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

Aaron Kaufman† Gary King‡ Mayya Komisarchik§

February 8, 2018

Abstract

To prevent gerrymandering, and to impose a form of democratic representation, many state constitutions and judicial opinions require US legislative districts to be "compact." Yet, the law offers few precise definitions other than "you know it when you see it," which effectively implies a common understanding of the concept. In contrast, academics have shown that the concept has multiple theoretical dimensions and have generated large numbers of conflicting empirical measures. This has proved extremely challenging for courts tasked with adjudicating compactness. We hypothesize that both are correct — that compactness is complex and multidimensional, but a common understanding exists across people. We develop a survey design to elicit this understanding, without bias in favor of one's own political views, and with high levels of reliability (in data where the standard paired comparisons 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 96 judges, justices, and public officials responsible for redistricting (and 102 redistricting consultants, expert witnesses, law professors, law students, graduate students, undergraduates, and Mechanical Turk workers). We also offer data on compactness from our validated measure for 18,215 state legislative and congressional districts, as well as software to compute this measure from any district. We then 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 j.mp/Compactness. Our thanks to Steve Ansolabehere, Ryan Enos, Dan Gilbert, Jim Griener, Bernie Grofman, Andrew Ho, Dan Ho, James Honaker, Justin Levitt, Luke Miratrix, Max Palmer, Stephen Pettigrew, Jamie Saxon, Steve Shavell, Anton Strezhnev, Wendy Tam, Rocio Titiunik, Larry Tribe, Robert Ward, participants in "A Causal Lab", and the audiences at the Society for Political Methodology Meetings, the Harvard Applied Statistics Workshop, and the ICPSR Summer Program 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 treated in the law as an important legal bulwark against gerrymandering. The Apportionment Act of 1901, many court decisions, and 18 state constitutions require compactness for U.S. House districts, and 37 states require their legislative districts to be compact (see j.mp/aRED). Compactness is also required in federal law 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)).[1]

Compactness is also important for the academic literature, where scholars seek to help the redistricting and litigation processes, and also to study age old political science questions such as the causes, consequences, and normative implications of compact districts over American history (**ForPla05**; **AnsSny12**; **AnsPal16**). Compactness intuitively refers to both how close a legislative district's boundaries are to its geographic center and how "regular" in shape a district appears to be.[2] But upon deeper study, scholars have shown that in fact compactness is a complicated multidimensional concept and have offered almost 100 measures of different features of it (**NieGroCar90**).[3]

While many state constitutions explicitly require compactness, the vast majority provide no definition or measure for how to detect violations of the standard. For example, the Constitution of Illinois says only "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 con-

[1] Claims about most other types of unfairness in redistricting all also seem to depend on a legal finding of noncompactness (*Davis v. Bandemer*, 478 U.S. 165; Justice Powell in *Vieth v. Jubilerer*, 541 U.S. 267 (2004) 176-177; *Kirkpatrick v. Preisler*, supra, at 394 U. S. 526, 538).

[2] For example, the Colorado State Constitution suggests that "compactness, as used in the constitutional sense relating to reapportionment, concerns a geographic area whose boundaries are as nearly equidistant as possible from the geographic center of the area being considered, allowing for variance caused by population density and distribution" (Co. Const. art. V, pt XLVII). The state of Idaho mandates "to the maximum extent possible, the [districting] plan should avoid drawing districts that are oddly shaped."

[3] The empirical claim sometimes implied in the law, that compactness requirements constrain racial or partisan gerrymandering, is the subject of active research program (**AltMcD12**; **BarJer04**; **CheRod13**), 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 (**Cain84**; **PilNie93**). But regardless of the outcome of these important debates, the degree of compactness of legislative districts will always have an essential role in defining the nature of representation and electoral competition in modern democracies.

tiguous to the extent practicable."[4]

The federal Courts have been similarly vague. They have acknowledged both the multitude of possible measures for compactness, and the fact that they often produce different conclusions.[5] Except in rare cases, they have not provided guidance on particular measures or seen the need for them. For example, Justice Souter stated that "it is not necessary now to say exactly how a district court would balance a good showing on one of these indices against a poor showing on another, for that sort of detail is best worked out case by case" (*Vieth v. Jubelirer*, 541 U.S. 267 (2004); Souter dissenting). And most famously, a Supreme Court opinion indicated "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)). Here, the Court at once laments the absence of a single quantitative standard while also implying that the concept is clear enough that all reasonable observers should understand it in the same objective way.

Consistently invoking the idea of "compactness" without a clear definition or required measure suggests two conclusions about the law. First, the law seems to imply that "compactness" is a single, coherent, and agreed upon concept, discernable simply by examining a district map. After all, how could the courts expect legislators to draw districts that comply with "compactness" without a shared understanding of what it means? And second, this lack of precision in the law has, simultaneously, enabled redistricters and litigants battling over legislative maps in specific cases to cherry pick their own self-serving def-

[4]Some states have passed laws highlighting certain features of compactness that may help with intuition but neither precision nor application. For example, Virginia Senate Joint Resolution 224 (1/14/2015, Article II, Section 6(5)) 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) mention some precise measures but not how to use this information.

[5]"Indeed," writes Justice Souter, dissenting in *Vieth v. Jubelirer*, "although compactness is at first blush the least likely of these [traditional redistricting] principles to yield precision, it can be measured quantitatively in terms of dispersion, perimeter, and population ratios, and the development of standards would thus be possible."

initions and measures to suit their claims. As such, the courts and policy makers do not benefit as much as they could from quantitative measures offered by social scientists.

We attempt to span this divide between the seemingly universal understanding of compactness proposed in or needed for the application of the law, and the theoretical complexity and multidimensionality revealed in the social science literature. We do this by inferring, measuring, and validating the single underlying dimension of compactness that practitioners may need to apply the law, and we find that people of all types seem to agree upon it. In other words, since compactness in the law is, for all practical purposes, defined by the judgment of human observers — including redistricters, experts, consultants, lawyers, judges, public officials, and ordinary citizens — the claim of an objective standard, measured on a single dimension, can only be supported if most educated people evaluated a district's compactness in the same way. We provide this objective measure and show that these and other groups of observers all view compactness in this way. This new dimension is not the average (or principal component) of existing measures but a new quantitative construction that accurately and reliably predicts human judgment.

