisons, the only way to do this would be to ask respondents to choose a single dimension explicitly and to keep that dimension in their heads while they answer 190 survey questions. Although the goal of any survey question is to be clear enough so respondents are answering the question intended by the researcher (i.e., on the dimension of interest), giving respondents multiple separate questions makes this difficult to achieve.

To test our hypothesis that ranking will work better than paired comparisons, we set it an especially difficult task. We go beyond the 3-4 items recommended in the literature, and past the 20 in our running example. Instead, we ask respondents to give a full rank order for 100 separate legislative districts by their degree of compactness.

To begin, we embed our 40 districts (which we used in 20 pairs in the experiments in Figures 3 and 4) among 60 others and ask a new set of respondents to rank all 100. To

[3]We also experimented with having two coders participate together in ranking each set of districts, on the theory that the social connections would make the task even more engaging. Our theory was supported, in that respondents spent about 30% more time together completing the task, but this engagement was unnecessary since it did not increase inter- or intracoder reliability.

compute a relative assessment of the two methods, we evaluated intercoder and intracoder reliability of the *implied* paired comparisons of how these 20 pairs were ordered by full ranking and compared them to reliability from the *actual* paired comparisons. That is, from full ranking, we record only which district in each pair of 20 comparisons is ranked higher. Then, to compute intracoder reliability, we waited two weeks, shuffled the rank ordering, and asked the same respondents to rank the same 100 districts, again only using the 20 designated pairs among these. We then computed the percent agreement over time in these implied paired comparisons exactly as we did for the actual paired comparisons. The results, which appear in the same two figures (salmon colored histogram, at the right of each figure), are far more clearly separated from the random placebo test and have much higher levels of intracoder reliability than the actual paired comparisons. For intercoder reliability, in Figure 3, we have 75% agreement on average, and for intracoder reliability, in Figure 4, we have 84% agreement on average.

Now that we have a method that bests paired comparisons for measuring compactness with respect to pairwise intracoder and intercoder reliability, we turn to evaluating full ranking on its own terms. We begin with intercoder reliability by correlating the ranks for 100 districts coded independently by (all possible) pairs of respondents. We then present in Figure 5 one scatterplot representing the pair of coders with the median correlation ($\rho = 0.77$ in the top left panel) as well as the pair with the first quartile (bottom left) and third quartile (top right). In the bottom right of the same figure (salmon colored), we also present a density estimate (using a kernel truncated at the minimum and maximum observed correlations) of all the correlations, along with a baseline density estimate of correlations among randomly generated ranks. The conclusion from this figure reveals high intercoder reliability, clearly distinguishable from chance, and with no systematic error patterns in any individual scatterplot.

We then repeat this process for intracoder reliability by correlating the ranks for each respondent with the same respondent, re-ranking the same districts, two weeks later. Figure 6 shows these results in the same format as Figure 5. As would be expected, our results here are even stronger than for intercoder reliability. The median correlation (top left) is

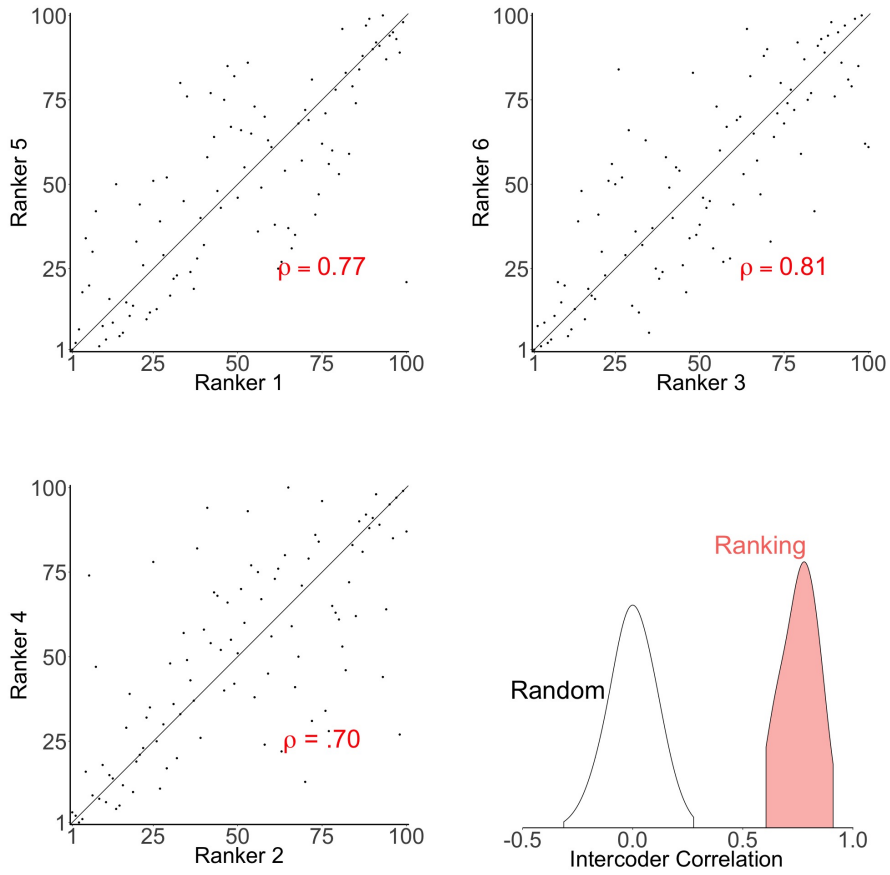


Figure 5: Intercoder Reliability for Full Ranking with 100 districts. Scatterplots are given for the median correlation (top left panel), first quartile (bottom left) and third quartile (top right). A histogram of all correlations, along with a placebo-based histogram appear at the bottom right.

