Sorting Algorithms for Qualitative Data to Recover Latent Dimensions with Crowdsourced Judgments: Measuring State Policies for Welfare Eligibility under TANF. *

James Honaker † Michael Berkman ‡ Chris Ojeda § Eric Plutzer ¶

Presented to the Applied Statistics Workshop, February 12, 2014

Abstract

The Quicksort and Bubble Sort algorithms are commonly implemented procedures in computer science for sorting a set of numbers from low to high in an efficient number of processes using only pairwise comparisons. Because of such algorithms' reliance on pairwise comparison, they lend themselves to any implementation where a simple judgment requires selecting a winner. We show how such algorithms, adapted for stochastic measurements, are an efficient way to harness human "crowdsourced" coders who are willing to make brief judgments comparing two pieces of qualitative information (here, sentences of text) to uncover the underlying dimension or structure of the qualitative sources.

As a demonstration of the ability of our approach, we show that correctly structured non-expert judgments of the level of democratization in countries recovers the same information that alternate expert scales of democratization –with large cost and time– estimate.

Our key motivating implementation involves a large collection of written policies describing conditions for eligibility for welfare (TANF) in each state in the US. An open question in the literature on welfare is the existence of a race-to-the-bottom, which necessitates measuring the generosity of complex sets of eligibility rules that differ from state to state and across time.

Existing approaches in the literature have attempted to scale or rank state generosity in welfare policies by either constructing large coding questionnaires (Fellowes and Rowe, 2004) and summing responses, or by attempting factor analysis of all possible raw data on welfare policy (De Jong et al, 2006). The first approach requires understanding all the dimensions that are relevant before constructing the survey implement. The second requires converting all policy documents and rules into quantitative measures.

We show how to obtain a ranking of state welfare generosity without doing harm to the qualitative nature of the sources, and without leveraging expert knowledge to sort the vast collection of textual sources. We present "crowdsourced" human coders sentences describing one policy measure in each of two states, and ask them which of the two is the more generous (or more flexible, or more lenient) welfare rule. The optimal set of pairwise comparisons is continuously chosen by the sorting algorithm. We compare the rankings of state policies

*We thank for helpful comments Justin Gross, Burt Monroe, Chris Zorn and participants of the Text Analysis in Political Science conference hosted at QuaSSI, Penn State, as well as the NSF for support (SES-1059723) and the Minerva Research Initiative.

†Senior Research Scientist, The Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street CGIS Knafel Building, Room 350, Cambridge, MA 02138 jhonaker@iq.harvard.edu

‡Professor and Director of Undergraduate Studies, The Pennsylvania State University, Department of Political Science Pond Laboratory University Park, PA 16802 mbb1@psu.edu

§Doctoral Candidate, The Pennsylvania State University, Department of Political Science Pond Laboratory University Park, PA 16802 cjo136@psu.edu

¶Professor of Political Science & Sociology, Academic Director, Survey Research Center. The Pennsylvania State University, Pond Laboratory University Park, PA 16802 exp12@psu.edu

created using judgments from paid human coders through Mechanical Turk, as well as more resource intensive rankings we created to replicate the scaled indices and factor scores used in the previous literature for the most recently available data.

We demonstrate that it is possible to reveal structure, and to organize textual information through "human processing" by relying on algorithms common to quantitative methods, but without any actual quantification of the qualitative textual sources. Moreover this is a highly resource efficient method to organize large corpora of written information.

We demonstrate the powerful performance of non-expert human intelligence, when given sufficiently small structured textual tasks, set out as pairwise comparisons, and show how well understood sorting algorithms can take these human judgments and uncover the latent quantitative ordering of the qualitative sources. The results are very cheap, incredibly fast measures that correlate as strongly with gold standard statistical methods as alternate statistical specifications scale with each other.

1 Introduction

A quantitative research program requires measurement. Oftentimes we need to measure central, fundamental concepts that are intuitively meaningful, but extremely difficult to precisely define. The standard approach is to use expert knowledge to set out all the important features of the concept, then measure those features for each observation, and then collapse all that coded information down to a reduced, manageable scale. We do not disagree that, properly implemented, this can be the gold standard target of quantitative measurement. Indeed, by situating this as the "gold standard" we allude to the fact that this can be a very costly process in terms of project resources, and researcher time. Rather, in this paper, we examine another method of quantification of vague, but intuitively meaningful concepts, that relies on simple, cheap (but highly structured) pairwise comparisons by non-experts making direct judgments of qualitative sources.

As the building block of our method, we present non-experts with qualitative information about two observations, and simply ask them to tell us which observation has more of some concept we are trying to measure. We go to lengths to argue that many meaningful social concepts can be effectively evaluated or judged in this manner by non-experts. An individual may not know Dahl's or Gurr's definitions of democracy, but if I give them descriptions of two countries, they can tell me which country is more democratic. An individual may not understand all the various intricacies of state policy, understand the evolving history of welfare in this country, or know the ways the states have divided themselves, but they can evaluate two rules and tell which one seems more generous, or which one they would rather live under if hypothetically they needed assistance.

Our central argument is that collectively these judgments are meaningful, informative, and well ordered. We then explore a number of sorting algorithms –algorithms that employed in other domains are quite foundational and well understood– that can arrange objects in a list using only up-or-down comparisons between pairs of objects. The trick is to correctly and efficiently choose the right pairs to judge so as to yield the most information. We demonstrate that, properly employed, these algorithms, coupled with non-expert judges, have the potential to cre-

ate measures and scales of difficult concepts that correlate extremely highly with elaborate coding projects coded by experts –indeed, as highly as different expert schemes correlate with each other– for a small fraction of the resources and time.

Conceptually, we also find the on-the-ground validity of qualitative, human generated sources, being directly judged in a qualitative fashion by a social population of individuals who rely on an instinctive understanding of the concept being measured, very appealing. We do not interpret our questions as opinion polls, as much as very tiny coding exercises, distributed across a large population, arranged and structured algorithmically to learn the most about the latent concept being measured. In other words, we are using this population, who has an intuitive, qualitative ability to make simple judgments, and then structuring the set of judgments that they see so that collectively this non-expert population is harnessed to form a scaling of observations in a latent dimension. Through this, we are accomplishing a scaling task that would normally require an immense amount of expert work, first to uncover the important dimensions, then to code them from the sources, and then reduce and scale from those dimensions back down to a single measure of the latent concept. In our examples, we compare both this approach, and the well understood "gold standard." Succinctly, we see the comparison as being between a costly, extremely accurate and expert measurement of lots of things orbiting the thing we actually desire to measure, versus an inexpensive, weak measure of *exactly the concept we are interested in*; and moreover a weak measure that can be repeated increasing many times for increasing precision.

2 Approaches to Quantification

We first discuss the abstract process behind a quantitative coding design. We discuss in detail what in most contexts could be assumed understood, so as to concretely define how our approach and methods differ. That is, we want to be able to describe our measurement design in the clearest contrast, and that first requires explicitly discussing the pragmatics of measurement.

Let us assume there is some object in the world that we want to measure: perhaps whether someone has voted, whether a nation is at war, how democratic a country is. These are commonplace variables in the political literature, but each are progressively complicated to measure in a quantitative fashion, even in the presence of voluminous information. Let us consider them in order

In the first example, measurement is trivially aligned with the observable object. The act of voting seems reasonably dichotomous, one votes or does not vote. Perhaps we need to decided how to deal with abstention, and modify our coding to the simpler act of turnout. But the object of interest fits into a few categories, and those categories are readily observable, and we can assign quantitative values as labels to each category we directly witness.

Sometimes measurement becomes difficult at the boundaries between categories. Even if we think that the concept of war is clear, and most countries that we witness can either be labelled as obviously not at war or clearly in midst of war, there are boundary issues. "War" is a vague predicate. There are countries with low, but not absent levels of violence, or the violence appears predominantly one sided, or the exact actors might not be states, and so measurement requires making some decisions about what the definition means, how it might be coded. This requires some effort on the part of experts to formulate a set of rules map from any set of observables to a simpler

scaled variable. One rule might be "a war exists when at least a thousand combat deaths have occurred, with more than one hundred on each side."

Sometimes measurement is exceptionally hard because no observations are readily labelled as any particular value, or the boundaries are perplexingly difficult. When the information we have available from which to create our measure is a large collection of qualitative, textual sources, this may require the expert reading through all these sources, noting all the different dimensions and attributes that can be measured, constructing meaningful scales on which to measure them, and then going into the raw sources and coding all of these values from all of the raw materials. Next, it is the step of some algorithm for data reduction to take all these large quantity of numbers and reduce them in complexity to a small number of dimensions, perhaps a single value, for each object or observation.

Collectively, this requires an enormous amount of expert knowledge and expert effort. First we require substantive expert knowledge to realize and describe all the dimensions that need to be measured. Next we need proficient thoroughness to then code these dimensions for every one of the observations. Then it requires quantitative knowledge to employ a scaling technique so as to collapse all these dimensions of coded data down to a measure of the latent concept. This might be a statistically derived model for data reduction, like principal components or an IRT model, or it might be a more substantively derived scaling method involving judgments about which variables will have high weight or low weights in some constructed sum or index.

In our motivating example, we needed a measure of how generous states were in their policies towards welfare recipients. As a research question we were interested in understanding how the underlying level of generosity in welfare rules in a state changed the way state welfare officers ("steet-level bureaucrats") conformed to the written statutes they were supposed to enforce. Each state had a long set of written rules describing conditions and exceptions and policies under which an individual could qualify to continue to receive welfare payments. Rather than there being merely thirteen attributes of mercy, after months of reading the source documents we identified 218 different variables that were necessary to describe all the features of a state welfare policy's level of generosity. These 218 variables coded all the possible ways that one state might measurably differ from another. These might be quite precisely defined variables with tiny scope, such as "Are mother's over the age of 21, who are without a high school diploma and who have children under the age of six eligible to count hours spent in job training toward required work hours", or they might be broader variables such as "Can the chronically ill get a waiver from work requirements." We then measured these 218 variables for each state from the qualitative text sources, Finally, we then developed an IRT model to estimate the latent level of generosity of each state given all the observed features of their measured rules. This was the primary research activity of four experts for eight months (Berkman et. al 2013). We are going to compare this to a measurement exercise using qualitative judgments that took less than three hours to perform.

We do not argue against the gold standard measurement method when available or feasible. But we do recognize that not every project has eight months and a team of experts to construct a measurement of a difficult concept. Rather than forego projects entirely, what is presented here maybe a good measurement technique for some tasks where standard quantitative approaches would be exorbitant or infeasible.

2.1 Judgment versus Quantification

However, just because the concept becomes more difficult to codify, or even becomes abstract, does not mean it becomes less immediately meaningful. Justice Stewart famously wrote in a concurring opinion on an obscenity ruling "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that."(Jacobellis v. Ohio, 1964) Colloquially, "But I know it when I see it" has come to mean an ironic subjective fluidity. But Justice Stewart was an experienced Supreme Court Judge who was struggling to provide a definition. And yet he knew he could make a judgment in a specific instance without constructing a set of rules defining all possibilities. In many respects, our framework is faithful to his exact meaning. There are many difficult, vague concepts, that are extremely hard to codify in such a way as to exhaustively define all possible situations that might arise, and yet some of these concepts can be immediately judged in individual cases.

Indeed, when we are dealing with socially meaningful terms, if these could not be judged to exist in the real world independent of a rigourous definition, the original concept would have no meaning. Ideas such as democracy, freedom, generosity, are social constructs that are meaningful in ordinary interactions among individuals in the world, without them precisely defining their terms. In these cases, it is precisely because they have ready meaning in the social world of mass behaviour that we as scholars are interested in studying them. Our ability as academics to uncover and reveal exactly what these terms mean is valuable work, but these concepts predate our rigorous definitions, and have *valid meaning among individuals who use the term, even if those individuals can not define the terms they use*, or ascribe why they make the particular judgments they firmly believe to be true.

Non-expert participants might not be able to define what makes a state democratic, or what makes a welfare policy generous or when exactly a state of war exists, but they know it when they see it. In fact, we are going to require even less of participants than this. We will simply require that given two observations, they can judge which possesses more of some quality. This requires less than the Justice Stewart approach. Deciding whether an observation does or does not obtain a particular quality (obscenity, democracy, generosity) requires first mentally depicting some vague boundary and then deciding whether the judged object is has greater or less of the quality than the depicted boundary. By giving individuals a pair objects, we are lessening the cognitive complexity of the task. Implicitly, one object immediately forms the boundary and, now as before, the participant judges whether the next object is above or below that given boundary. Moreover, we avoid issues of anchoring that arise if we worry that different individuals would construct different vague boundaries.