In four sections, we proceed by *conceptualizing*, *measuring*, *validating*, and *interpreting* our derived dimension of compactness. Section 2 inductively defines the underlying dimension by building on the encyclopedia of existing diverse measures, adding new ones that show how humans perceive objects like district shapes, and providing intuition about the commonly perceived dimension we seek to measure. Section 3 then develops a way to measure this concept by eliciting views of the compactness of specific districts from respondents using a novel survey approach to rank order districts according to their compactness. We are 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 a unidimensionality hypothesis (and suggests 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 the districts' shapes.

Our 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 and 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. Application of this same principle also enables us to provide the first uncertainty estimates for a measure of compactness offered in the literature (see Appendix C). Section 5 then offers interpretations of the resulting measure, and Section 6 concludes.

2 Conceptualizing

We now attempt to inductively characterize the concept of compactness that most laws, constitutions, judicial opinions, and participants in redistricting at least implicitly 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 this measurement. This is the most common basis for a compactness definition, dating well before the famous "Gerry-Mander" cartoon (**Tisdale1812**), but not the only one possible. In other words, our goal is to define and estimate *absolute* compactness based on district shape alone. 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 would be a key component in that calculation.

We attempt to characterize the compactness of each district separately. Although changing the boundaries of one district obviously affects neighboring districts, separate

measurement follows major redistricting litigation, which typically evaluates the compactness of districts individually or in a small group rather than for an entire state redistricting plan all at once (e.g., *Shaw v. Reno*, 509 U.S. 630 (1993), pp. 637, 647, 656). This strategy is especially useful for the most fine grained scholarly research on the causes and consequences of compactness.

Many aspects of the overall methodology we develop here can also be applied to some other redistricting criteria, when additional data are available (or to concepts unrelated to redistricting 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 (**FryHol11**); 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 (**GroKin07**; **GelKin94b**), and racial fairness (**KinBruGel96**). Redistricting also influences more personalistic factors common 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

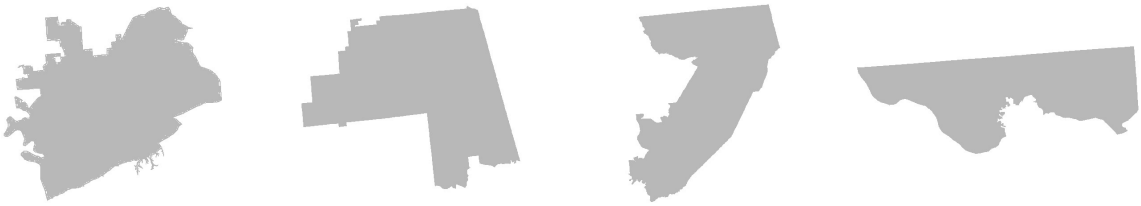
Numerous specific compactness measures have been proposed in the academic literature, each one fitting different qualitative conceptual definitions and intuitions for certain geographical configurations and violating it for others (**Altman98**; **Stoddard65**; **NieGroCar90**; **Young88**). These measures are based on geometric concepts such as perimeters, areas, vertices, and centroids, often in comparison with some pure form geo-

metric 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 (**Harris64** 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 (**Reock61**); *Polsby-Popper*, the ratio of the area of the district to the area of the circle with the same perimeter as the district (**PolPop91**; **Schwartzberg65**); and (modified) *Boyce-Clark*, the (normalized) mean absolute deviation in the radial lines from the centroid of the district to its vertices (**BoyCla64**; **Kaiser66**; **MacEachren85**). 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 (**AnsPal16**; **NieGroCar90**).

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 (we discuss X-Symmetry and significant corners in Section 2.2). 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 D), 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
Reock	1	2	3	4
Convex Hull	4	3	2	1
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, but in the absence of political or personal biases. We continue to proceed inductively, with Section 3 devoted to measuring this concept. We do this in three ways, followed by a characterization of the dimension of interest.

First, our goal is to elicit views about compactness, but without the biases psychologists have long demonstrated skew human judgments in the direction of our own political and other preferences. Although this may be the goal of lawyers advocating on behalf of their clients, research has shown that subject matter experts are as vulnerable to bias as

nonexperts, and more overconfident in the belief that they can avoid it. The only reliable solution has been to remove even the possibility of bias by instituting formal procedures (such as double blind experiments). (**Kahneman11**). We thus avoid a key form of bias here by eliciting views about compactness, without revealing to respondents how their decisions in any one situation might benefit one political party or another. This is a critical point: Because individual judges, advocates, redistricters, and experts do not have access to the mental processes in their own thinking that would enable them to evaluate and avoid these biases (**WilBre94**), they would also be unable to come to the same judgment as our measure in the context of a real redistricting contest by merely looking at a district shape.

Second, note that all existing compactness measures are *rotationally invariant*, meaning that if we rotate a district, say 45 degrees, a measure will have the same value. Although this is a reasonable normative standard from some perspectives — and we discuss below how to easily adjust our methods to impose this restriction if desired — human beings (including judges) do not evaluate districts in this way. In fact, human perception is famously sensitive to the rotation of objects: even familiar faces can become unrecognizable when viewed upside down (**MauLeGMon02**). Our own experimentation suggests that people view long thin district shapes located on a diagonal () as less compact than the same shape located along the horizontal axis ().[6] 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 virtually always use this single standard orientation and do not rotate districts when evaluating compactness; as a result, their decisions are not rotationally invariant. In other words, since the usual orientation of a district has precedence in how humans interpret it, some of our measures need to pick up on these features.

Thus, primarily for illustration in this section, and later as a measurable feature of district shape that can be included (and if desired controlled) 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

[6] This pattern may be related to the "horizontal-vertical illusion" discovered in psychology (**PriGet93**).

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.