$\rho = 0.9$, with not much spread around the median (see salmon colored histogram in the bottom right panel). None of the scatterplots show any systematic patterns in deviations from the $45°$ line, and all indicate high levels of intracoder reliability.

3.3 A Statistical Measurement Model

To construct our ultimate measure of compactness, we take a set of districts and elicit the views of respondents via our full ranking survey approach. We average away random error by using the first principal component of these data, preserving the ranked scale. This forms the outcome variable in our statistical model. We then code geometric features of the districts as explanatory variables, including the seven compactness indicators in

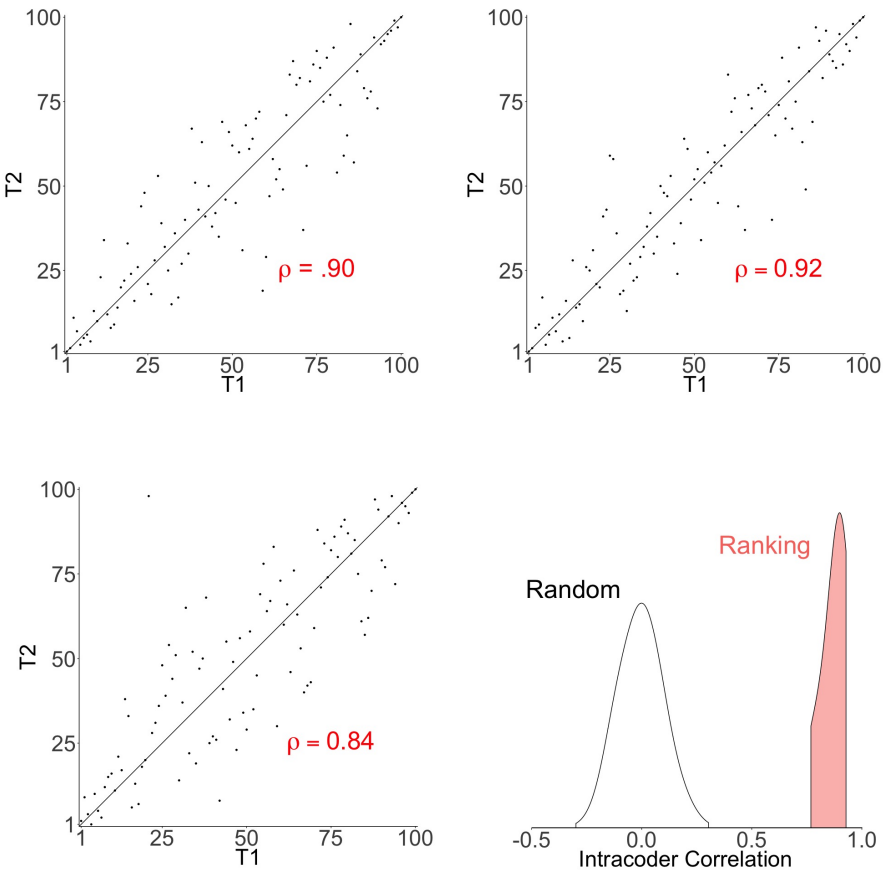


Figure 6: Intracoder Reliability for Full Ranking, following the same heuristics as Figure 5.

Table 1 and many others given in Appendix B. Finally, we train an ensemble of predictive methods with these data, consisting of least squares, AdaBoosted decision trees, support vector machines, and random forests. (All details and code are available in our replication data file which will accompany this paper.)

4 Validating Our Measures

Via cross-validation (in Section 4.1) and out-of-sample prediction in diverse populations (in Section 4.2), we now evaluate our single, unidimensional compactness measure and confirm our concomitant hypothesis that the theoretical concept we are measuring is the same one people know when they see. The data for this section come from diverse populations ranging from far away to a participant involved in decision making about legislative

redistricting.

4.1 Cross-validation

We evaluate our model here with cross-validation using 100 districts each. To do this, we use six groups of survey respondents, potentially making it harder for our model by mixing size of group, mode of administration, and type of respondent: (1) two pairs of undergraduates (the two within each pair working together) and one pair of graduate students; (2) one pair of undergraduates, one individual undergraduate, and one pair of graduate students; (3) 5 individual undergraduates, 5 pairs of undergraduates, and 16 Mechanical Turk workers; (4) 5 pairs and five individual undergraduates; (5) 8 undergraduates; (6) 8 undergraduates. (We found ex post that respondents gave similar rankings regardless of whether they worked alone or in pairs. Similarly, Mechanical Turk workers, undergraduates, and graduate students gave similar rankings on the same sets of districts.)

We then trained our model on groups 1–5 of respondents taken together, and predicted the remaining "test set" of respondents in group 6; we repeated this six times in total, with each group taking its turn as the test set and the remaining groups as the training set. The prediction from this model uses all information from the training set but only the district geometry (i.e., no survey information) from the test set. Figure 7 evaluates the performance of this procedure by providing six scatterplots corresponding to each of our training set-based predictions (horizontally) by the true test set values (vertically). As is evident, these cross-validation results indicate very high predictive accuracy. Correlations between predictions and test set values range from 0.91 to 0.96, with no noticeable systematic error patterns in any graph.