It should be clear, also, what we are not expecting of participants. We do not expect them to be able to judge qualities that are academically coined, but have no daily meaning to respondents. The concept of "Anocracy" is a useful academically constructed concept that usefully organizes and categorizes some groups of countries, but as a word of technical coinage we would not expect non-experts to make judgments surrounding its use. Moreover, we would not expect non-experts to be able to rank-order a large set of objects. Perhaps they could feasibly order three objects highest to lowest, or in some contexts more, but we do not assume that given a large number of qualitative objects, that differ in many different dimensions, that non-experts could lay them out in one scale. In our examples we only ask individuals to compare pairs of observations. It suits our algorithms, and conceptually we like the idea

of constructing a high-information coding from very many low-informational tasks. But in terms of improved efficiency, speed and cost, it might be possible to successfully tax the abilities of participants to higher degree.

3 Sorting Algorithms

In quantitative social science, we learn a large number of statistical models, and also generally learn something about the algorithms that implement them. Although we learn them hand-in-hand, these are different things. One could write down a statistical model and prove that a particular estimator has desirable properties under certain given assumptions, and if there is no algorithm that can compute that estimator, the statistical model is still valid but not of practical use. In the other direction, we can often write down algorithms that seem intuitively appealing in their design, and which have nice numerical properties in terms of speed and computability, but perhaps have no statistical model to justify them. In the machine learning literature, such algorithms are increasingly common.

Typically, however, we learn these things hand-in-hand. We learn the Gauss Markov theorem and the properties of the linear regression statistical model. And at the same time we also learn the algorithm to obtain regression coefficients easily expressed in linear algebra as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{Y}'\mathbf{Y}$. Perhaps, at an even more foundational level of understanding, we also learn the relative merits of different algorithms for taking the necessary inverse, such as Gauss-Jordan elimination, or the use of determinants and cofactors.

By and large, the algorithms we learn in quantitative methods training tend to focus on different approaches to optimization: analytical optimization justified by calculus (as in the above example), various hill climbing approaches useful in maximizing likelihood or other objective functions, or even more stochastic approaches in the MCMC literature. This focus has a good payoff, as understanding something of these algorithms gives some pragmatic sensibilities as to how estimates obtained by these algorithms can fail in practice to match the statistical models they are paired to. But the algorithms themselves are distinct from the statistical models.

3.1 Comparison Sort Algorithms

An enormous, and critical body of algorithmic understanding are the numerous algorithms for sorting numbers. Sorting algorithms are the foundational algorithms of computer science. Efficient methods to sort numbers are central to merging objects, conforming objects' type (canonicalization), and memory allocation. The object of sorting algorithms is to take an arbitrary object and rewrite it so it obeys some objective; for example, taking a set of numbers and ordering them so each one is greater than the last, or taking a set of words and alphabetizing them. Often, the primary goal of a sorting algorithm is to be fast, that is, reach a sorted state in a small number of computational operations. A secondary goal might be to be algorithmically simple (easy to code) or transparent (easy to understand). The majority of the sorting literature is concerned with the class of algorithms of *comparison sorts*, that is, they iteratively use some binary function (for example a function that could be interpreted as the logical $>$ "greater than" sign), rather than any more involved or high level function (for example, a function that positions objects onto a number line). Comparison sorts are often simple and transparent, but most importantly they are fast in terms of total machine operations required of the processor.

We are interested in comparison sorting algorithms because we believe pairwise comparisons are a function that non-expert human intelligence is often capable of performing, and we want to be able to efficiently harness all the pairwise judgments to form an underlying ranking. In what follows, we very briefly describe the mechanics of two popular sorting algorithms, and then compare them to a much slower and costly, but more transparent approach. Then we will demonstrate how we implement them with human judgment as the locus of the comparison function.

The *Quicksort* algorithm (Hoare, 1962) is widely used in many real computational applications. An element is selected at random from the set (perhaps the first element in a list). Then all the other elements are compared to that item. Elements that are less than the referent are moved below that item and items that are greater than the referent are moved above the item. The referent is now fixed, and the two new groups, below and above the referent, are now recursively rearranged in the same fashion. An example of this shown in the left column of table 1. The numbers 1 through 5 are unsorted in the first line. The first number, 3, is picked as the referent. In the comparison in the first line, 5 is greater than 3, so stays to the left of 3. In the next line, 1 is less than 3, and so moves to the left of 3. Finally when all the numbers have been moved to the left or right, the 3 is now fixed and those two smaller groups are now sorted in the same fashion. From this behavior, we see the *divide-and-conquer* nature of quicksort, in that the 3 divides the set in two, and relative to the 3, these numbers will always remain on the same side. If the data is originally random, each iteration of quicksort will occur on a set that is about half as small as the previous, and so the sets get exponentially small until they are singletons and need no more sorting. For this reason, quicksort is generally quite fast. In lists of length n that are initially random, the total number of comparisons required will be some function where $n\log(n)$ is the largest term; We say then that the *order of operation* of quicksort is $n\log(n)$ or compactly, $O(n\log(n)$.[1] It tends to slow down if the data is already partially sorted as each division moves fewer observations around. In the worse case, when Quicksort sets about to order a set that is already sorted, then the algorithm is $O(n^2)$, which for large n is many more operations than $O(n\log(n)$. We will also look at a much older divide-and-conquer algorithm, *Mergesort*, (Goldstine and von Neumann 1948) which starts off with pairs of observations and orders them. Pairs of these pairs are then merged together, preserving order, to form larger ordered sets, which are iteratively merged until the entire set is merged together. While typically slower than Quicksort, Mergesort distributes the paired comparisons much more evenly across the set of objects which we thought might be a desirable property for some of the purposes examined.

Bubble Sort is an algorithm that is often used as an introductory example to teach the concepts of sorting, (for a detailed history, see Astrachan, 2003) although it is often derided as an algorithm that is always taught and never used.[2] In brief, bubble sort compares the first and second elements in an unordered set. If they are in the wrong order their positions are swapped; if they are in the correct order, their positions are unaltered. Then the second and third elements are compared, and then the third and fourth until the end of the set is reached. The entire process is repeated until no items move. Bubble sort is shown operating on the same unordered set in the right of table 1. First 3 and 5 are compared, which are in the right order. Next 5 and 1 are compared, and since $5 > 1$ they are swapped in

[1] For example, if the number of terms required in some particular implementation of quicksort were $an\log(n) + bn = c$ as n gets large, the first term quickly dominates, and thus we would say the order of operation was $O(n\log(n)$.

[2] Knuth famously opined "In short, the bubble sort seems to have nothing to recommend it, except a catchy name and the fact that it leads to some interesting theoretical problems." (Knuth, 1998)

	quicksort		bubble sort	
$(3 < 5)$	3 5 1 2 4		3✓5 1 2 4	$(3 < 5)$
$(3 > 1)$	3 5 1 2 4		3 5×1 2 4	$(5 > 1)$
$(3 < 2)$	1 3 5 2 4		3 1 5×2 4	$(5 > 2)$
$(3 < 4)$	1 2 3 5 4		3 1 2 5×4	$(5 > 4)$
$(1 < 2)$	1 2 3 4 5		3×1 2 4 5	$(3 > 1)$
$(5 > 4)$	1 2 3 5 4		1 3×2 4 5	$(3 > 2)$
	1 2 3 4 5		1 2 3✓4 5	$(3 < 4)$
			1 2 3 4✓5	$(4 < 5)$ Repeat evaluation
			1✓2 3 4 5	$(1 < 2)$
			1 2✓3 4 5	$(2 < 3)$ Repeat evaluation
			1 2 3✓4 5	$(3 < 4)$ Repeat evaluation
			1 2 3 4✓5	$(4 < 5)$ Repeat evaluation

Table 1: *Examples of the same unordered set being sorted by the Quicksort and Bubble Sort algorithms. Each line represents a single pairwise comparison.*

position. Now 5 is compared to 2, and another swap occurs. The name of the algorithm refers to this phenomenon that large objects (like the 5) will rise up the list like a bubble in liquid. Bubble Sort is very slow, requiring $O(n^2)$ operations to sort a random set (although one small advantage is that in nearly ordered sets that have only adjacent pairs reversed, it is quite fast as $O(n)$).

One final but very inefficient method is the *Full Count* sort. If there are n objects to sort, a object is selected, and compared to every other object in the set. The number of times, k, that that object is found to be greater than another object, across all these comparisons is counted. The object is then placed in new sorted sequence in the kth position.[3] This requires exactly $n(n-1)/2$ unique comparisons, so is also $O(n^2)$.

Generally, sorting algorithms are chosen so as to implement fast in the type of problem that forms the anticipated use case. Speed can be a factor of the number of comparisons, as already discussed, but also many other algorithmic operations, such as the number of times the memory needs to be transfered as items are sorted or copied, or the ability of subprocesses of the sorting algorithm to work independently, and thus be distributed in parallel. As we will see, while all these influence the typical understanding of the speed of a sorting algorithm, the only criteria we are interested in minimizing is the number of unique paired comparisons.

[3]Note, the smallest object will be greater than no other objects, and so have $k = 0$, the largest object will have $k = n - 1$. It is common in many languages for the first position in an array to be position zero, even if it is unintuitive from the conventions of set theory. If a sequence of k from 1 to n is desired, obviously we add one to this count (or curiously define any object as greater than itself and add that to the count).

3.2 Qualitative Comparison Sort Implementation

We want to lean on the fact that many difficult concepts can be meaningfully judged by non-experts, as a method to cheaply quantify concepts that would otherwise be prohibitively expensive to measure. We are going to use comparison sort algorithms, where instead of numbers, the objects to be sorted are short qualitative texts, and instead of a deterministic simple mathematical function like "greater than", we are going to use the average judgment of a set of non-experts. However, the algorithm remains unchanged, even though the objects the algorithm sorts are unconventional and the functions the algorithm uses are computed within human intelligence.

4 MTurk

To recruit our non-expert participants, and to display the qualitative sources, we used Amazon Mechanical Turk (hereafter MTurk). This service has a large number of users around the world who are willing, over the internet, to perform small tasks for small payments. People with tasks, called *requesters* in MTurk parlance, post these tasks with short descriptions and a specified payment. Users who are interested, called *providers*, self select to fulfill tasks for which they are interested. Some small degree of control over the pool of qualified providers can be set by requesters, such as the number or fraction of tasks a user has successfully fulfilled, the user's location (judged by their IP address). Arbitrary tasks can be programmed by requesters for providers to fulfill online, but simple tasks can be set up through the MTurk interface (API) which requires only some knowledge of html. When the task has been fulfilled it is submitted back to the requester, who can then review tasks and pays for tasks where they approve the work submitted.[4]

We payed 2.00 (US) per task. The median time to completion of any of our tasks was under seven minutes, meaning we were paying our median coders an hourly rate greater than 17 (US) an hour. A task involved two judgments we were interested in, and a third control judgment, which was a very easy judgment we thought everyone should readily agree on if they were reading the qualitative sources provided. Users of this service often fear that some providers will submit nonsense with the greatest possible speed, or even create automated "bots" that fulfill tasks randomly, so as complete tasks fast and harvest payments for minimal effort. We used the control task as a way to measure the degree to which this might be a problem.

4.1 Previous uses of MTurk

By far the most common tasks available on MTurk are simple jobs that can be electronically presented, but not easily automated, and for which computer algorithms are not currently well suited, but which are relatively easy for humans to quickly understand. One of the original purposes the architecture was first developed by Amazon was to find a way to remove duplicate listings of the same product from the online Amazon catalog; A provider would be presented with the description and images of two products for sale, and the task is simply to decide if they

[4] A pool of money is submitted to Amazon Web Services, sufficient to cover payment for all tasks requested. When a fulfilled task is approved, Amazon takes payment from the pool and credits the user. Amazon takes an additional 10 percent fee as payment (20 percent for providers who have qualified as "expert categorizers.")

are the same item. Duplicate removal is still a common task in the MTurk listings. Other common tasks are typed transcription of audio text, color and picture tagging and annotation of text.