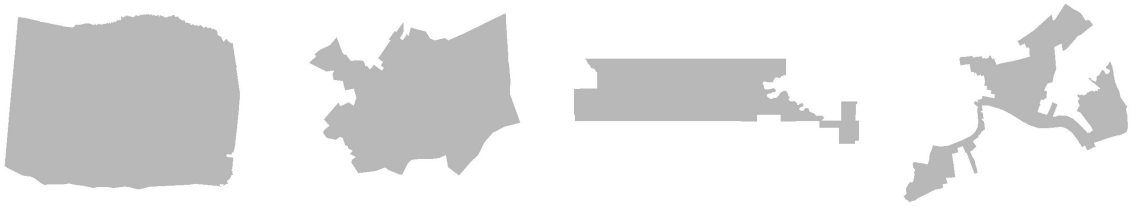
Since we are attempting to quantify human perception, we try to avoid imposing theoretical notions of what compactness should be, what might be rational, or what meets various mathematically "pure" standards that implicate one normative preference or another. Finding the common objective measure that exists in minds of districting authorities, the courts, and others requires respecting how humans think rather replacing it with alternative normative preferences. Although the courts have never addressed the issue, in all likelihood those who drafted compactness requirements in legislative statues, judicial opinions, and state constitutions, that imply that the concept is so simple that you know it when you see it, were not assuming rotational invariance. However, if a rotational invariant measure is desirable or at some point required, we can easily impose it using a procedure analogous to what we do for avoiding political bias. Thus, we would use all the procedures described in this paper except that we would simply display districts at random rotational angles when eliciting compactness evaluations.

Third, 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 border, which of the small border inlets and peninsulas count as notable 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 (**HarSte88**), 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

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 (**AngPar11**; **ChoKimMur14**). Section 3.1 attempts this by applying best current practices in survey research — using a modern version (**David88**) of Thurstone's venerable paired comparisons approach (**Thurstone27** a method that dates at least to 1860; see **Fechner66**). 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 D.

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 **MitGopCar11**). For these reasons, **ConPre86** 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" (**Lior12**).

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., **MitGopCar11**). 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[7]. 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,

[7] All districts are visualized at maximally high resolution to ensure that no features such as coastline are lost.

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.)

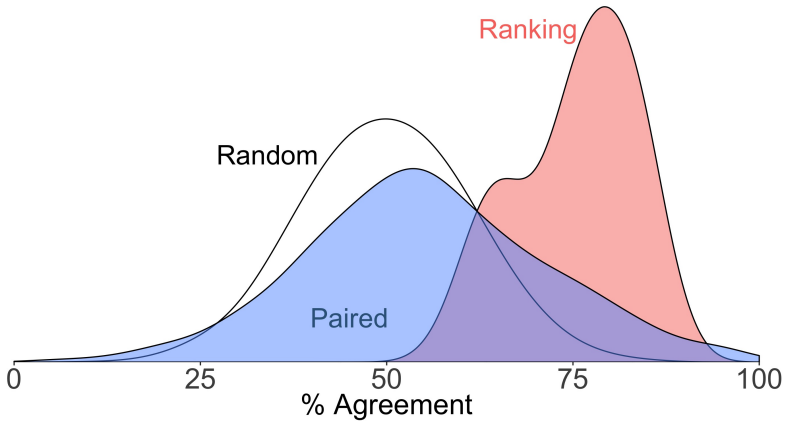


Figure 3: Intercoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

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.)

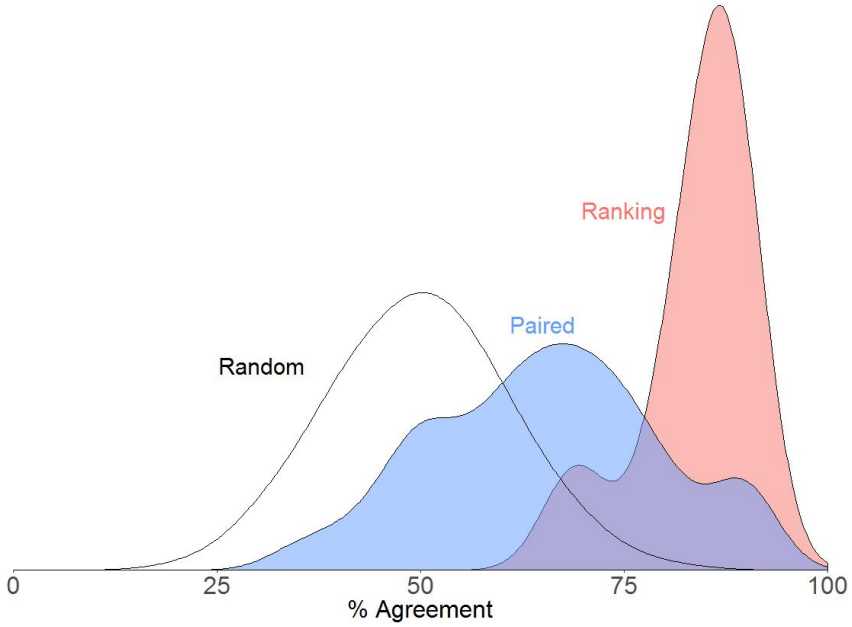


Figure 4: Intracoder Reliability of Thurstone's Paired Comparisons (blue histogram), full ranking (salmon histogram), and a random placebo distribution (white histogram), all using density estimation.

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 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 also tried full ranking with districts printed on paper and arrayed on a long table, as well as via an online system we built that allows districts to be dragged and dropped to their chosen location; we find no evidence that the mode of administration matters either (**Blasius12**).

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 (**Fabbris13**). Our own cognitive debriefing strongly supports the advantages of ranking in this regard.[8]

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 (**IpKwaChi07**). As a hint that this might work, **Krosnick99** (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 (**NieGroCar90**). 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.

[8]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.

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. In other words, while "roundness" may be the deciding factor for compactness in one given pair of districts, length vs. width may be the relevant question in the next pair, and so forth. 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 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 $\rho = 0.9$, 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 Table 1 and many others given in Appendix A. Finally, we train an ensemble of predictive