4.2 Predictive Validation in Diverse Populations

The statistical model in Section 3.3 is designed to predict human judgment about the compactness of any set of districts, given only the geometric shapes of the districts. Our model will make predictions for any legislative district shape, including new districts and those which do not appear in a training set.

Our hypothesis is that any informed human being will judge the compactness of a set

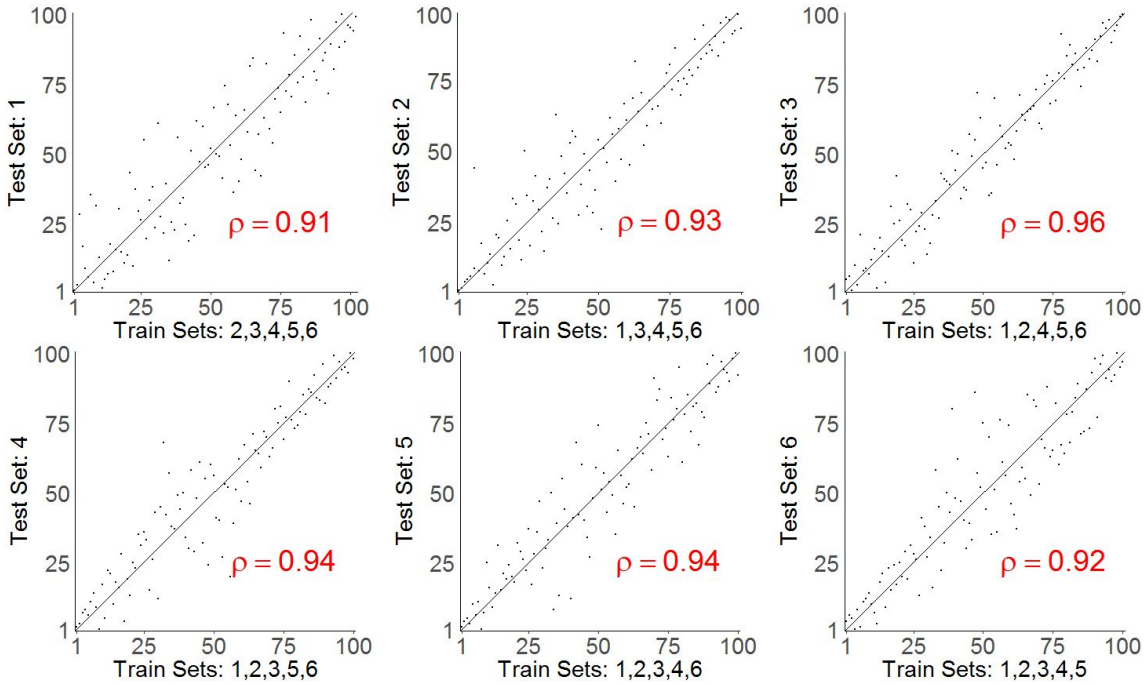


Figure 7: Cross-Validation of Model Predictions

of districts in almost the same way, thus admitting to high levels of statistical reliability. We now test this hypothesis by asking a wide range of groups to evaluate the compactness of different sets of legislative districts. We can order the different groups of people by the degree to which they have been involved in, or knowledgeable about, legislative redistricting. From less to more involved, these include (1) Mechanical Turk workers, who received small monetary payments, (2) undergraduates, (3) political science PhD students, (4) law students, (5) law faculty, (6) redistricting consultants and expert witnesses, (7) lawyers involved in legislative redistricting cases, and (8) public officials who have some responsibility for redistricting and (9) judges from diverse jurisdictions and positions who decide redistricting cases.

We promised our respondents complete confidentiality, including their responses and the fact of their participation. This was most obviously a concern in recruiting judges, who decide redistricting cases, and other public officials, who have decision making authority in or substantial influence on the process. It turned out to be of no less a concern for some lawyers who try redistricting cases, and some consultants and expert witnesses who are held to account for their previous statements and opinions. For these reasons, we are

not able to make these data or all information about them available publicly, although we do make available the software we designed to let respondents sort districts online and all our specific experimental protocols. All these steps were approved by our university Institutional Review Board.

In this experiment, we asked each respondent to rank order twenty legislative districts by their degree of compactness and represent the degree of predictive accuracy by a simple correlation with our predictions. We portray our results in Figure 8 with a histogram for each of nine categories of people. As a baseline, we present a density estimate (in blue) of the percent agreement among random rankings, which is of course centered at zero, and the variance of which conveys uncertainty given $n = 20$ districts. The (salmon-colored) histogram is for Mechanical Turk workers. The remaining histograms of correlations appear in white, with black outlines. We do not distinguish among these for a further level of confidentiality, but they all lead to the same conclusion of very high levels of predictive accuracy.

We found no statistically significant differences between the size of the correlations among different groups of respondents. The main predictor of the strength of the correlations was the time spent on the task, with longer times yielding higher correlations. This accounts for the larger variance of Mechanical Turk workers, as they are paid by the completed task regardless of how long they spend. (We did not pay any of the other groups to participate, except for some undergraduates.) After initial experimentation, we changed the definition of a "completed" task for all our groups by requiring at least ten district reorderings (operationally, the submit button on our online application was grayed out until ten districts were dragged and dropped to a different order; we then subtly changed the button afterwards to allow hitting submit but not so obviously that we started to encourage stopping at ten).