The pool of providers has been an increasingly alluring subject pool for social scientists. The category of "research purposes" is a label a requester can use under which to list their task. MTurkers are a responsive, inexpensive population that provide results with great speed. Across the fields of social science, experiments have been replicated using MTurk providers as test subjects or survey respondents, and compared to responses using randomly sampled populations. Within linguistics, the use of MTurk for annotation and judgment of text is described by Snow, O'Connor, Jurafsky & Ng (2008) and Callison-Burch and Dredze. (2010). Using MTurk tasks to conduct experimental economics studies are described by Paolacci et al (2010) and Horton et al (2011). Applications in experimental psychology are described by Buhrmester et al (2011) and survey experiments in political science are addressed by Berinsky, Huber, Lenz (2012). The results are mixed. Within the United States these users are more highly educated and much younger than the general population. Generally the differences seen are as would be expected from a younger, better educated population, but generally not as skewed or exaggerated as samples conducted within undergraduate student populations, another common, inexpensive and readily available subject pool extensively utilized in some social sciences.

To reemphasize a point, however, in contrast with these above studies we do not conceptualize our MTurker providers as subjects in an experiment, or respondents in a survey, but rather as coders or research assistants, performing very small coding tasks for us. Of all of the above, our task is closest to the academic linguists who use MTurkers to annotate text documents to build corpora. To the extent that these are better educated or younger than the general population we do not see a cause for concern (except if we push our method to some boundary where we might try to measure a particular latent concept that is hard for the young, inexperienced or educated to judge). In this regards we also mirror recent work, discussions and proposals by Benoit, Conway, Laver and Mikhaylow (2013), and D'Orazio (2013) for crowdsourcing as an avenue for inexpensive coding of large datasets. However, each of these is a methodology for a conventional coding exercise in an unconventional setting, where the variables and attributes to be recorded are previously defined by the expert setting out the task. To reiterate, our approach does not aim to place individual observations in a measured, coordinate space, but to harness very simple pairwise comparisons *between* objects, aiming to avoid the need to predefine the various attributes to measure, to avoid anchoring these measures across coders, and indeed to avoid a quantitative measurement process entirely in all the intermediate steps up to creating the measure of the latent structure.

5 A Statistical Model of Vague Judgements

The sorting algorithm creates an ordering of the objects from pairwise comparisons. This gives us an ordinal representation of the position of the objects in the latent space. In some cases, with the same observed information, we can also statistically estimate a continuous, cardinal representation of the latent space. We can derive the same equivalent model from several perspectives: as a measurement error model, as a large dimensional IRT model, and as Gaussian channel. Each perspective reduces to the same equivalent model, and while the last of these is explicitly true to our definition of the judgement process, the measurement and IRT derivations are commonly understood models and provide intuitions for the meaning and behaviour of the estimated parameters, so we present these also.

Assume that each object i has latent position β_i. We could set up the problem of judgement as a measurement error process with additional censoring. When making a judgement between objects i and j, individual n takes measurements, x, of the latent positions with Normally distributed unbiased errors:

$$x_{in} = \beta_i + \mu_{in}; \qquad \mu_{in} \sim \mathcal{N}(0, \sigma_i^2) \tag{1}$$

$$x_{jn} = \beta_j + \mu_{jn}; \qquad \mu_{jn} \sim \mathcal{N}(0, \sigma_j^2) \tag{2}$$

And records the difference, $y_n^* = x_{in} - x_{jn}$, of which we observe a censored version $y = \{1, y^* > 0; 0, y^* \leq 0$. Their observed judgement is the n-th comparison in the data between objects i_n and j_n, coded 1 if i_n is judged greater and 0 if j_n is judged greater. The probability of any observed judgement is then:

$$\Pr(y_n = 1 | i_n, j_n) = \Pr(x_{in} - x_{jn} > 0) = \int_0^\infty \mathcal{N}(\beta_{i(n)} - \beta_{j(n)}, \sigma_i^2 + \sigma_j^2) = \Phi\left[\frac{\beta_{i(n)} - \beta_{j(n)}}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right] \tag{3}$$

A function of the true latent positions of the two objects judged, and their associated object specific error variances. The first principles in equations 1 and 2 imply that individuals are able to make judgements about an object's position in the latent dimension, where as we have premised our study simply on the ability of individuals to make pairwise judgements of relative order. However, this is simply a familiar way to set out the model. If instead we assume that individuals are the receivers at the end of a Gaussian channel, which is sending two values, and their informational goal is to decode which value is greater from the signals that emerge from the channel, then exactly the same result is derived (for Gaussian channels, see Cover and Thomas, 1991, Chapter 10). Similarly, we could set up this model as a complicated IRT model where each possible country-pair, leading to potentially n^2 items sharing only n parameters, and obtain the same expression.

(4)

From either approach then, the model becomes a heteroskedastic Probit with object specific parameters that estimate the specific positions in the space. For identification, we minimally need to peg some set of values and we choose two objects in the space and allow a free variance parameter, as this makes visualization and comparison across bootstrapped samples easiest. As it may be the case that some items perfectly order in small samples, we may

need to impose additional restrictions of the range of the latent space, which we set aside momentarily.

We might worry that some individuals in the data are simply answering randomly. If π fraction of the population is answering randomly (flipping coins), and the rest are as previously modeled, the likelihood for y_n is:

$$\Pr(y_n = 1|i_n, j_n, \beta_{i(n)}, \beta_{j(n)}) = \frac{\pi}{2} + (1-\pi)\ \Phi\left[\frac{\beta_{i(n)} - \beta_{j(n)}}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right] \tag{5}$$

We can parameterize π on some observed characteristics, z_n of the individual making comparison n. If we have each individual making many comparisons, these might be individual specific fixed effects. In our data, we generally have many individuals who each make few judgements. However, there are many heuristic rules among users of MTurk to detect and then eliminate individuals who are not contributing effort. For example, whether they fail simple control questions, how much time they take to complete the task. We can code these various tests as $\mathbf{z}$ and see if they statistically contribute to π thus telling us whether we have evidence in the data that the rules used are valid or working. We can also model π as a function of the task payment and attempt to estimate what is an efficient level of payment to receive the most information for a fixed budget. In our example, all workers received one common question that we felt was an easy judgement with a correct answer that any individual worker should. We took workers from several predefined populations, explained in more detail below, which also serve as covariates, and slightly adjusted the level of payment.

Together, we set this up as a penalized likelihood function, which is equivalent to setting a ridge shrinkage prior on the β's. We could have also set a prior on the σ's but chose to assume a common σ for all objects, leaving estimation of object specific σ's –perhaps as a function of word length and sentence complexity– as a topic for further investigation. This simplifies to:

$$\begin{aligned}\text{logL}(Y|\beta,\pi) = \sum_{n=1}^{N}\Bigg[\, & y_n \log\left(\frac{\pi}{2} + (1-\pi)\ \Phi\left[\frac{\beta_{i(n)} - \beta_{j(n)}}{\sqrt{\sigma_{i(n)}^2 + \sigma_{j(n)}^2}}\right]\right) \\ & + (1-y_n)\log\left(\frac{\pi}{2} + (1-\pi)\ \Phi\left[\frac{\beta_{j(n)} - \beta_{i(n)}}{\sqrt{\sigma_{i(n)}^2 + \sigma_{j(n)}^2}}\right]\right)\Bigg] - \frac{\gamma_2}{K}\sum_{k=1}^{K}(\bar{\beta} - \beta_k)^2\end{aligned} \tag{6}$$

6 Application: Democracy Scores

To test our method in a well understood setting, we first attempted to measure the level of democracy in countries. Characterizing and scaling the level of democracy in states has been a longstanding enterprise in political science; it is a very mature literature, alternate measures abound, and the specific choice of measure can influence findings (See Caspar and Tufis 2003, Pemstein, Meserve and Melton 2011, Honaker and Wright 2013 for discussions of alternate measures available). The most commonly used measure comes from the longstanding Polity project, which we use as our point of comparison.

Our implementation was as follows. We chose a random sample of countries, stratified so as to sample one country at random at every two-point increment along the -10 to $+10$ Polity scale. The eleven countries used in our pilot, spread across the range of democratization were (in order of increasing democratization): Quatar, Bahrain, Kazakhstan, Tunisia, Chad, Burkina Faso, Djibouti, Nigeria, Liberia, Indonesia, Hungary. We took three written paragraphs about these countries from the Country Reports from the annual *Freedom in the World* publication of the Freedom House project. These country reports are written in reasonably similarly structured fashion across countries, so we took the last paragraph of the "Overview," which tends to describe the most recent elections, and the first two paragraphs of the section which immediately follows titled "Political Rights and Civil Liberties" which typically focuses on the distribution of power among branches of the government, and the freedom to form parties.

For each participant, we simply took these paragraphs, placed them side-by-side, and asked the respondents to read them and tell us which country was more democratic.[5] We called the countries "Country A" and "Country B" and replaced all names of the countries with these labels so as to avoid having the participant draw on prior outside knowledge about these countries.[6] Specifically we structured the question as:

> *Compare the available information and decide which country is more democratic.*
> *Read the information on Country A and Country B.*
> *In your best judgment, decide which of the following best describes a comparison of the countries:*
> . • Country A is much more democratic than Country B.
> . • Country A is slightly more democratic than Country B.
> . • Country B is slightly more democratic than Country A.
> . • Country B is much more democratic than Country A.

We gave the respondents a choice of four gradations because we thought the availability of more subtle distinctions would incline the participant to a closer, more careful reading of the text, but for the analysis we simply simply collapsed this down into the dichotomy of which country was more democratic.

[5] For recruitment purposes we titled the task: Reading text to compare the level of democracy in pairs of countries, and used the description: Read a series of text extracts about elections and political institutions and categorize which country is more democratic, and for search purposes, keywords: categorization, political, coding

[6] The names of party leaders were left unchanged, but we assume most participants do not gain much information from this, and although it is an online task, we believe participants are unlikely to look them up.

Figure 1 gives an actual screenshot as the task appears from the interface of an MTurk participant. We posed two such questions on each assignment, and a third question which was used as a control. The control question was the same for every single task. It was constructed from two countries not among the set we were actually interested in, but intentionally selected so as to be an easy judgment. Specifically we used Chile (a reasonably clear democracy) and Azerbaijan (a reasonably clear authoritarian state). We expected that individuals who were reading the prompts carefully would classify Chile as more democratic than Azerbaijan, but were willing to let participants make any judgment they saw fit. The important use of this question was to attempt to identify individuals who were answering randomly and answering many tasks to quickly harvest many payments. Individuals reading the prompt should make consistent answers in each task they attempt when they see the control question. Individuals clicking randomly would have a cycle of different answers to the same control question across tasks. Figure 2 gives an screenshot as the control question appears from the interface of an MTurk participant for this comparison.

Democracy Judgements

For each of three cases below, compare the available information and decide which country is more democratic.

Read the information on Country A and Country B.