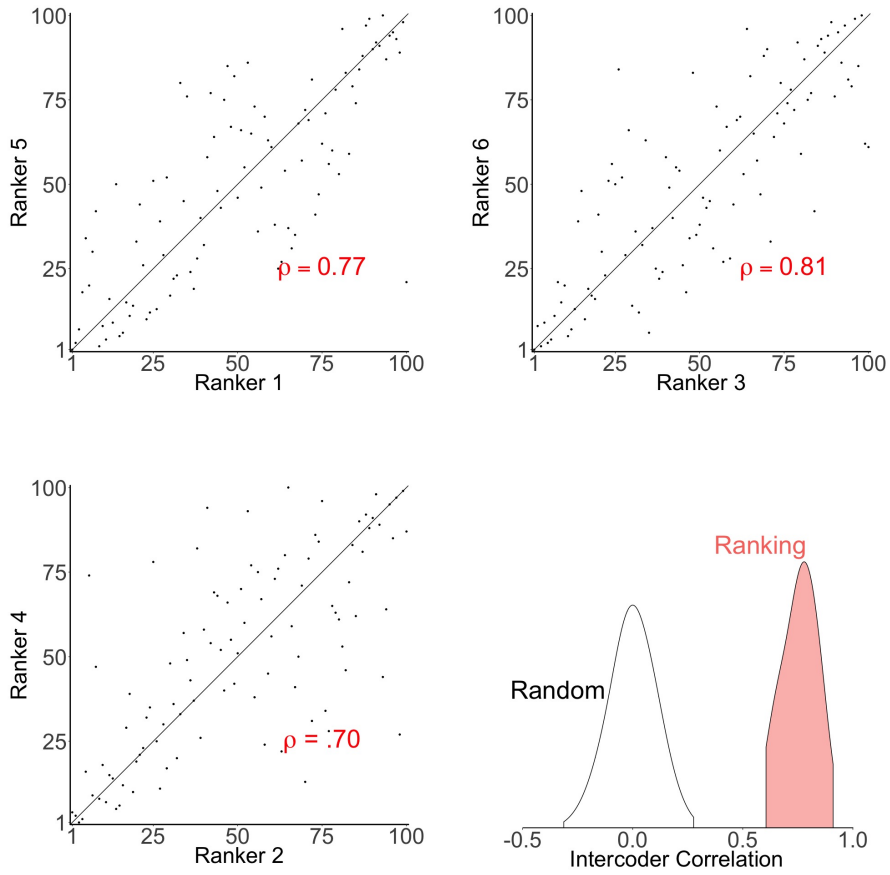


Figure 5: Intercoder Reliability for Full Ranking with 100 districts. Scatterplots are given for the median correlation (top left panel), first quartile (bottom left) and third quartile (top right). A histogram of all correlations, along with a placebo-based histogram appear at the bottom right.

methods with these data, consisting of least squares, AdaBoosted decision trees, support vector machines, and random forests. We detail our modeling approach in the Appendix, Section B; all further details and code are available in our replication data file which will accompany this paper. While our training data consist of integer ranks from 1 to 100, our ensemble model produces continuous outputs on the same scale.

4 Validating

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

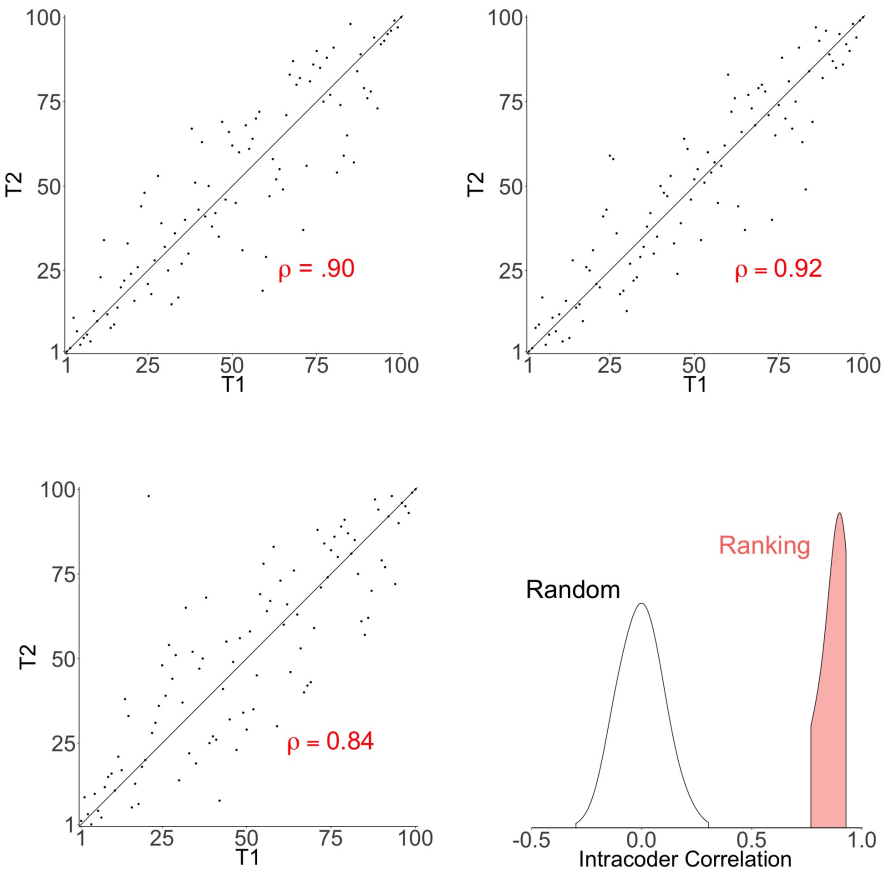


Figure 6: Intracoder Reliability for Full Ranking, following the same heuristics as Figure 5.

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.