5 Concluding Remarks

Approaches developed here for measuring an ill-defined concept that you know only when you see may be applicable to other difficult-to-define concepts. These include measure-

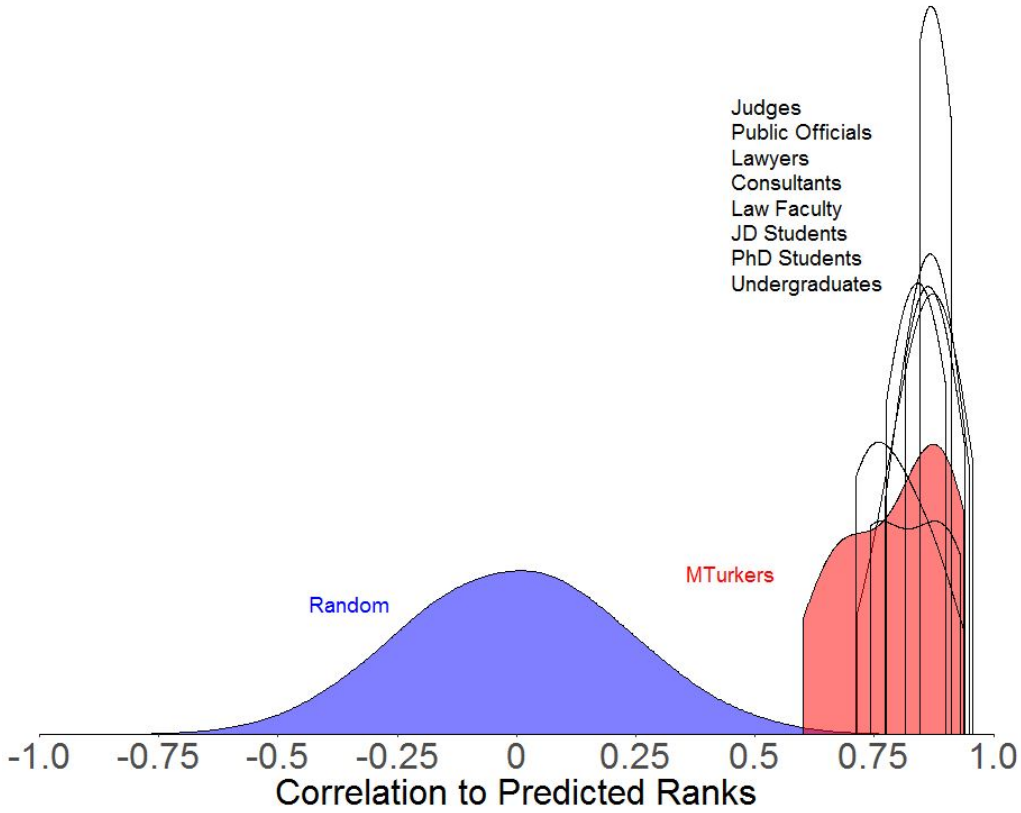


Figure 8: Histograms (via density estimates) of correlations between predictions from our model and answers to survey questions from nine different groups of respondents.

ment by full ranking rather than paired comparisons, which saves time and turns out, in our application, to have much higher levels of intra- and intercoder reliability; the incorporation in a model rather than replacement of most existing measures and approaches; and formalization into a statistical model of an approach that predicts the views of a wide range of different types of people.

The key aspect of our approach to measuring compactness is defining the concept of interest separately from the measure used to estimate it, so that our measure becomes vulnerable to being proven wrong and, as a result, our approach can improve over time. Indeed, we encourage others to take up this challenge and improve on the methods we propose, and develop statistical methods that outperform ours; this should now be possible, as performance standards now exist. New features measuring compactness can also be included in our approach as covariates in our statistical model, which may well be improved. Other criteria for redistricting can also be incorporated (see the introduction to

Section 2).

The dimension of compactness we measure here is on an absolute, rather than relative, scale and, as such, is designed to be affected by fixed features of the geography of states, such as rivers, highways, and others. For example, a district drawn on a highly irregular coastline may have a maximum possible level of compactness that is below that in the Iowa farmlands. This is not an issue to correct, but it is a characteristics of absolute measures to be aware of in using the data accompanying this article to compare across states or other geographic areas. Put differently, the canvas affects the artwork — by design. Finally, our measure also considers one district at a time; because a full redistricting plan affects more than one district and often an entire state one may need to average our results over districts within a plan may for certain purposes.

Appendix A Geometric Features of Legislative Districts

We define many useful existing compactness measures, and other geometric features of legislative districts we introduce. We use all of these quantities in Section 3.3. We begin with basic notation used in many of the measures and then turn to the measures.