In your best judgment, decide which of the following best describes a comparison of the countries:

- ○ Country A is much more democratic than Country B
- ○ Country A is slightly more democratic than Country B
- ○ Country B is slightly more democratic than Country A
- ○ Country B is much more democratic than Country A

Country A	Country B
• The 2002 National Assembly elections were the first conducted without a significant opposition boycott. Compaore`s Congress for Democracy and Progress (CDP) party won only 57 of 111 seats, compared with 101 in 1997. Compaore secured a third term as president in 2005, though it was shortened to five years by a 2000 constitutional amendment. A 2001 amendment had imposed a two-term limit for presidents, but the CDP argued that it was not retroactive. The country`s first municipal elections were held in 2006, with the CDP capturing nearly two-thirds of the local council seats. The CDP gained 16 seats in the 2007 National Assembly elections, for a total of 73, while the largest opposition party, the Alliance for Democracy (ADF-RDA) lost three seats, for a total of 14. In January 2009, Compaore pardoned Thibault Nana, leader of the opposition Democratic and Popular Rally (RDP) party, who had been sentenced to three years in prison in 2008 for allegedly orchestrating violent protests against high food prices that year.	• The April 2007 elections were marred by bloodshed and eyewitness reports of massive vote-rigging and fraud. At least 200 people were killed in election-related violence, with victims including police and several candidates. International and local election monitors were highly critical of the vote, and opposition parties refused to accept the results, which gave Yar`Adua 70 percent of the presidential ballots, Buhari 19 percent, Abubakar 8 percent, and the Progressive People`s Alliance candidate, Orji Uzor Kalu, 2 percent.In the parliamentary vote, the PDP took 87 out of 109 Senate seats and 263 out of 360 House seats. The ANPP took 14 Senate seats and 63 House seats, while the AC took 6 Senate seats and 30 House seats; the remainder went to three smaller parties. The PDP also led the state elections, taking 29 out of 36 governorships. The official results drew a raft of legal challenges that were adjudicated by election officials as well as the court system, with many appeals stretching well into 2008. In December 2008, the Supreme Court delivered its final ruling on the presidential contest, repudiating the opposition complaints and upholding Yar`Adua`s victory. Separately, in a rare instance of an opposition candidate unseating a PDP rival through the appeals system, an appeals court in November overturned the election of the Edo State governor based on `voting irregularities,` declaring the AC candidate the rightful governor. A February 2009 ruling annulled the gubernatorial victory of the PDP`s Segun Oni in Ekiti State, calling for a rerun of the 2007 vote. However, political violence and misconduct attributed to PDP operatives accompanied the April 2009 runoff between Oni and the AC`s Kayode Fayemi, and official results confirmed Oni as the winner.
• International monitors have judged the most recent presidential, municipal, and legislative elections to be generally free but not entirely fair, due to the ruling CDP`s privileged access to state resources and the media. President Blaise Compaore is currently serving his third term in office, and he is expected to seek another five-year term in 2010. The 111-seat National Assembly is unicameral, and members serve five-year terms. The legislature is independent, but subject to executive influence.	• According to the constitution, the president is elected by popular vote for no more than two four-year terms. Members of the bicameral National Assembly, consisting of the 109-seat Senate and the 360-seat House of Representatives, are elected for four-year terms. The Brussels-based International Crisis Group found that the general elections of April 2007, in the view of most voters and the many international observers alike, were the most poorly organized and massively rigged in the country`s history. Civil society organizations reported numerous, widespread incidents of political harassment and violence surrounding the elections in six neighboring states, with the majority committed by PDP supporters or criminal gangs acting on behalf of PDP politicians.
• The constitution guarantees the right to form political parties, and 13 parties are currently represented in the legislature. Opposition members have argued that 2004 revisions to the electoral code, which tripled the number of electoral districts, gave an undue advantage to larger parties, particularly the CDP. Some civil society groups have also criticized the 2009 electoral reforms, which established a gender quota and extended suffrage to citizens living abroad. The CDP has notably higher numbers of female members, and there are concerns that the overseas polling will be managed exclusively by embassies, with fewer monitoring opportunities for the opposition and civil society. Opposition parties remain weak; in the 2007 legislative elections, only two parties, the CDP and ADF-RDA, reached the 5 percent vote threshold needed to qualify for campaign financing. Another April 2009 electoral reform reduced that threshold to 3 percent of the vote so as to include a greater number of parties.	• Nearly 50 parties participated in the 2007 elections. The three major political parties are the ruling PDP; the ANPP, which is the largest opposition party and draws its strongest support from the Muslim north; and the AC, an opposition party formed from smaller groups ahead of the 2007 elections. Three other parties are represented in the federal legislature: the Progressive People`s Alliance, the Labour Party, and the Accord Party. Although political parties represent a wide array of policy positions and openly engage in debate, they continue to be marginalized by the PDP. Many opposition parties have argued that the Independent National Electoral Commission is effectively an extension of the PDP.

Figure 1: *An example of a task as it appears on Amazon Mechanical Turk comparing the levels of democracy between two countries based on qualitative paragraphs taken from Freedom House country reports.*

Read the information on Country E and Country F.

In your best judgment, decide which of the following best describes a comparison of the countries:

- ○ Country E is much more democratic than Country F
- ○ Country E is slightly more democratic than Country F
- ○ Country F is slightly more democratic than Country E
- ○ Country F is much more democratic than Country E

Country E

- Aliyev easily won a second term in the October 2008 presidential election, taking 89 percent of the vote amid 75 percent turnout, according to official results. Most of the political opposition chose to boycott the poll, citing barriers to meaningful media access and the overwhelming influence of administrative resources deployed by the YAP. In March 2009, a constitutional amendment that removed term limits for the president reportedly passed a referendum with more than 90 percent of the vote, allowing Aliyev to run again in 2013. The country`s constitution provides for a strong presidency, and the parliament, the 125-member Milli Majlis, exercises little or no independence from the executive branch. The president and members of parliament serve five-year terms, and a referendum held in March 2009 eliminated presidential term limits. Although the PACE indicated that the vote was transparent, well organized, and held in a peaceful atmosphere, it criticized the lack of public debate on the issue in the media.

- Elections since the early 1990s have been considered neither free nor fair by international observers. The most recent parliamentary elections, in 2005, were afflicted by extensive irregularities. The OSCE cited the interference of local authorities, disproportionate use of force to thwart rallies, arbitrary detentions, restrictive interpretations of campaign provisions and an unbalanced composition of election commissions.

- The 2008 presidential election, though largely peaceful, was no exception to this pattern. The OSCE`s monitoring report noted a number of problems, including a lack of robust competition, a lack of vibrant political discourse, and a restrictive media environment. President Ilham Aliyev said he would not campaign personally, but he reportedly stepped up his official activities and opened a number of infrastructure projects during the campaign period, garnering extensive coverage from the biased media. The OSCE also noted that public officials and YAP operatives worked cooperatively to mobilize support and increase turnout.

Country F

- Michelle Bachelet, previously the health and defense minister, was elected president in January 2006. Because of her party`s (Concertacion) strong performance in the 2005 legislative elections and a reform that eliminated the institution of unelected senators, she became the first president to govern with majorities in both houses of Congress. However, this advantage was relatively short-lived. In December 2007, the Christian Democratic Party suffered a serious split, causing six of its lawmakers to break away and end Concertacion`s majority.

- Elections are considered free and fair. The constitution, which took effect in 1981 and has been amended several times, currently calls for a president elected for a single four-year term, and a bicameral National Congress. The Senate`s 38 members serve eight-year terms, with half coming up for election every four years, and the 120-member Chamber of Deputies is elected for four years.

- In 2005, the Senate passed reforms that repealed some of the last vestiges of military rule, ending authoritarian curbs on the legislative branch and restoring the president`s right to remove top military commanders. The reform package included the abolition of the Senate`s nine unelected seats and reduced the presidential term from six years to four. In September 2009, a bill was introduced to repeal another relic of the former regime, the Copper Reserve Law, which obliged the state-owned copper producer Codelco to transfer 10 percent of its earnings to the military.

If you have any concerns about the tasks above, or found anything difficult or unclear, please comment below. Thanks for your assistance.

Submit

Figure 2: *An example of the control question asked as the third judgment in every task, which should be simple to answer, but more importantly, should be answered the same way if the same individual answers multiple tasks. Note also the comment box for feedback.*

6.1 Results

To gain a complete understanding of the performance of different sorting approaches, we tasked all possible pairs of countries nine times each for a total of 495 judgments ($9 \times 11(10)/2$). We posed these tasks in three available pools which MTurk easily allows restrictions to: Respondents from any location, Respondents from the US, and Master Categorizers. The first two pools are constructed by examining the IP address of the provider. The latter is a special category of individuals who have successfully completed a large number of "categorization" tasks.[7]

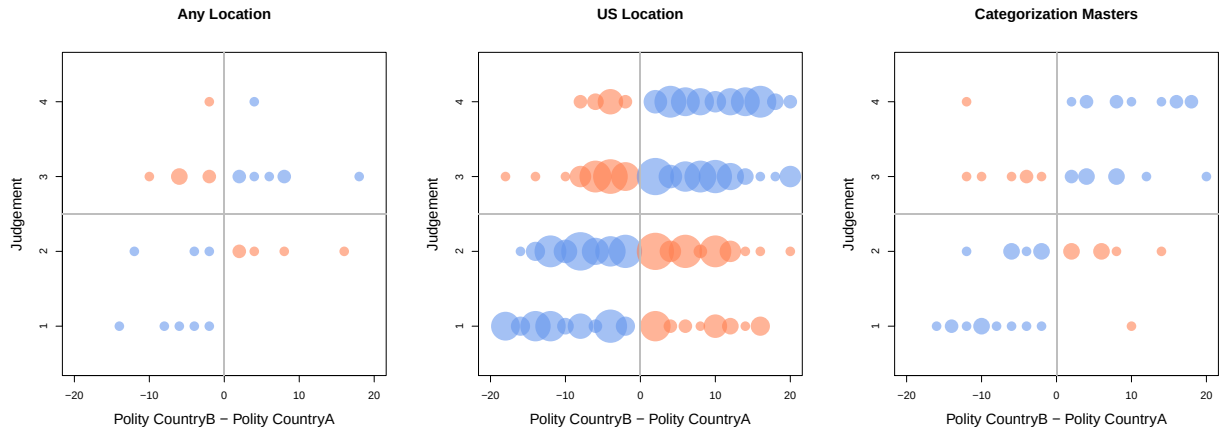

Figure 3: *Pairwise non-expert judgments, plotted against the difference in Polity score, for three populations of potential MTurk providers. Points in blue represent judgments that agree with the ordering Polity would offer, while red disagree. The area of each point represents the number of observations as each location and the percentage below captures the fraction of all individual level judgments that agree with the ranking from Polity scores.*

Figure 3 graphs the comparison of judgments across all responses compared to the differences of these countries in their Polity scores, in each of these three respondent pools. For any paired comparison, the x-axis measures the difference in the Polity score between Country A and Country B. The y-axis gives the judgment on the four point scale of the reader, where a code of 4 is "Country A is much more democratic than Country B" and a code of 1 is "Country B is much more democratic than Country A". Any judgment that falls in the first or fourth quadrant of the graph (upper-right or lower-left) is a judgment where the reader of the qualitative text makes a judgment that happens to agree with the ordering of the two countries in the Polity index. These are colored blue. Judgments from the text that run counter to the Polity index are in the second and third quadrants and colored red. Because of the discrete nature of both dimensions, many points will fall in exactly the same position on the graph, so the size (area) of each dot represents the number of individuals who made such a judgment. We see in the middle and right graph, that most of the area is blue, that is, most of the participants are making the same judgment that we

[7] Access to this pool requires a higher surcharge paid to Amazon of 20 percent of any approved payment. The payments themselves do not have to higher, but there is a smaller pool of individuals, and the competitive rate tends to be higher. Our rates, figured as an hourly rate were still higher than usual in this pool.

would have gotten if we had instead used the Polity scale. The few red observations are generally close to the vertical line, which represents a Polity difference of zero, so judgments that disagree with the relative ranking in Polity are generally only occurring for countries that are close in their Polity ranking.[8]

The actual rate of individual agreement, is not the key factor in determining how the aggregate, or "crowd-sourced," judgment will line up with the expert codings. This number could arise because all individuals agree with one another (all five answers are the same), but individuals see about a third of the cases opposite to how they line up in Polity. Or, the individuals could have some internal disagreement, but the average answer lines up exactly the same as the quantitatively derived estimates. In the extreme, it would be possible for the population agreement to be some tiny ϵ above 50 percent, and still have the average judgment, given enough individual responses, agree with the statistical estimates. Or in such an extreme, there could be no agreement at all, and they agree with the same probability that two random coin flips are the same. That is, 50.1 percent agreement could represent an exact match at the country-pair average to Polity, or no agreement whatsoever.

What is key then, is the fraction of all judgements, of some particular country-pair, that agree with the relative polity scores of those states. This information is presented in figure 4. The eleven states in our study are arranged from low Polity (Quatar) to high (Hungary). Each number in the figure represents the fraction of the estimates that stated the column country was more democratic than the row country, and thus gave a judgement that corresponded to the ordering the Polity score would give. Only the upper diagonal is presented, as the lower diagonal would contain no additional information. For ease of interpretation, where the value is greater than 0.5, the cells are colored in increasing bright green, while the cells where the average judgement does not correspond with the analytical Polity score are increasingly red.[9] We can see in the top right, where the strongest democracies are compared to the least democratic states, that the fractions are all consistently very high. In the bottom right, where democracies are compared against each other, the fractions are generally dropping, and there is some disagreement in the judgements as to which state is more democratic, although with two exceptions the average judgement aligns correctly with the Polity score. There are more disagreements with the Polity score in the top-left of the figure where the least democratic states are being compared to each other. This might be because relative levels of democratization are hard to judge in descriptions of less democratic states, or it might be that there is more than one dimension of authoritarian structure (personalist, single party, military) that make different authoritarian states more difficult to compare (Honaker and Wright 2013).