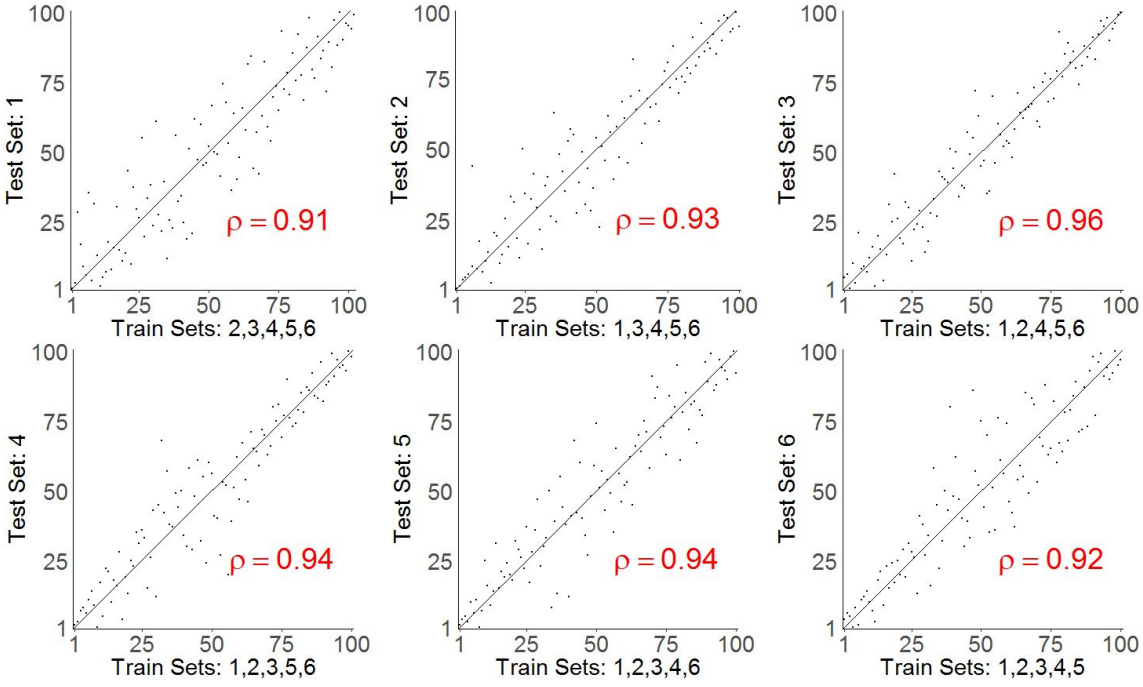


Figure 7: Cross-Validation of Model Predictions

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 can make a prediction for any legislative district shape, including new districts and those that do not appear in our training set.

Our hypothesis is that any informed human being will judge the compactness of a set 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 and comparing these evaluations to our predictions. Our main test comes from 96 sitting justices, judges, and public officials, all with some responsibility for redistricting or deciding redistricting cases. We also elicited the views of 102 others, ranging from less to more involved in and knowledgeable about redistricting, including Mechanical Turk workers, who received small monetary payments, undergraduates, some of whom received hourly wages, and others who not paid, including political science PhD students, law students, law faculty, redistricting consultants and expert witnesses, and lawyers involved in legislative redistricting cases.

We promised our respondents confidentiality, including their responses and the fact of their participation. This was most obviously a concern in recruiting judges and justices, 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 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. (We have also prepared and field tested teaching exercises for American government classes that use our districts, enable students do the ranking exercise themselves, and compare them to our predictions.)

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.

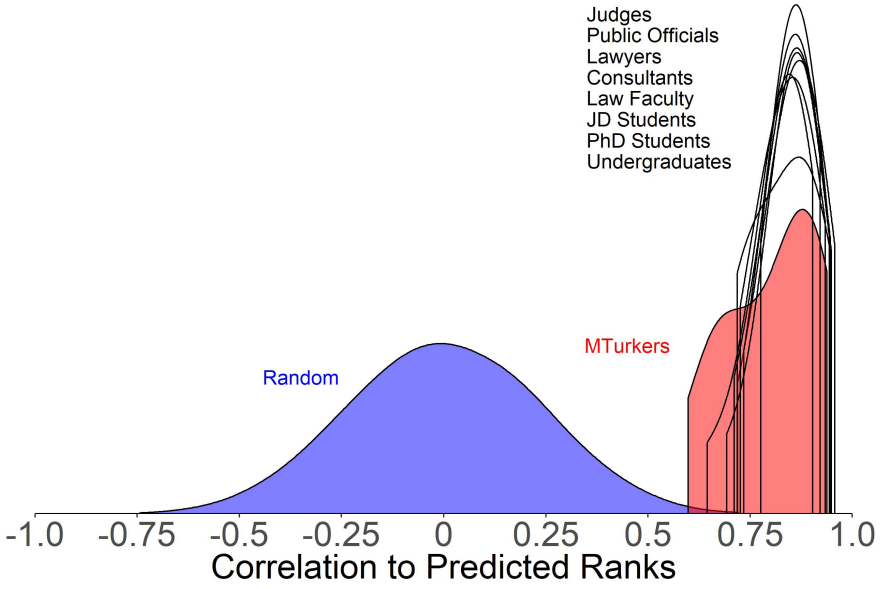


Figure 8: Histograms (via density estimates) of correlations between predictions from our model and answers to survey questions from nine different groups of respondents.

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 Interpreting

Having conceptualized, measured, and validated our estimate of compactness, we now interpret the result. Of course, we already have one interpretation — that we know it when we see it. That is, our fully automated quantification of the compactness of a district geography reproduces how informed human observers evaluate a never-before-seen district shape. Our model can do this instantly for millions of potential districts in ways no human could ever do, but the quantity being estimated by our model and by people is the same.

Nevertheless, a reasonable question is whether we can understand compactness via some simpler top down geometric approach, analogous to any of the existing measures. The common difficulty of explaining how we as humans (or statistical models that approximate them) perform sophisticated tasks — recognizing a friend's face, developing a scientific hypothesis, judging compactness when we see it, etc. — is known as "Polanyi's paradox," that "we know more than we can tell" (**Polanyi66**; **Autor14**). We have studied, in considerable detail and in many ways, how to simplify our measure and find that indeed the simplest way to know what we see is merely to look or to use our measure. A theoretically simpler version may even be an illusory goal, since humans use such sophisticated combinations of these mathematical simplifications rather than any one. We analyze this point in three ways, and then discuss whether other approaches to this question might be possible.

First, we could consider correlations between our measures and several existing ones, but the question is in what data do we perform this correlation. Since they are different measures, it would be easy to construct a data set where the correlations take on any values at all. We thus study the question within different real world groupings, and see whether

one dominates. To do this, we construct 773 data sets of districts, formed from the cross-product of all states, legislative chambers, and years in our collection (e.g., all districts in the Alabama State Senate in 1962). We then compute the percent of times, across data sets, where each existing measure has the highest correlation with our measure. The measure that winds up in the top position most often is Polsby-Popper, but this occurs in only 43.9% of the data sets — followed by the convex hull in 39.9%, Grofman in 21.3%, Reock in 10.4%, Boyce-Clark in 7.8%, and the length/width ratio in 7.5%. In other words, any existing measure can come out on top in approximating our measure, depending on the particular features of the districts in the group, and so none of these measures alone (or in simple combinations, which we have also tried) can be used as a simpler replacement or even as a rule of thumb, at least not without checking the relationship first.