Notation Denote a generic legislative district as D, and define it as a non-self-intersecting closed polygon with n vertices, each labeled (x_i, y_i) and numbered i in clockwise order (for $i = 1, \ldots, n$). We choose an arbitrary starting vertex for label $i = 1$ and (using clock or modular algebra) define $i = n + 1 = 1$. The length of the line segment from vertex i to $i + 1$ is then $L_i = ||(x_i, y_i), (x_{i+1}, y_{i+1})||$ where $||(a, b), (c, d)|| = \sqrt{(a - c)^2 + (b - d)^2}$. Denote the set of all horizontal vertex coordinates as $X = \{x_i : i = 1, \ldots, n\}$, vertical vertex coordinates as $Y = \{y_i : i = 1, \ldots, n\}$, and line lengths as $L = \{L_i : i = 1, \ldots, n\}$.

Then the area of D is $A(D) = \frac{1}{2} \sum_{i=1}^{n} (x_i y_{i+1} - x_{i+1} y_i)$ and perimeter is $P(D) = \sum_{i=1}^{n} L_i$. Occasionally, as in the case of islands, D is composed of multiple polygons. In these cases, $A(D)$ and $P(D)$ are the sums of the areas and perimeters of all the polygons in D, and all subsequent notation refers to all vertices in all polygons taken together.

Denote the district centroid as $C(D)$, defined by a vertex with coordinates $C(D)_x = \frac{1}{6A(D)} \sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$ and $C(D)_y = \frac{1}{6A(D)} \sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$, and radii $r_i = ||[C(D)_x, C(D)_y], (x_i, y_i)||$. Then denote as Circle$(D)$ the minimum bounding circle (Nielsen and Nock, 2008) and as Hull(D) the minimum bounding convex hull (King and Zeng, 2006; Kong, Everett, and Toussaint, 1990). Finally, for set S with cardinality $\#S$, denote the mean over i of function $g(i)$ as $\text{mean}_{i \in S}[g(i)] = \frac{1}{\#S} \sum_{i=1}^{\#S} g(i)$, the variance as $\text{var}_{i \in S}[g(i)] = \text{mean}_{i \in S}\left[\{g(i) - \text{mean}_{j \in S}[g(j)]\}^2\right]$, and the mean absolute deviation as $\text{mad}[g(i)] = \frac{1}{\#S} \sum_{i=1}^{\#S} |g(i) - \text{mean}[g(i)]|$.

Feature List The perimeter of the minimum bounding circle is PC $= P($Circle$(D))$ and minimum bounding convex hull is PCH $= P($Hull$(D))$. The area of each is the AC $=$ $A($Circle$(D))$ and ACH $= A($Circle$(D))$. The number of polygons is PARTS and vertices, or sides, is SIDES $= n$ (Timmerman, 100 N.Y.S. 57, 51 Misc. Rep. 192 (N.Y. Sup. 1906)). We then have REOCK $= A(D)/A($Circle$(D))$; GROFMAN $= P(D)/\sqrt{(A(D))}$; HULL RATIO $= A(D)/A($Hull$(D))$; SCHWARTZBERG $= P(D)/(2\pi\sqrt{A(D)/\pi})$ and the mathematically related POLSBYPOPPER $= 4\pi A(D)/P(D)^2$; the variation in the coordinates of the x-axis, XVAR $= \text{var}_{i \in X}[x_i]$, and y-axis, YVAR $= \text{var}_{i \in Y}[y_i]$; the average, AVGLL $= P(D)/n = \text{mean}_{i \in L} L_i$, and variance, VARLL $= \text{var}[L_i]$, of the polygon line segment lengths; LENGTH-WIDTH RATIO $= [\max_i(x_i) - \min_i(x_i)]/[\max_i(y_i) - \min_i(y_i)]$; (our simplified expression of modified) BOYCE-CLARK $= 1 - \frac{1}{2\,\text{mean}_i[r_i]} \text{mad}_i[r_i]$ (MacEachren, 1985, p.56); POINTS $= n$ for the district polygon defined by the official US Census shapefile; using the Harris Corner Detector algorithm (C. Harris and Stephens, 1988), we also have the number of significant "corners" (i.e., vertices), CORNERS, and the variance in the x-coordinate XVARCORNERS and y-coordinate YVARCORNERS of each corner. The EQUAL-LAND-AREA CIRCLE, defines noncompactness as a threshold occurring when a circle with origin at $C(D)$ and area $A(D)$, i.e. with radius $\sqrt{A(D)/\pi}$, captures less than half the area of D (Angel and Parent, 2011, p.93). Finally, we have Y-SYMMETRY, the area of district D overlapping with the reflection of D around a vertical line going through $C(D)$, divided by $A(D)$, and X-SYMMETRY, which is the same except for reflection of D around a horizontal line going through $C(D)$.

Appendix B Compactness Data and Software

We offer additional details here of how we collected the data for our experiments and then outline a large collection of data we make available as a companion to this paper on the compactness of numerous state legislative and congressional districts.

Data Collection To construct training and test sets for our various experiments, we use a set of 18,215 district shapes, including all congressional districts 1823–2013 and the last two cycles of state legislative districts. We obtained the shape files and other geographic data for congressional districts from Lewis et al. (2013) and state legislative districts from McMaster, Lindberg, and Van Riper (2003).

To avoid focusing on imperceptible differences among districts, we begin with a rough preliminary compactness ranking by ordering these districts based on an average of each district's Reock, Polsby-Popper, and Convex Hull scores. We create six groups of districts using systematic random sampling — to ensure a spread over the entire range of compactness — using a random start without replacement across groups — to avoid overlap among the groups. For the cross-validation in Section 4.1, we drew 100 districts. For our out-of-sample validations in Section 4.2, we collected 20 districts (to accommodate respondent time constraints).