From all of these paired judgments, we computed the order of the states using the full count rule, and the Quicksort and Bubble Sort algorithms discussed in section 3.1, as well as a Mergesort. To reiterate, the eleven states are the objects for the sorting algorithms to order. The matrix in figure 4 served as the comparator, that is, everytime a pair objects needed a paired comparison to select a winner, the average value of all judgements was used; when this was greater than 0.5, the column state was determined more democratic, and when below 0.5, the row state was

[8] There are many more observations in the center graph. Given that the respondents from "Any Location" had lower agreement levels, and the "Master Categorizer" agreement had only slight increase but cost more and took substantially longer to collect due to a small pool, we focused on the ordinary US resident pool in later rounds of judgement collection. Also, we began our collection effort in the spring of 2012, but in the beginning of 2013 Amazon significantly restricted the ability of non-US responders to join or contribute to tasks.

[9] The astute reader will notice there are many cells with value 0.4, and many with value 0.6, and none at 0.5. There are nine judgements, so $4/9 = .444$ is rounded down, and $5/9 = .556$ is rounded up, thus 0.5 is not possible from 9 judgements.

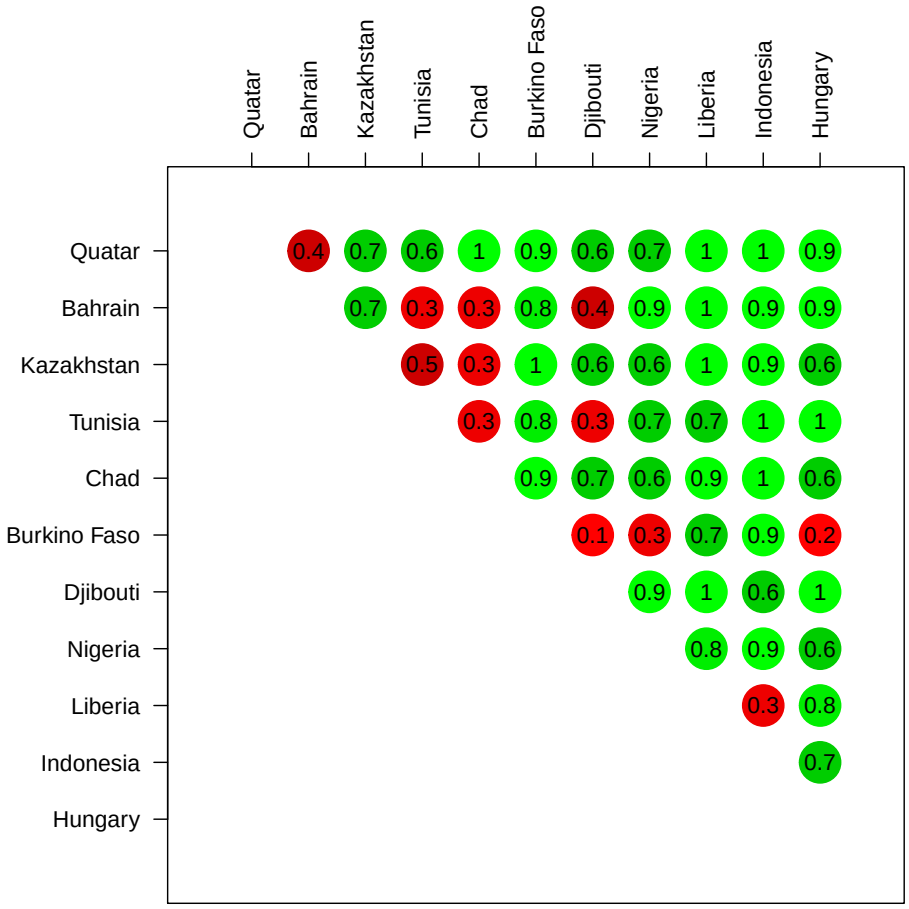

Figure 4: *Fraction of all pairwise judgements that agree with the ordering of the states in the Polity score, that is, the column state is judged more democratic than the row state. Cells in green show the judgements where the majority of respondents agree with the ordering that would come from the Polity score.*

considered more democratic. The initial order of the states was randomly shuffled, and then the order sorted by each algorithm, using only the particular paired comparisons called for by that algorithm. If the comparator was a strong ordering, then each of these algorithms would always result in the same final ordering of states, regardless of the initial shuffling of states, or the algorithm used. However, there are intransitivities present in the comparator we have from the average judgements, possibly due to actual intransitivity in judgement, or also likely due to the stochastic nature of the small sample size we have for each point. If we had hundreds of judgements for each country-pair, likely some of the cells presently below 0.5 would move above 0.5 (and possibly some of the those presently above 0.5 might also fall below that mark). With this intransitivity, the final ordering of the states will depend on the algorithm used, and the initial shuffling of the states. We ran each algorithm 1000 times on random initial permutations to order, and recorded the order that resulted, as well as the number of country-pair comparisons that were needed to develop the ordering.

We computed the correlation of each algorithm's distribution of rankings across these simulations with the Polity score, and the correlation matrix is shown in figure 5. The full count rule, which is the most expensive to compute as it uses all judgments of possible pairs, correlates with the Polity score at 0.85. To judge whether this is high or low, we compute the correlation of the Polity scale to the Freedom House scale for these same countries, for the same year; this is also 0.85. That is, our measure of democratization computed entirely with distributed judgments of non-experts, correlates with the Polity score at about the level as costly expert "gold standard" measures correlate with each other. We also implemented the more efficient algorithms described in section 3. The Bubble Sort, Quicksort and Mergesort algorithms, which use fewer of the available judgments, and thus would be cheaper to compute if we only collected the judgments they individually required, have lower levels of correlation with Polity, with Quicksort fractionally leading in performance. Bubble Sort, although simple and conceptually attractive, is generally looked down on in applied situations owing to the larger number of computations (and generally worse order of operation) in situations where comparisons are perfect. One useful feature in Bubble Sort for our application is that pairs that are "incorrectly" placed at some round in the sort are not permanently set (as in divide-and-conquer algorithms), but can be adjusted in future rounds. The fact that some pairs are repeatedly compared in the algorithm, costs computational time in conventional numeric settings, but in our application, we do not necessarily need to rerun pairwise comparisons among non-experts which have already been judged, so these repeated computational operations do not translate into increased costs or rounds of task on MTurk.

We note also, that all our measures correlate with Freedom House at a lower level than with Polity. At first glance, this is intriguing when we recall that we are using qualitative text, pulled from reports written by Freedom House, to give our participants. However, we pulled those paragraphs that focused on recent elections and parties, which are focuses of the Polity conception of democracy, whereas the Freedom House score has a greater focus on rights. We feel that the greater correlation of our measure based on the intent and focus of the qualitative sources, rather than the particular institute that wrote the qualitative sources, is an interesting form of face validity for our method.

The previous orderings are all algorithmic interpretations of the data. Figure 6 shows the estimated latent positions of the states in a democracy dimension, using the statistical model described in section 5. They are aligned along the x-axis in polity score, and the y-axis gives the estimated positions using all of the paired judgements

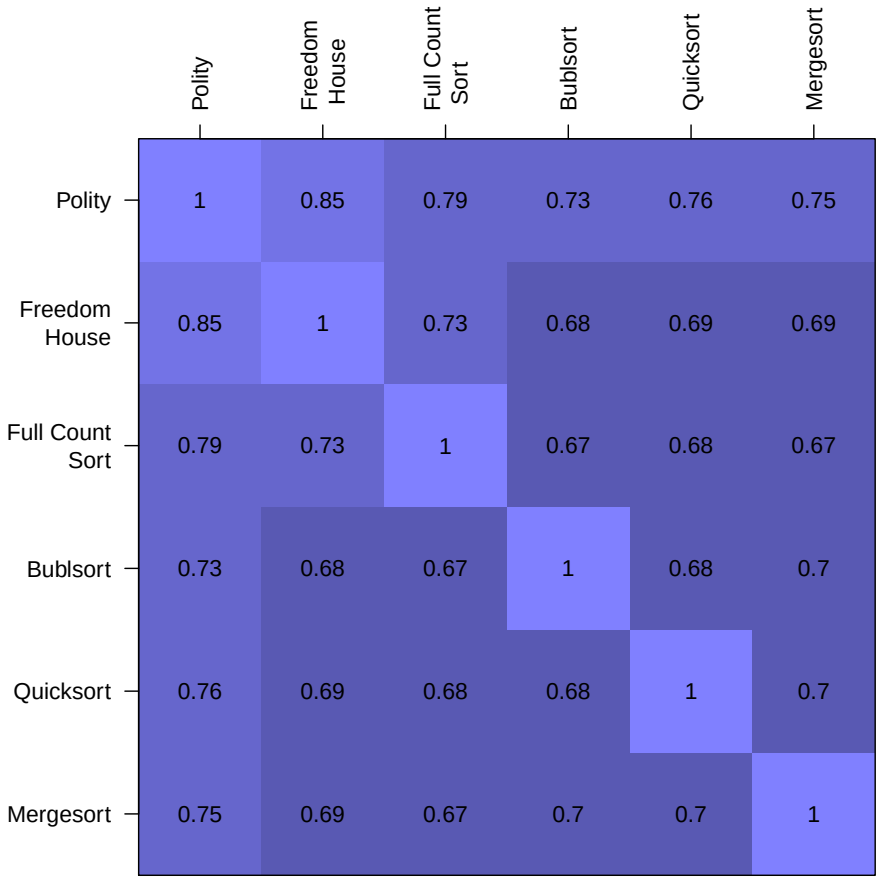

Figure 5: *Correlations between the Polity score, the alternate Freedom House measure of political freedom, and three rankings constructed using algorithms to sort using non-expert pairwise judgments.*

(as well as 80 percent confidence intervals). The states at either end are pegged parameters to identify the model, so have no uncertainty. The states' estimated positions generally trend upwards. Burkina Faso is most notably divergent from this pattern, which makes sense from what we saw in figure 4, as the average judgement placed it more democratic than all states with a lower polity score, but also more democratic than Djibouti, Nigeria and even Hungary, which have higher Polity scores than itself. The correlation of the estimated positions from the judgement data, to the Polity score, is now 0.91. This slightly outperforms the fullcount sort. While the fullcount sort uses the judgements on every possible pair of states, it only uses the information as to whether the average judgement is above or below 0.5. This statistical model uses all of the individual judgements, and leverages the information of the rate of disagreement; states that are far apart should have low levels of disagreement among judgements, whereas states that are close together should have high levels of disagreement. This additional information gives some leverage to the statistical model over the simplistic fullcount rule which we previously saw correlated at ****.

However, both approaches show the potential for non-expert pairwise judgements to construct an ordering of the latent space that well measures the underlying concept of democracy from textual sources, with no definition and coding of the quantitatively measurable attributes of democracy.