Second, we offer illustrations of the nature of the agreements and disagreements between our measure and each of the seven existing measures we discussed in Section 2. For each existing measure, we construct a 2×2 cross-tabulation of example districts that reflect agreements (compact and noncompact) and disagreements (where the existing measure says noncompact and ours compact, and the reverse). We array horizontally the four cells of this 2×2 table for each measure in a row in Table 2. To generate this table, we define "compact" districts as having a predicted compactness rank in the top 15 (of 100) and "noncompact" as 85 or lower. (If no district appears in a cell of the cross-tabulation, we expand our definition from 15 and 85 to 20 and 80, etc.) Then, to avoid cherry picking, we choose the first in alphabetical order[9] among all districts defined by each cell in each table.

The results in Table 2 are striking. The agreements appear in the first two columns: Column one includes seven obviously compact districts, and column two includes seven clearly noncompact districts. The last two columns reflect disagreements between our measure and an existing one. The first of these (in the third column) are districts that our measure indicates are noncompact and an existing measure says are compact. Most human observers agree with our measure (by design) that these are in fact highly non-

[9]We define alphabetical order according to a specific naming convention. All districts receive an identifier which includes state, district set (upper chamber, lower chamber or Congress), district number, and year. For example, Alaska's first congressional district from 2010 is 01_CD_001_2010.

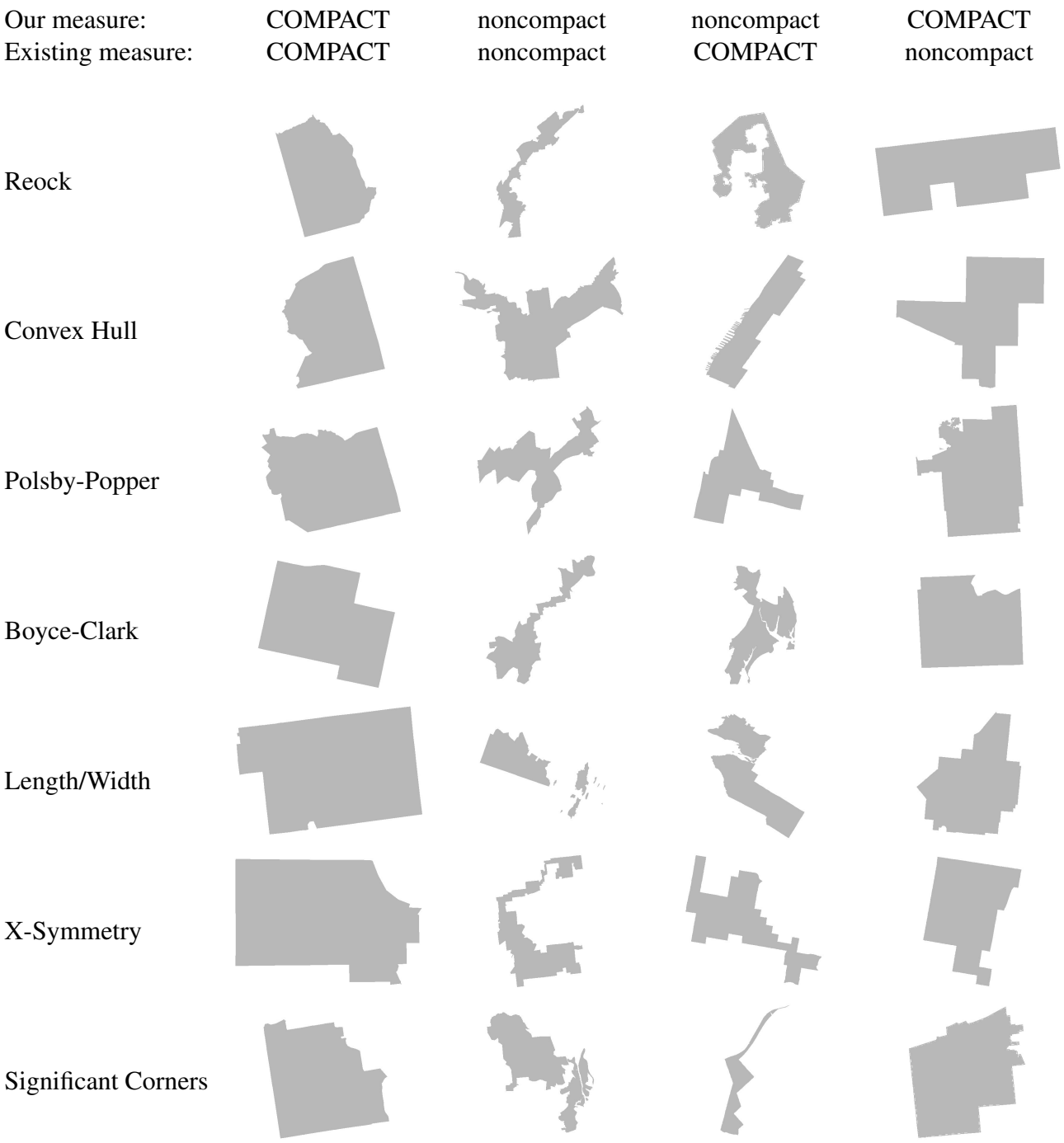


Table 2: Illustrations of agreements (in the first two columns) and disagreements (in the last two columns) about the degree of compactness between each of seven existing measures and our measure. Each row represents a 2×2 table of our measure by an existing measure, with a dichotomized compactness summary, displaying one example district in each cell arbitrarily chosen via alphabetical order.

compact districts. Similarly, the final column includes districts judged as noncompact by an existing measure, but compact by ours. This table clearly reveals how each existing measure picks up important features of the compactness of legislative districts and omits others. The features each measure picks up or misses are those widely discussed in the existing compactness literature as benefits or failures of each measure, since in practice this theoretical literature is using the standard from which our measure was derived (you know it when you see it) to judge their own measures. In contrast, our measure seems to pick up all the features identified throughout the literature as desirable, without obviously missing any feature of a district shape generally seen as important.

Finally, we take a computational approach and consider whether our specific machine learning ensemble can be made more parsimonious. As it happens, the best practice in choosing predictive models that are likely to work out of sample, which we followed, involves finding the most parsimonious model that predicts accurately; as such, we are by definition unable to find an even more parsimonious without giving up some degree of predictive accuracy. Thus, we tried here to find a model that was substantially more parsimonious but which degraded performance by only a small amount. Unfortunately, we found no large discontinuity in the relationship between parsimony and performance, and so did not find a much simpler model that also predicted almost as well as our more sophisticated one. (A straightforward principal component analysis of the existing measures also does not yield a simple solution; indeed, the third principal component correlates more highly with our measure than the first two, although this result varies over data sets as well.)