We tested a variety of different instructions to our respondents. Here is a simple version we used for our online administration for full ranking. [We found the sentences in square brackets below useful for respondents, such as some from Mechanical Turk, who are not as familiar with the concept of compactness or the idea of legislative districts. Experiments we conducted among those familiar indicate that these passages do not affect the resulting rankings.]

> The law requires that legislative districts for the US congress and many state legislatures be "compact". The law does not say exactly what district compactness is, but generally, people think they know it when they see it. [One dictionary definition of compactness is "joined or packed together closely and firmly united; dense; arranged efficiently within a relatively small

space." Some characteristics of districts people view as noncompact are wiggles, arms, noncontiguous segments, river-like features, or being much longer than wide. Compact districts look more densely packed, like rectangles, circles, or hexagons.]

Here's your task: Below is a group of legislative districts, randomly ordered. Order the districts from most compact (at the top left) to least compact (at the bottom right) according to your own best judgement, by dragging and dropping. [We have many individuals performing this task, and the more your ranks are similar to others', the better you will have done.]

For paired comparisons, we changed the second paragraph to ask respondents to choose the more compact district of the two presented to them.

Our undergraduate respondents ranked 100 districts in a conference room with a long set of connected tables. We printed out pictures of each district, along with an identifying number, on a card measuring 4.25×5.5" (one quarter of a standard 8×11.5" paper). We asked each respondent to order the cards from most to least compact and then to enter the final results in a spreadsheet. As described in Section 3.2, we experimented with different sets of instructions, and with respondents working alone and in pairs, but we found no difference in intercoder or intracoder reliability as a result.

We asked the Mechanical Turk workers who ranked 100 districts to print out twenty-five sheets of paper with four districts each, and then to cut each in quarters and to follow the same instructions we gave our undergraduates. We asked for and received cell phone photos from the Turkers at each stage, to help ensure the task was completed as designed.

The undergraduates and Mechanical Turk respondents each took about 45–90 minutes to rank 100 districts. In order to reach a larger number of respondents, and especially to avoid charges of diverting public officials from performing their duties, we conducted our out-of-sample predictions with 20 districts. We chose this number by repeated experimentation with undergraduates, until we were able to get the time necessary to complete the task to under ten minutes. Most took 7–10 minutes.

Data Availability and Future Research For each of the 18,215 congressional and state legislative districts in our collection, we compute the degree of compactness by applying the model in Section 3.3. We make all these data publicly available as a companion to this paper, as well as software that implements this model that others can use to estimate compactness in new districts. We think further analyses of these data may shed light on many of the venerable questions scholars have asked about compactness and its relationship to other variables, such as the balance between the parties, the existence of partisan gerrymandering, and the extent of racial fairness.

The data also seem to suggest many other important questions worthy of further analysis. As one example, we examine results for four states frequently mentioned in the press as examples political gerrymandering. In Maryland's 2016 congressional elections, 37% of the state's two-party vote share went to Republicans. Despite this, Republicans managed to win only one congressional seat in Maryland—leaving the state with a 7–1 delegation in favor of the Democrats. In Pennsylvania, despite winning approximately 46% of the two-party vote share in 2016, Democrats won only 5 of 18 congressional districts. In North Carolina, Democrats won 47% of the two-party vote share in 2016, but hold only 3 of 13 congressional seats. Similarly, in Ohio, the Democratic share of the two-party vote was 42% whereas Democrats hold only 3 of Ohio's 16 seats. A full partisan symmetry analysis would need to be conducted to evaluate whether these results were fair to the political parties (Gelman and King, 1994a), but this prima facie evidence certainly suggests further analysis is worthwhile.

Our model predicts the rank a district would be given by a human coder (given only the shape of the district), with rank 1 being most compact and higher numbers indicating higher levels of noncompactness. We thus compute this *noncompactness* measure, using our methods, for each congressional district in each of these four states, for every new redistricting since 1893. We then average all the districts within each state and, in Figure 9, plot the averages over time.

Interestingly, noncompactness dramatically increases in Ohio and Pennsylvania beginning in the mid-1960s, shortly after Baker v. Carr (1962) mandated redistricting to