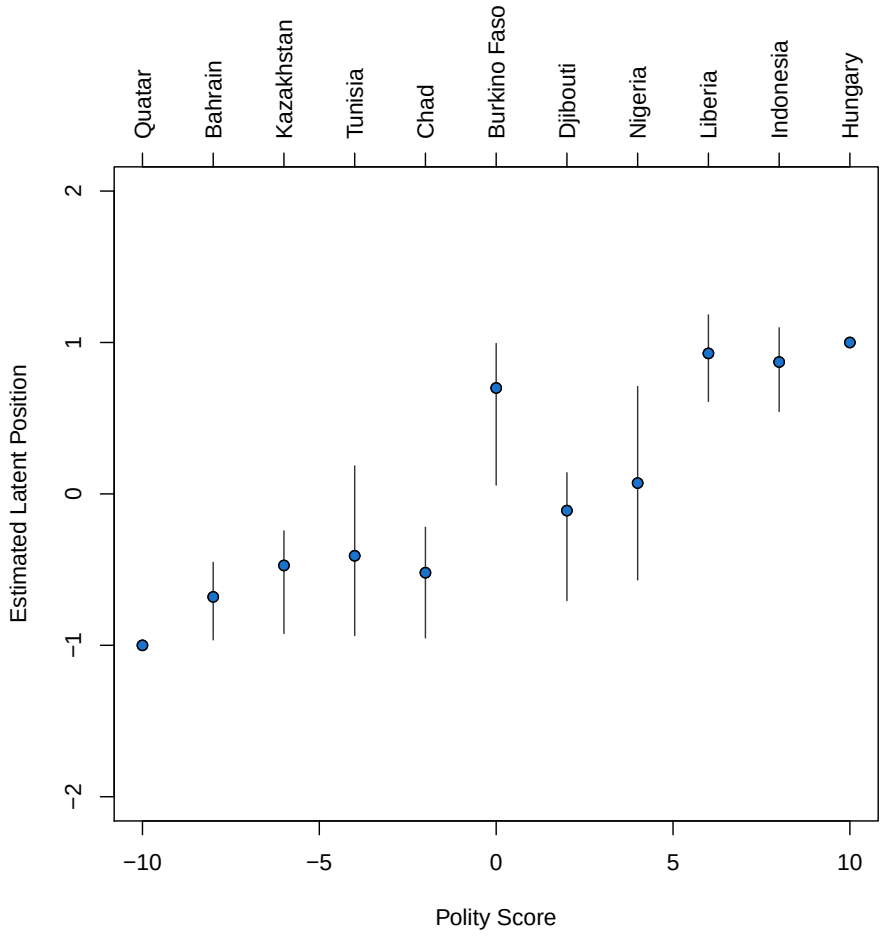

Figure 6: *Estimated latent democracy using the qualitative pairwise judgements and the Gaussian channel model of the previous section. The endpoints are pegged parameters, while the middle points trend positively in estimated position as we move across the Polity score of states. (Bootstrapped confidence intervals are shown at 80 percent.)*

7 Application: State Welfare Generosity

Our motivating application for this project, was to form a measure of the relative generosity of state policies towards welfare (Temporary Assistance to Needy Families, or hereafter, TANF) recipients, since the devolution of policy to the states under the welfare reform of the Clinton administration. Two major studies have attempted to measure this before. Fellows and Rowe (2004) constructed from theory a set of 12 questions that could be coded of each state's policy, and then summed these together to form a scale they called *flexibility*.[10] De Jong et al (2006) attempted a more data-driven, atheoretical approach by conducting a factor analysis of every quantitative variable that had been collected in a prior dataset distributed by the Urban Institute (Welfare Rules Database, 2011).

Our approach was somewhere between the two. Major criticisms of De Jong's approach point to the lack of substantive guidance in the data reduction model (Allard 2006, Soss et al 2006, Cadena et al 2006). We attempted to construct gold standard data by reading through all the rule descriptions in the sources to see all the features that needed measuring, as well as heavily leveraging the extensive literature on the history of welfare policy to understand the major groups or dimensions of policy that states define. In addition, we also constructed another dataset set that was an attempt to replicate Fellows and Rowe's coding and extend it to later years. These are both described in more detail below, but a complete description is given in Berkman et al (2013).

7.1 Quantitative Coding of State Policy Data

The Fellows and Rowe-style data were constructed using information from their 2004 article entitled "Politics and the New American Welfare States." They identify two primary components of state welfare policy: eligibility and flexibility. The eligibility index refers to "the rules that govern the initial eligibility of applicants" (pg. 365) and is comprised of 28 items. Higher scores reflect stricter eligibility laws. The flexibility index refers to the "flexibility of new welfare work requirements" (pg. 365) and is comprised of 12 items. Higher scores reflect greater flexibility of work requirements. The data for these 40 items come from the Welfare Rules Database (2011, hereafter WRD) maintained by the Urban Institute. To reconstruct their data, we use information from their 2004 article. We match each item in their indices to rules found in the WRD. We then extract this information from WRD and code it to mimic the Fellows and Rowe (F&R) data.[11]

In addition to the F&R-style data, we built our own separate multi-dimensional dataset of state welfare policies between the years 1996-2010. In the full dataset we focus on the requirements, exemptions, and sanctions put forth by a state. Requirements refer to actions a recipient must undertake in order to maintain benefits. Exemp-

[10] The authors also summed together 28 questions they constructed and coded that they saw as a separate dimension measuring *eligibility* or how broad was the pool of individuals that could be considered for TANF payments.

[11] Our reconstructed data still differ from the F&R data in a number of ways, which means we do not have a direct replication of their indices. We code policies from 2000-2009, while F&R examine policies from 1997-2000. Additionally, F&R never explicitly state which rules from WRD are used, which means there is some guesswork in matching items from the article to rules from WRD. In the most extreme case, we do not have any information for eligibility item #28 because F&R use ambiguous language to describe it, thereby making it impossible for us to match the item to a WRD rule.3 Finally, F&R do not explain how they handle missing data in their indices, despite its presence. In our experience, WRD does not have full information about states welfare policies. While the missingness is small, F&R report analyses as if they had full information, implicitly suggesting they imputed information, but without saying how. However, we consider our reconstruction a replication in spirit of their approach and their key indicators in the present time period.

tions refer to rules that allow groups of individuals to receive assistance without meeting specific requirements. Sanctions refer to the consequences, financial or otherwise, imposed by a state when a recipient fails to fulfill a requirement. Requirements, and the corresponding exemptions and sanctions, span several domains, including work requirements, child care requirements, school policies for dependent children, requirements for minor parents, immunization requirements, and requirements about contracts. In general, we code policies to reflect the demands of the requirement contained therein. Policies that require recipients to engage in numerous activities are more demanding than policies that require less recipient action. Policies that provide fewer work options are more demanding than policies that allow recipients to select from many work options.

Within each domain of a state policy, we also make distinctions between how the policy differs across types of recipients. For example, the work requirements within a single state may differ depending on the age and education of the recipient as well as the age of the recipient's child. Recipients with a high school degree may have fewer work options to select from than do recipients without a high school degree, for example. Put simply, different requirements may be different for different types of recipients and we capture this variation in our policy dimensions.

To then collapse all of this large number of coded variables into a scale, we applied an IRT model to the raw data. As the variables were grouped by substantive domains that we could identify from theory, and further broken up by classes of individuals that we witnessed being treated differently in the sources, we believe this data reduction technique is amenable to the core criticisms leveled at De Jong, even though in the abstract an IRT model is reasonably similar to a factor analysis model. What is key is that our variables are grouped in dimensions described in the literature on welfare policy, and constructed and structured to measure the facets of policy we observed in the text sources, rather than simply a dataset that absorbs or accumulates any available precoded quantitative variables.

7.2 Statistical Estimation of the Latent State Generosity

A brief graphical summary and visualization of our IRT model is presented in figure 7, which we explain shortly. This particular model is an IRT for all the work requirements set out by states for individuals who have a high school diploma or GED. We used the state policies in 2008 for the analysis that follows, simply because it was central in the time period we most interested in for the larger project. Owing to its theoretical roots, the common analogy, or working context for IRT models are datasets of educational testing, where we are attempting to measure the latent ability of test takers from their answers to a set of test questions. For any given question, high ability test takers should be more likely to correctly answer the question; some easy questions may be correctly answered by nearly everyone, some hard questions may have only be correctly answered by the very best. The IRT model simultaneously attempts to estimate the latent ability of every test taker, as well as the probability distribution that any question will be correctly answered by any individual of given ability. In our implementation of this model, "test questions" are observed policy attributes, and "latent ability" is the underlying level of generosity in welfare policy. The S-shaped curves at the bottom half of figure 7 each correspond to one measurable facet or variable or a particular policy, such as "Are English as a Second Language classes allowable activities to fulfill work requirements" or "Is providing child care for others an allowable activity." The horizontal x-axis represents the underlying latent dimension of generosity or leniency, and the height of any S-shaped curve gives the probability that any state will

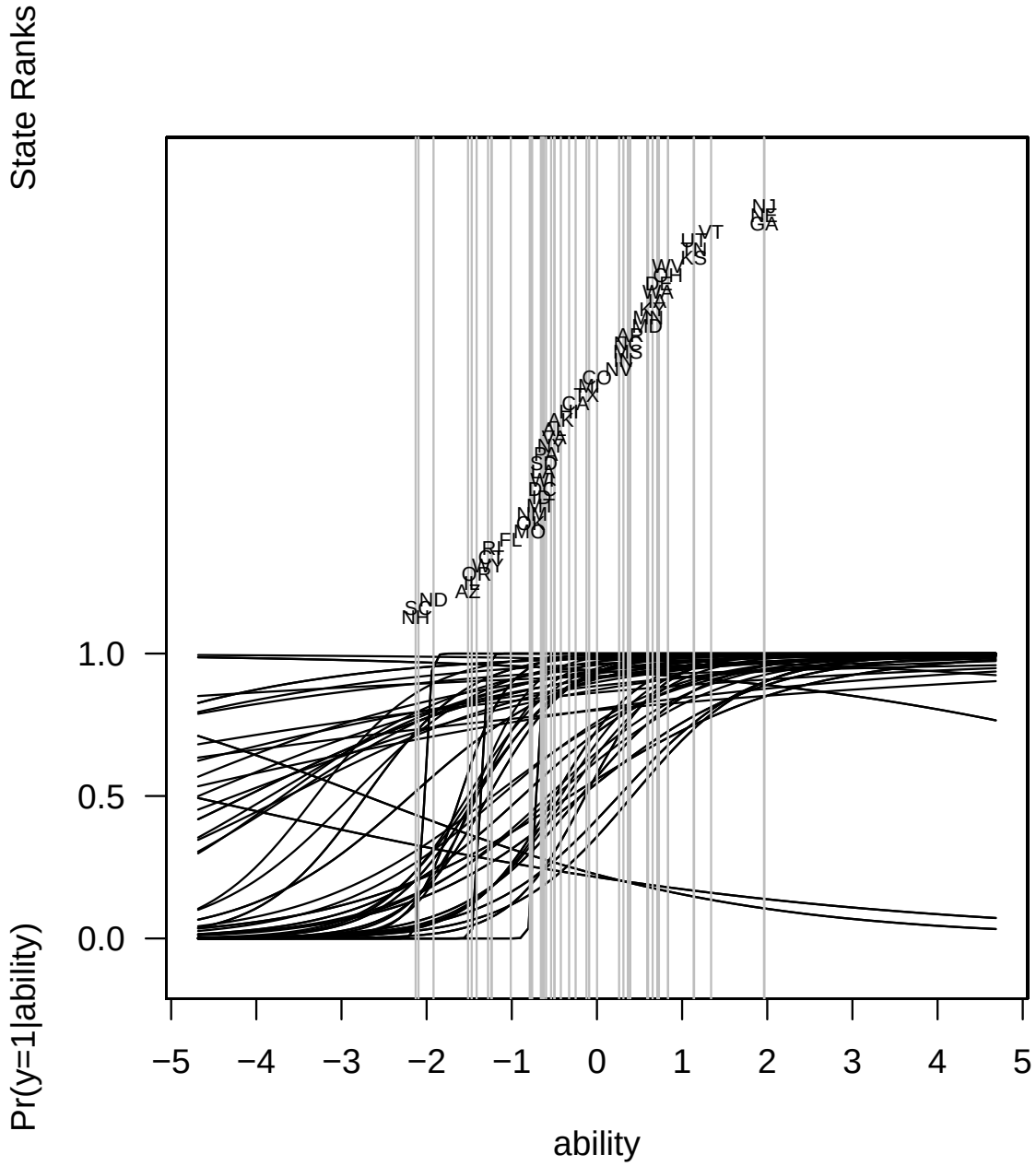

Figure 7: *A representation of the estimated latent dimension of state policy generosity or leniency in work requirements for individuals with high school diploma or GED. The x-axis is the latent dimension. The S-shaped curves give the probability that any quantitatively coded policy will be adopted in some state at some level of generosity, and the vertical lines represent the estimated generosity of each state. Lines are label with state abbreviations above, in descending order. A more detailed graph of the S-shaped curves is also found in appendix B.*

have that particular policy in their rules at any given level of generosity. With three exceptions, all of these S-shaped curves scale in the same direction, that is as states become more generous they are more likely to adopt each rule.

That they overwhelmingly scale in the same direction adds some face validity to the coding process (a more detailed figure which labels the exceptions can be found in appendix B). An S-shaped curve that looks like a step-function is a rule that has high discrimination in determining the position of state: states to the right of the jump will always have that rule, and states below that point will never have that rule. S-shaped curves are rules that have very little ability to determine or distinguish the level of state generosity: across all levels of generosity the probability of adopting that rule is similar or flat. While two of our rules resemble step functions, most of our rules smoothly transition. States at the high level of latent generosity have a very high probability of having almost all permissive policies; states at the low end of the latent dimension have a low probability of having each policy. The three rules that seem to move in the wrong direction are rather flat thus not too troubling.