In summary, this section demonstrates that none of the existing measures, and no measure we were able to come up with, offer a simple geometric representation for what humans know when they see. To be clear, however, we have not proved that creating such a measure is impossible. We thus leave this as an open question and encourage future researchers to seek such a simplifying geometric definition, if that turns out to be possible. This may follow from the literature on interpreting machine learning models (**LetRudMcC15**; **VelMarLis12**). Or perhaps scholars in geometry could derive an un-

derstandable parametric form that takes a large range of polygons (approximating even if not equaling real districts) as inputs and outputs compactness. One complicating factor in this work is that the sample space of possible compact district shapes are apparently less numerous than noncompact shapes, and those with middling levels of compactness may even be the most numerous.

6 Concluding Remarks

We conclude that the measure derived here reflects the underlying viewpoint held about the concept of compactness by everyone from educated Americans to public officials, judges, and justices. This measure appears to confirm and reflect the single, universally recognizable standard implicit in legal compactness requirements of state constitutions, federal and state legislation, and court decisions. Although "we know more than we can tell" about how humans perceive compactness, this measure quantifies "what we know when we see." The measure is also visibly different (as per Table 2) from any existing measure and, by design, much closer to how human beings perceive compactness.

Approaches developed here for measuring an ill-defined concept that you know only when you see may also be applicable to other difficult-to-define concepts. These include measurement 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 here 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. In this light, we encourage others to take up this challenge and improve on the methods we propose, and develop statistical methods that outperform ours; this may now be possible, as clear performance standards now exist. New features measuring compactness can also be included in our approach as additional covariates in our statistical model, which may well be improved.

We hope the large collection of compactness data we make available with this paper (for 18,215 state legislative and congressional districts) and software that makes it easy to compute compactness on any new district enable future researchers to study a wide range of questions related to this crucial concept (see Appendix D). As well, we hope that having a single measure of compactness that all agree on will begin to constrain some aspects of unbridled advocacy during the redistricting process and subsequent litigation.

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 define 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 (**NieNoc08**) and as Hull(D) the minimum bounding convex hull (**KinZen06**; **KonEveTou90**). 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 $\text{mad}[g(i)] = \frac{1}{\#S} \sum_{i=1}^{\#S} |g(i) - \text{mean}[g(i)]|$.

Measures The perimeter of the minimum bounding circle is PC $= P(\text{Circle}(D))$ and minimum bounding convex hull is PCH $= P(\text{Hull}(D))$. The area of each is the AC $= A(\text{Circle}(D))$ and ACH $= A(\text{Hull}(D))$. 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 $= A(D)/A(\text{Circle}(D))$; GROFMAN $= P(D)/\sqrt{(A(D))}$; HULL RATIO $= A(D)/A(\text{Hull}(D))$; 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 $= 1 - \frac{1}{2\,\text{mean}_i[r_i]} \text{mad}_i[r_i]$ (**MacEachren85**); POINTS $= n$ for the district polygon defined by the official US Census shapefile; using the Harris Corner Detector algorithm (**HarSte88**), 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 (**AngPar11**). 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 Ensemble Modeling

Our model training and evaluation procedure involves five steps: (1) Partition the data into training and test sets; (2) fit each of four models separately on the same training

data: linear regression with variable selection, random forests, AdaBoosted Decision Trees (ADT), and support vector machines (SVM); (3) Calculate each model's predictions for the test set; (4) average each model's test set predictions; and (5) compare the averaged prediction vector to the true labels for the test data. We offer detailed information about each step in the replication data file that accompanies this paper and details about each model in Step (2) here:

Linear regression with variable selection To perform variable selection on our standard OLS model, we experimented with many included covariate sets and selected the best one via cross-validation. Beginning with the full set of covariates and interactions, we iteratively dropped the worst-performing covariate and, as we did so, observed the cross-validation accuracy increase. We followed this procedure until the cross-validation accuracy began to decrease.

The selected main variables are: Polsby-Popper, Boyce-Clark, Convex Hull, Significant Corners, X Symmetry, Y Symmetry, District Area, Variance of Corners' X coordinate, Variance of Corners' Y Coordinate, Variation in Line Segment Length. As well, included are the following interactions: Polsby-Popper * Convex Hull, Polsby-Popper * X Symmetry, Polsby-Popper * Y Symmetry, X Symmetry * Y Symmetry, Polsby-Popper * Significant Corners, Convex Hull * Significant Corners, Polsby-Popper * X Symmetry * Y symmetry.

Random Forest Random Forests, which consist of bootstrap-aggregated decision trees, are among the most commonly used machine learning models in practice, and show strong performance without tuning. We train our random forest using 2,000 trees and the default settings in the `randomForest` library (**LiaWie02**).

ADT ADTs hold great promise for the social sciences (**KauKraSen18**). They are ensembles of trees, similar to random forests, but each tree is trained on a version of the data set reweighted according to the previous tree's residuals. We train our ADT using 2,000 trees, interaction depth of 3, and otherwise default settings in the `gbm` library

(**Ridgeway06**).

SVM Support vector machine regression is also widely applicable and requires little tuning. They utilize one of a variety of kernels to obtain nonlinear estimation and sparsity. We train our SVM using the default settings for the `e1071` library (**MeyDimHor14**). The default kernel used is the radial kernel.

Appendix C Uncertainty Estimation

Prior approaches to compactness do not define theoretical quantities of interest separate from their proposed empirical measures. As a result, the statistical properties of these measures have not been defined or applied. And without this key distinction, estimates of uncertainty (based on deviations from a quantity of interest) have not been introduced. If a quantity of interest were defined for a compactness measure, uncertainty estimates could then be generated using bootstrap, frequentist, likelihood, Bayesian, predictive, or another theory of inference.