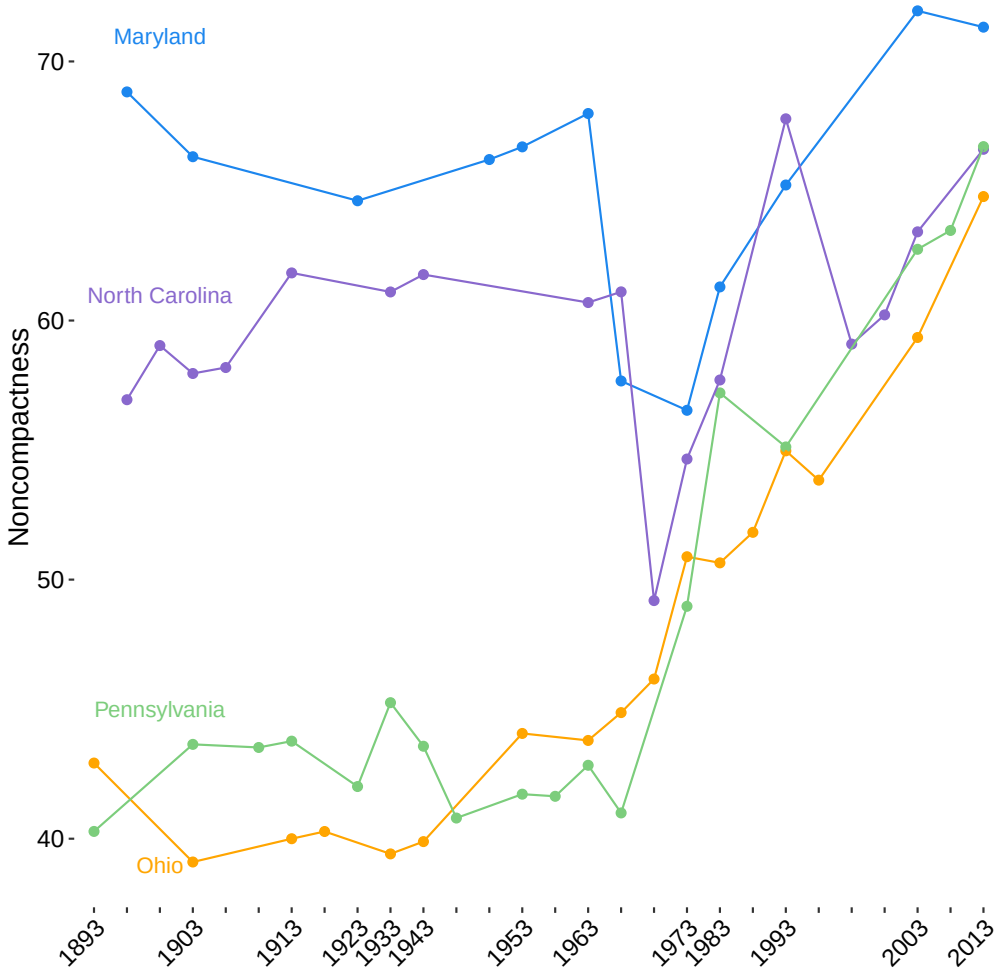


Figure 9: Time series plots of average district compactness in congressional districts for four states often claimed in the media to be political gerrymanders.

achieve equal district populations. Maryland and North Carolina, in contrast, show no such increase. Is this because these states had high noncompactness levels to begin with? Could noncompactness have been at an effective maximum? Did redistricters from the majority parties in Ohio and Pennsylvania take advantage in ways those in North Carolina and Maryland did not? Did the progress (or overreaching) on behalf of minorities in two of the states take a different path than in the other two? Or might the differences be due to other factors, such as local political subdivisions, communities of interest, or natural features of the states being taken into account in districting in different ways? We encourage future researchers to delve into these and the numerous other questions these

data suggest.

References

Altman, Micah (1998): "Modeling the effect of mandatory district compactness on partisan gerrymanders". In: *Political Geography*, no. 8, vol. 17, pp. 989–1012.

Altman, Micah and Michael P McDonald (2012): "Redistricting principles for the twenty-first century". In: *Case W. Res. L. Rev.* Vol. 62, p. 1179.

Angel, Shlomo and Jason Parent (2011): "Non-compactness as voter exchange: Towards a constitutional cure for gerrymandering". In: *Northwestern Interdisciplary Law Review*, vol. 4, p. 89.

Ansolabehere, Stephen and Maxwell Palmer (2016): "A Two Hundred-Year Statistical History of the Gerrymander". In: *Ohio St. LJ*, vol. 77, pp. 741–867.

Barabas, Jason and Jennifer Jerit (2004): "Redistricting principles and racial representation". In: *State Politics & Policy Quarterly*, no. 4, vol. 4, pp. 415–435.

Boyce, Ronald R and William AV Clark (1964): "The concept of shape in geography". In: *Geographical Review*, no. 4, vol. 54, pp. 561–572.

Cain, Bruce (1984): *The Reapportionment Puzzle*. Berkeley: University of California Press.

Chen, Jowei, Jonathan Rodden, et al. (2013): "Unintentional gerrymandering: Political geography and electoral bias in legislatures". In: *Quarterly Journal of Political Science*, no. 3, vol. 8, pp. 239–269.

Chou, Christine et al. (2014): "On empirical validation of compactness measures for electoral redistricting and its significance for application of models in the social sciences". In: *Social Science Computer Review*, no. 4, vol. 32, pp. 534–543.

Converse, Jean M. and Stanley Presser (1986): *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage Publications.

Cox, EP (1927): "A method of assigning numerical and percentage values to the degree of roundness of sand grains". In: *Journal of Paleontology*, no. 3, vol. 1, pp. 179–183.

David, H. A. (1988): *The Method of Paired Comparisons, 2nd ed.* London: Oxford University Press.

Fechner, Gustav (1966): "Elements of psychophysics. Vol. I. [Originally published 1860]". In:

Fryer Jr, Roland G and Richard Holden (2011): "Measuring the compactness of political districting plans". In: *The Journal of Law and Economics*, no. 3, vol. 54, pp. 493–535.

Gelman, Andrew and Gary King (May 1994a): "A Unified Method of Evaluating Electoral Systems and Redistricting Plans". In: *American Journal of Political Science*, no. 2, vol. 38, pp. 514–554. URL: j.mp/unifiedEc.