The vertical gray lines represent the estimate of each state's latent level of generosity, given their rules and the estimated curves for each rule. States to the left have very few allowable ways to meet the requirements for TANF payments. In these states one simply has to work. States to the right have a large number of allowable activities that can count towards the required hours, and thus seem more lenient, or generous. Each vertical line is labeled with the state name, and as the state labels move from the bottom to the top, the latent level is increasing, thus reading through the state labels gives their relative ranks for their policies in 2008.

In conclusion, the estimated latent position of each state, noted by its vertical gray bar in the horizontal x-axis is scaled measure of generosity, specifically with respect to work requirements for individuals with some form of diploma for secondary education. This is the scaling of the states that we needed for our larger project.[12] Reading through the sources, identifying from the literature and the sources all of the measurable attributes in which states could differ, and then hand coding each state, was the primary research focus of four experts for eight months.

7.3 Scaling by Qualitative Judgments

In comparison to this exhaustive quantitative exercise, we describe our attempt to recover this coding using only pairwise judgments by non-experts. We sampled a subset of states, so that again we could compute all paired comparisons and evaluate how different sorting algorithms (which might require different sets of comparisons) perform. We sampled the following states, ordered from low to high latent leniency: Arizona, Florida, South Dakota, Virginia, Colorado, Minnesota, Delaware, Vermont, New Jersey. These were stratified to be roughly equally spaced along the latent dimension estimated in the previous section (here, 0.5 units in this arbitrary scaling). Participants were asked their judgement with the following prompt:[13]

> Public Policy Judgments
>
> For each of three cases below, compare the available information and decide which state policy is more generous or lenient.

[12] Specifically, we were interested in a number of questions, such as whether generosity could be predicted by various state-level demographics, and whether welfare officers, the "street level bureaucrats" that have to enforce these rules, behaved differently in states that were more or less generous.

[13] For recruitment, the title of the task was given as "Reading text to compare the level of generosity in pairs of state public policies" with an attached description of "Read a series of text extracts about public policy rules and categorize which state is more generous or lenient." and keywords for searching of "categorization, political, coding".

Under the current Temporary Assistance to Needy Families (TANF) rules, often known as 'welfare', each State is allowed to set its own requirements about what activities an individual must accomplish in order to receive assistance payments. The following paragraphs describe what requirements different States have in their rules. For example these activities might be that the individual has to work a certain number of hours a week, or engage in job training or other activities in order to continue to be eligible for assistance payments. States can have very different requirements for assistance.

Read the information describing two states, labeled State A and State B. These describe who each rule applies to, how many hours the rule requires, and what activities satisfy those hours.

In your best judgment, decide which of the following best describes a comparison of the two states:

. • State A is much more generous or lenient than State B
. • State A is slightly more generous or lenient than State B
. • State B is slightly more generous or lenient than State A
. • State B is much more generous or lenient than State A

Average time per assignment was 7 minutes which at the same payment of 2 (US) per assignment is equivalent to an hourly wage just above 17 (US). A subset of participants were paid 1 (US) so as to allow an examination of response quality to rate of payment. As with the democracy example, each assignment consisted of two comparisons from the sampled states, and a control comparison selected to be straightforward and simple. This was constructed using slightly simplified text from the Urban Institute descriptions. Again, the key purpose of the control was not to evaluate ability, but to look for participants who were answering multiple assignments randomly to harvest payments, and thus answering the same, simple, repeated question in an inconsistent fashion. These individuals, of whom there were two, were eliminated from the data and their tasks resubmitted for evaluation.[14] As this had proved successful in the democracy example, we limited participants to those with US IP addresses and moderate experience (greater than 50 previously successful tasks, and greater than a 95 percent acceptance rate of submitted tasks).

[14] Notably, these individuals also spent far less time on each assignment than others in the survey.

Public Policy Judgements

For each of three cases below, compare the available information and decide which state policy is more generous or lenient.

Under the current Temporary Assistance to Needy Families (TANF) rules, often known as *'welfare'*, each State is allowed to set its own requirements about what activities an individual must accomplish in order to receive assistance payments. The following paragraphs describe what requirements different States have in their rules. For example these activities might be that the individual has to work a certain number of hours a week, or engage in job training or other activities in order to continue to be eligible for assistance payments. States can have very different requirements for assistance.

Read the information describing two states, labelled State A and State B. These describe **who** each rule applies to, **how many hours** the rule requires, and **what activities** satisfy those hours.

In your best judgment, decide which of the following best describes a comparison of the two states:

- State A is much more generous or lenient than State B
- State A is slightly more generous or lenient than State B
- State B is slightly more generous or lenient than State A
- State B is much more generous or lenient than State A

State A	State B
• **Who:** Non-exempt recipients 18 years of age and older with a high school diploma or a GED. **How Many Hours:** 30 for recipients with a child age 6 or older; 20 for recipients with a child under the age of 6. **What Counts:** The following activities may count towards the first 15 required hours: Job skills training, Job search, On-the-job training, Unsubsidized job, Work supplement/subsidized job, Self-employment, Child care provider for others, Community service. The following activities may count toward the remaining required hours: Post-secondary education	• **Who:** (1) Non-exempt recipients under age 20 not pursuing a high school degree/GED and (2) Non-exempt recipients age 20 and older **How Many Hours:** 31; those who are participating exclusively in job search must do so for a minimum of 30 hours per week. Single parent families with children under six years old are only required to participate for a minimum of 21 hours. The number of hours required for victims of family violence are determined based on the circumstances of the family. **What Counts:** The following activities may count towards the required hours: English as 2nd language, Post-secondary education, Job skills training, Job readiness activities, Job development and placement, Job search, On-the-job training, Unsubsidized job, Work supplement/subsidized job, CWEP/AWEP, Self-employment, Child care provider for others, Counseling, Life skills training, Community service. Participants who participate in work activities for at least 20 hours a week are encouraged to participate in education and training activities.
•	• **Who:** Recipients who are (1) ill or incapacitated, (2) caring for an ill or incapacitated family member, (3) pregnant if the pregnancy has resulted in an incapacity that prevents the woman from obtaining or retaining employment, or (4) age 60 or over. **How Many Hours:** Determined on a case-by-case basic. Recipients are expected to work as much as they are able, as determined by a qualified medical professional, in whatever activities will move them toward self-sufficiency. **What Counts:** The following activities may count towards the required hours: Basic or remedial education, High school/GED, English as 2nd language, Post-secondary education, Job skills training, Job readiness activities, Job development and placement, Job search, On-the-job training, Unsubsidized job, Work supplement/subsidized job, CWEP/AWEP, Self-employment, Child care provider for others, Counseling, Life skills training, Community service, and Others.
•	•

Figure 8: *An example of the first paired comparison asked of participants. A short prompt is included in this first question, and omitted hereafter. As with the democracy example participants are given a four point scale which is collapsed into a dichotomy for analysis.*

Read the information on State C and State D. These describe **who** each rule applies to, **how many hours** the rule requires, and **what activities** satisfy those hours.

In your best judgment, decide which of the following best describes a comparison of the two States:

- ○ State C is much more generous or lenient than State D
- ○ State C is slightly more generous or lenient than State D
- ○ State D is slightly more generous or lenient than State C
- ○ State D is much more generous or lenient than State C

State C	State D
• **Who:** All non-exempt recipients **How Many Hours:** Minimum of 22 hours per week (maximum 40 hours per week) **What Counts:** The following activities may count towards the required hours: Job readiness activities, Job search.	• **Who:** All non-exempt, unemployed recipients. **How Many Hours:** Core hours are equal to the sum of the welfare and food stamp benefits divided by the state minimum wage, up to a maximum of 30 hours a week. Total hours are based on grant amounts regardless of sanctions, and are calculated based on what the total grant would have been without the sanction. 10 hours of approved employment-related activities are required in addition to those hours. **What Counts:** The following activities may count towards the required core hours: CWEP/AWEP. The following activities may count towards the remaining 10 hours of the requirement: English as 2nd language, Job skills training, Job readiness activities, Job search, On-the-job training, Unsubsidized employment, Work supplement/subsidized job, Child care provision for others, Counseling, Life skills training, Community Service, and Other.
• **Who:** Non-exempt recipients under age 20 **How Many Hours:** Minimum 20 hours per week. Determined on a case-by-case basis by the caseworker. **What Counts:** The following activities may count towards the required hours: English as 2nd language, Post-secondary education, Job skills training (see note below), Job readiness activities, Job search, On-the-job training, Work supplement/subsidized job, unsubsidized job, CWEP/AWEP, Child care provider for others, Counseling, Life skills training, Community service (see note below). The following activities are countable (i.e., count towards the federal participation rate): basic or remedial education, high school/GED, ESL, and job skills training (see note below), job readiness, job search, on-the-job training, unsubsidized job, work supplement/subsidized job, CWEP/AWEP	• **Who:** All non-exempt, employed recipients. **How Many Hours:** 30 (20 if individual has a child under age 6). **What Counts:** The following activities may count towards the required hours: unsubsidized employment.
• **Who:** Non-exempt recipients over 20 years old **How Many Hours:** Unknown **What Counts:** The following activities may count towards the required hours: Post-secondary education, Job skills training, Job readiness activities, Job search, On-the-job training, Unsubsidized job, Work supplement/subsidized job, CWEP/AWEP, Child care provider for others, Community service.	• **Who:** Non-exempt full-time students who do not hold a baccalaureate degree **How Many Hours:** Combination of credit hours and work hours must equal at least 20 hours per week. **What Counts:** The following activities may count toward the requirement: Post-secondary education, Job skills training and Other.

Figure 9: *An example of one of the second paired comparisons, which now has a shorter prompt.*

Read the information on State E and State F. These describe **who** each rule applies to, **how many hours** the rule requires, and **what activities** satisfy those hours.

In your best judgment, decide which of the following best describes a comparison of the two States:

- ○ State E is much more generous or lenient than State F
- ○ State E is slightly more generous or lenient than State F
- ○ State F is slightly more generous or lenient than State E
- ○ State F is much more generous or lenient than State E

State E

- **Who:** Non-exempt recipients
  How Many Hours: 40 (field supervisors may approve a reduction to 30 hours if they deem appropriate)
  What Counts: The following activities may count toward the first 20 hours of the requirement: job skills training, job readiness activities, job search, On-the-Job-Training, unsubsidized employment, work supplementation/subsidized employment, and CWEP/AWEP. In addition to these activities, the following activities may count toward the remaining hours of the requirement: job skills training

State F

- **Who:** Non-exempt recipients
  How Many Hours: 30 hours, 20 hours if caring for a child between the age of 12 weeks and 6 years.
  What Counts: At least 20 hours per week must come from participation in the following core activities: Post-secondary education, Job skills training, Job readiness activities, Job search, On-the-job training, Work supplement/subsidized job, Self-employment, Life skills training, Unsubsidized employment, CWEP/AWEP, Community service, Counseling, and Child care provider for an another. In addition, up to 10 of the required hours can come from participation in: Basic or remedial education, High school/GED, and English as 2nd language.

If you have any concerns about the tasks above, or found anything difficult or unclear, please comment below. Thanks for your assistance.

Submit

Figure 10: *The third question asked is simply a control, which has no value to the algorithm, but all respondents should agree on (and respondents answering multiple questions should give the same answer to), and is useful for identifying potential respondents clicking random answers to quickly accumulate payments. Note also the comment box for feedback.*

7.4 Results

Every pair of states therefore, has a judgment on this four point scale from non-experts directly addressing the qualitative data, and also has a difference in our latent space estimated using IRT on very many variables developed after reading all the sources and coded by experts. We repeated each comparison among the 9 states, 5 times for a total of 180 judgments.[15] A comparison is presented in figure 11. As before, any individual judgment of a pair of states that agrees with the IRT model, that is, sees the state with the higher latent dimension as having greater generosity, would be a point in the first or fourth quadrants, and shaded blue. Because of the discrete nature of the data, multiple observations would be plotted in the same location, so the size of each plot point corresponds to the number of observations at that location. An average of 63.3 percent of individual responses were in the same direction as the latent evaluations. This is a lower fraction than in the democracy example, which seems to acknowledge this is a more difficult, and less clearly defined task.