Our theoretical quantity of interest is perceived compactness, which we theorize is common across educated people. Like all existing compactness measures, our proposed measure is a deterministic function of only the district shape. We treat our measure as a prediction of perceived compactness and evaluate its uncertainty based on the success of this prediction. We thus base our uncertainty estimates on prediction error, which a function of (a) measurement error in eliciting views of compactness from any individual, (b) actual variation across individuals in their views, and (c) predictive inaccuracy. (One could also add uncertainty based on other theories of inference but this seems unnecessary for most applications.)

We offer uncertainty measurements for a single compactness measure and for the difference in two compactness measures. For a single measure, we plot all our data used to evaluate out-of-sample our compactness predictions in Figure 9 (left panel), with our predicted compactness horizontally by the absolute deviation from the truth vertically. We then sore these data into 20 bins defined on the horizontal axis. Then we calculate for

each bin the quantiles of the absolute deviations from the out-of-sample truth. We record the 20 points that are at the 50% quantile and the 20 at the 95% quantile. Each fairly closely follows a quadratic curve and so we fit a polynomial regression and add these to the graph (black for 50% and red for 95%). The height of the black curve then represents the average amount of uncertainty we should expect and the height of the red curve indicates, for any given prediction, half the width of the 95% predictive interval. The red curve happens to have a relatively simple and easy-to-use form. Let c denote predictive compactness. Then half the 95% confidence interval is simply $c - 2 - 0.01c^2$. So for a highly noncompact district with a score of 90, the 95% interval is ± 7.[10]

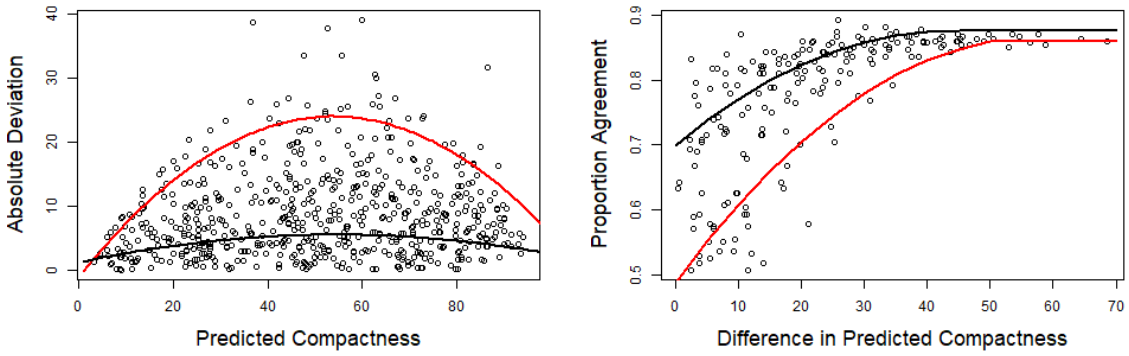


Figure 9: Uncertainty Intervals for a single compactness score (left panel) and for the difference in two compactness scores (right panel). Each graph plots the uncertainty on average (black line) and that which bounds 95% of likely outcomes (red line).

Finally, Figure 9 (right panel) gives uncertainty estimates for differences between two predicted compactness values. We do this by computing the percent agreement on those two districts (vertically) by absolute differences in predicted compactness for two districts (horizontally). We then again create 20 bins on the horizontal axis and compute the 50% and 95% quantiles, and fit smoothed lines (which are also quadratics, except to the top right with few data points). For example, the red line indicates, for a difference of 10 in predicted compactness between districts i and j, on average 75 evaluators out of 100 will agree that district i is more compact than district j, and only 5 of 100 will fewer than 60

[10] We also perform this procedure treating positive and negative errors separately, producing two separate quadratics rather than one. This less efficient procedure produces similar but less conservative predictive intervals, and so stick to the procedure in the text.

judges out of 100 agree.

Appendix D 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 **LewDevPit13** and state legislative districts from **McMLinVan03**.

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, as well as an estimate of uncertainty using a bootstrapping pro-

cedure described in Section C. 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 (**KinBro87**; **GelKin94**), 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 10, 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 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?

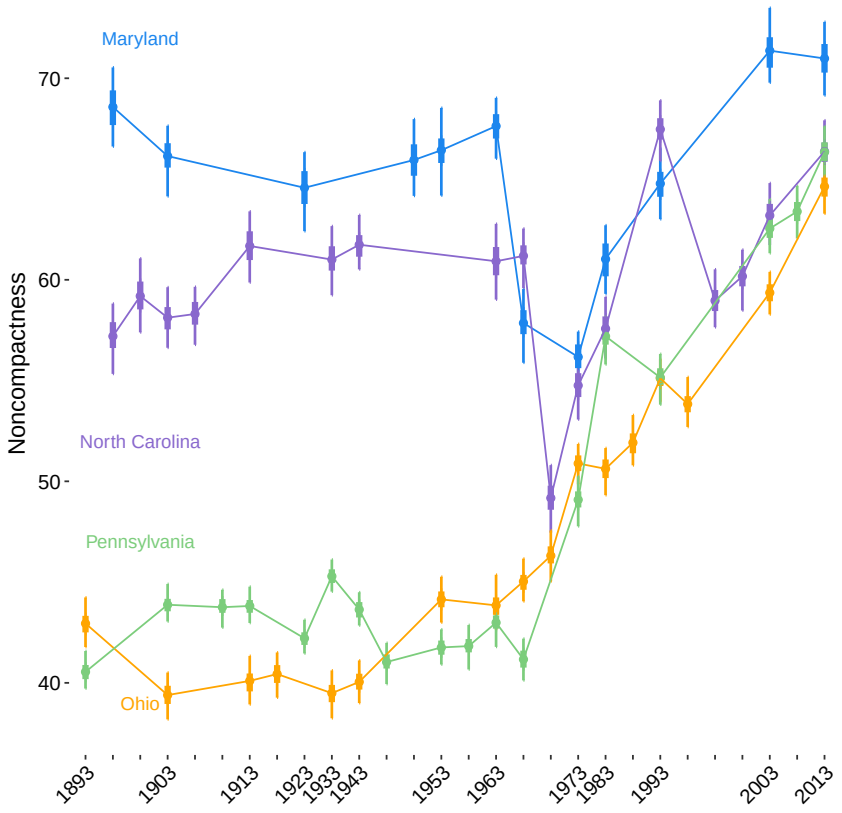


Figure 10: Time series plots of average district compactness in congressional districts for four states often claimed in the media to be political gerrymanders.

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.