— (Sept. 1994b): "Enhancing Democracy Through Legislative Redistricting". In: *American Political Science Review*, no. 3, vol. 88, pp. 541–559. URL: j.mp/redenh.

Gideon, Lior (2012): "The art of question phrasing". In: *Handbook of survey methodology for the social sciences*. Springer, pp. 91–107.

Grofman, Bernard and Gary King (Jan. 2007): "The Future of Partisan Symmetry as a Judicial Test for Partisan Gerrymandering after LULAC v. Perry". In: *Election Law Journal*, no. 1, vol. 6. http://gking.harvard.edu/files/abs/jp-abs.shtml, pp. 2–35.

Harris, Chris and Mike Stephens (1988): "A combined corner and edge detector." In: *Alvey vision conference*. Vol. 15. 50. Citeseer, pp. 10–5244.

Harris, Curtis C (1964): "A scientific method of districting". In: *Behavioral Science*, no. 3, vol. 9, pp. 219–225.

Hsu, Mei-Ling and Arthur Howard Robinson (1970): *The fidelity of isopleth maps: An experimental study*. U of Minnesota Press.

Ip, WC, YK Kwan, and LL Chiu (2007): "Modification and simplification of thurstone scaling method, and its demonstration with a crime seriousness assessment". In: *Social indicators research*, no. 3, vol. 82, pp. 433–442.

Kahneman, Daniel (2011): *Thinking, fast and slow*. Macmillan.

Kaiser, Henry F (1966): "An objective method for establishing legislative districts". In: *Midwest Journal of Political Science*, no. 2, vol. 10, pp. 200–213.

King, Gary, John Bruce, and Andrew Gelman (1996): "Racial Fairness in Legislative Redistricting". In: ed. by ed. Paul E. Peterson. Princeton University Press. URL: j.mp/Fairrace.

King, Gary and Langche Zeng (2006): "The Dangers of Extreme Counterfactuals". In: *Political Analysis*, no. 2, vol. 14, pp. 131–159. URL: j.mp/dangerEC.

Kong, Xianshu, Hazel Everett, and Godfried Toussaint (1990): "The Graham scan triangulates simple polygons". In: *Pattern Recognition Letters*, no. 11, vol. 11, pp. 713–716.

Krosnick, Jon A. (1999): "Survey Research". In: *Annual Review of Psychology*, no. 1, vol. 50, pp. 537–567.

Lee, David R and G Thomas Sallee (1970): "A method of measuring shape". In: *Geographical Review*, pp. 555–563.

Lewis, Jeffrey B et al. (2013): "Digital boundary definitions of united states congressional districts, 1789–2012". In: *Data file and code book*. URL: cdmaps.polisci.ucla.edu.

MacEachren, Alan M (1985): "Compactness of geographic shape: Comparison and evaluation of measures". In: *Geografiska Annaler. Series B. Human Geography*, pp. 53–67.

Maurer, Daphne, Richard Le Grand, and Catherine J Mondloch (2002): "The many faces of configural processing". In: *Trends in cognitive sciences*, no. 6, vol. 6, pp. 255–260.

McMaster, Robert B, Mark Lindberg, and David Van Riper (2003): "The national historical geographic information system (NHGIS), Version 11.0". In: *Proceedings 21st International Cartographic Conference*, pp. 821–828.

Mitliagkas, Ioannis et al. (2011): "User rankings from comparisons: Learning permutations in high dimensions". In: *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, pp. 1143–1150.

Nielsen, Frank and Richard Nock (2008): "On the smallest enclosing information disk". In: *Information Processing Letters*, no. 3, vol. 105, pp. 93–97.

Niemi, Richard G et al. (1990): "Measuring compactness and the role of a compactness standard in a test for partisan and racial gerrymandering". In: *The Journal of Politics*, no. 4, vol. 52, pp. 1155–1181.

Pildes, Richard H and Richard G Niemi (1993): “Expressive Harms, Bizarre Districts, and Voting Rights: Evaluating Election-District Appearances After Shaw v. Reno”. In: *Michigan Law Review*, no. 3, vol. 92, pp. 483–587.

Polsby, Daniel D and Robert D Popper (1991): “The third criterion: Compactness as a procedural safeguard against partisan gerrymandering”. In: *Yale Law & Policy Review*, no. 2, vol. 9, pp. 301–353.

Reock, Ernest C (1961): “A note: Measuring compactness as a requirement of legislative apportionment”. In: *Midwest Journal of Political Science*, no. 1, vol. 5, pp. 70–74.

Schwartzberg, Joseph E (1965): “Reapportionment, gerrymanders, and the notion of compactness”. In: *Minn. L. Rev.* Vol. 50, p. 443.

Stoddart, David R (1965): “The shape of atolls”. In: *Marine Geology*, no. 5, vol. 3, pp. 369–383.

Thurstone, Louis L (1927): “The method of paired comparisons for social values.” In: *The Journal of Abnormal and Social Psychology*, no. 4, vol. 21, p. 384.

Tisdale, Elkanah (1812): “The Gerry-Mander”. In: *Boston Gazette*.

Tsai, Yu-Hsin (2005): “Quantifying urban form: compactness versus sprawl”. In: *Urban studies*, no. 1, vol. 42, pp. 141–161.

Young, H Peyton (1988): “Measuring the compactness of legislative districts”. In: *Legislative Studies Quarterly*, pp. 105–115.