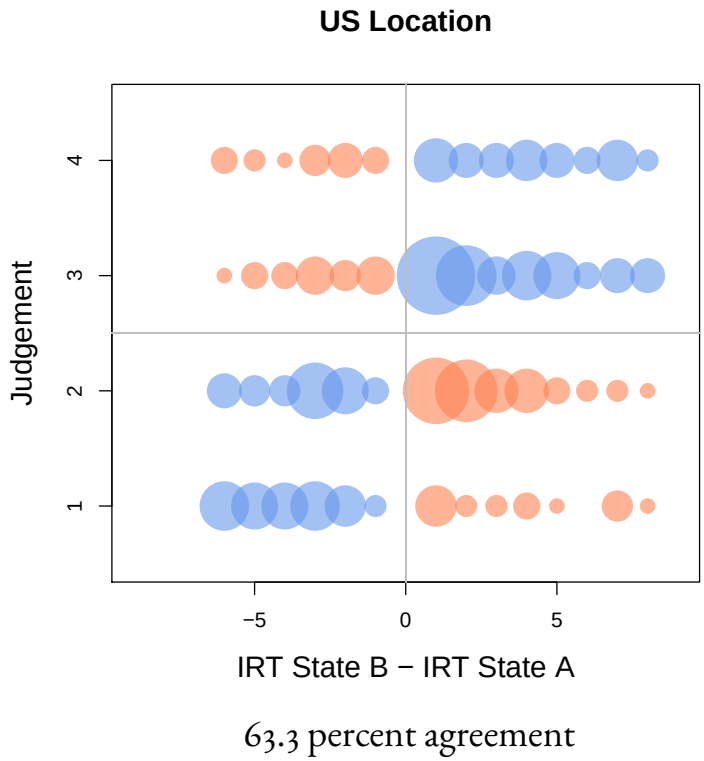

Figure 11: *Pairwise nonexpert judgments, plotted against the difference in IRT model estimates. Points in blue (63.3percent) represent judgments that agree with the ordering IRT would offer, while red (36.7 percent) disagree. The area of each point represents the number of observations as each location.*

The actual rate of individual agreement, is not the key factor in determining how the aggregate, or "crowd-

[15] $(5 \times 9 \times 8/2)$ One notable point is that not only did these 180 judgments cost simply 180 (US), but they were completed in under two hours.

sourced," judgment will line up with the statistical estimates. This number could arise because all individuals agree with one another (all five answers are the same), but individuals see about a third of the cases opposite to how they line up in the latent space. Or, the individuals could have some internal disagreement, but the average answer lines up exactly the same as the statistically derived estimates.

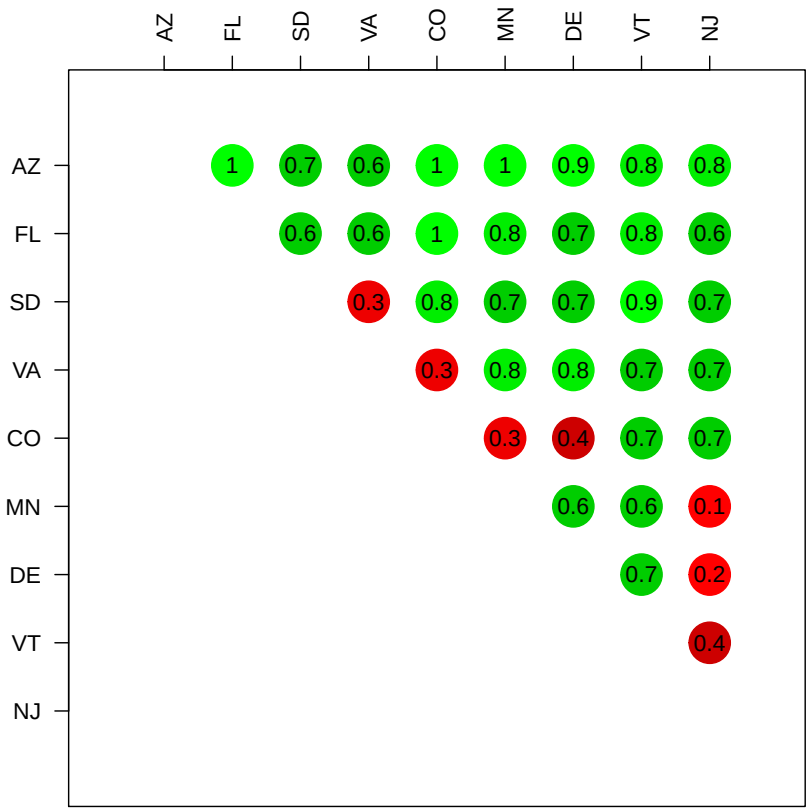

Figure 12: *A visual summary of all of the average judgments between all pairs of states. States are aligned in order of latent generosity as measured by the IRT model. Each number represents the fraction (of five) judgments that agree with the IRT ranking, that that column state is more generous than the row state. When the average judgment agrees (disagrees) with the IRT model, that point in the matrix is colored green (red).*

What is key then, is to measure how the average judgment lines up with the relative positions in the IRT estimated latent space. Figure 12 provides exactly this information represented in a matrix visualization. The states are ordered low to high, AZ to NJ respectively, along each axis. Each point describes the fraction of participants that judge the column state more generous than the row state. When this value is above 0.5, the majority of participants give judgments in the same direction as the IRT statistical model run on the quantitatively coded dataset. These are coded also colored in green. Red points are those for which the fraction of judgments for that pair that line up

with the statistical model is below 0.5, thus the majority disagree with the latent ordering. For parsimony, only the upper diagonal is shown, as all possible information is contained here. *If the average non-expert judgment always agreed with statistical model, every point should be colored green.*

We can see a number of state specific results from this graph. In the uppermost rows, every point is green, therefore the average pairwise comparisons between either Arizona or Florida, and all other states agree with the orderings from the IRT estimates, and always places these as less generous. Reading down the rightmost column, we see that New Jersey is ranked by our judgments as more generous than AZ, FL, SD, VA and CO, but less generous than Minnesota, Delaware and Vermont disagreeing in those three states with the ordering from the IRT model. Other than NJ, most of the other disagreements between the aggregation of our judgments, and the IRT model are along the diagonal, so some states that have small differences in the IRT model are ranked in a manner that disagree with the IRT model. If we assume the IRT model is the true gold standard data, then this might mean that our coders have difficulty with states that are close. Remember, states here were stratified to be equally spaced, so responders are having trouble with pairwise comparisons where the difference in the latent space is 1/8th the range of the measure (or in this example, approximately 1/3rd of a standard deviation apart). Notice also, that most of the averages along the diagonal are between 0.3 to 0.7 as a fraction of the judgments agreeing with the IRT model's direction, or specifically, between 3 to 6 individuals out of 9. Thus some of these errors we see on the diagonal might simply be the stochastic product of small samples. This suggests observations that appear to be ranked close to each other should have a greater number of participant judges than states which are easily seen to be greatly separated. If we had 10 or 20 judgments here, we might see a small fraction greater than 0.5 judging in the same direction as the IRT model.

Finally, figure 13 shows the correlation of the final rankings that are generated by each of our sorting algorithms. The IRT gold standard data is in the first row or column. We constructed an alternate statistical measure by taking all of the variables that went into the IRT model and extracting the first principal component. This is another reasonable data reduction technique, and mirrors in spirit the factor analysis model estimated by De Jong et al (2006). We also included our replication of the Fellows and Rowe coding. The Full Count Sort is in the fourth row or column.[16] Again, as in the democracy example, the correlation with the gold standard data from our aggregation of non-expert pairwise judgments is very high (0.88). This is much higher than the scale developed by Fellows and Rowe, for example, which is a very simple 13 point scale that experts devised as a proxy simple measure, has a much lower correlation to the gold standard than our non-experts achieve. However it is also quite close to the difference in correlation that results some simple changes in model specification, as the principal component analysis on the exact same variables can be seen to be correlated at 0.91. Again, the Bubble Sort and Quicksort algorithms, which don't use all of the judgments shown in figure 12 and thus require less comparisons, also have high correlations with the IRT latent estimates, but again, not as good performance as the more costly Full Count Sort.

[16] With the visualization presented in figure 12 we can now think of the Full Count as the count along any column (or row) of the number of red points below (above) the diagonal, and the number of green points above (below) the diagonal.

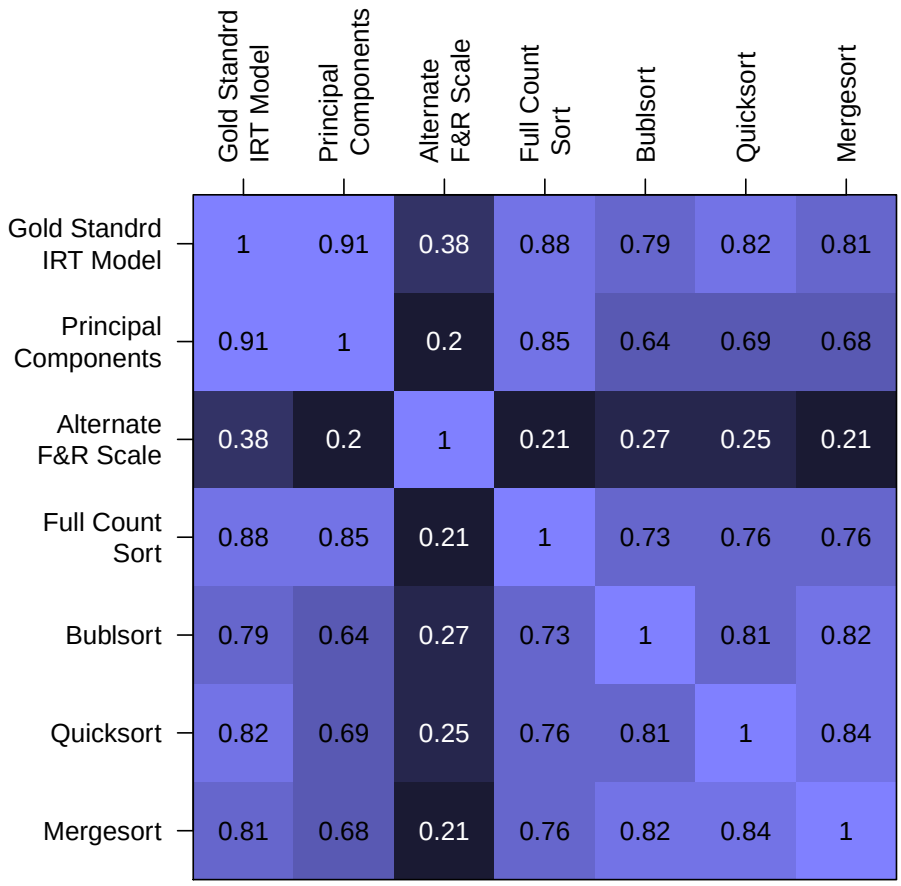

Figure 13: *Correlations of the gold standard IRT estimates, with a conventional and simple expert proxy (F&R) and our three sorting algorithms using non-expert pairwise judgments.*

7.5 Algorithmic Performance

There are several pointers we can learn from the behaviour and performance of the sorting algorithms in these applications. First, and most importantly, pairwise comparisons generated from non-experts have the potential to reveal the latent structure of qualitative data sources. We spent many months of many experts time generating gold standard measurements following the best practices of quantitative methods pooling substantive and statistical experts. We were able to create a reasonable proxy for this with no experts, no statistical models, and a tiny budget in under two hours. Human intelligence has a capacity for judging qualitative sources that is not well harnessed and underutilized, and a proficient way to organize or scale qualitative sources. The best results were obtained with the Full Count sort that requires all $n(n-1)/2$ unique pairwise comparisons, or with a statistical model utilizing this same observed information. As n grows this may be infeasible. The three sorting algorithms examined utilize only a subset of all comparisons. Below are presented the total number of unique comparisons required in each simulation

in each of our applications. The histograms for the three algorithms are presented interwoven with each other for ease of comparison, with each algorithm a different color. Quicksort required the fewest number of comparisons on average, although Mergesort was quite similar in its demands; Mergesort had very small performance advantages in the observed correlations with the gold standard data. Bubble sort has a much higher mean number of comparisons required, and much higher variance, as well as worse performance in correlation to the sorted orderings to the gold standard data in the applications.

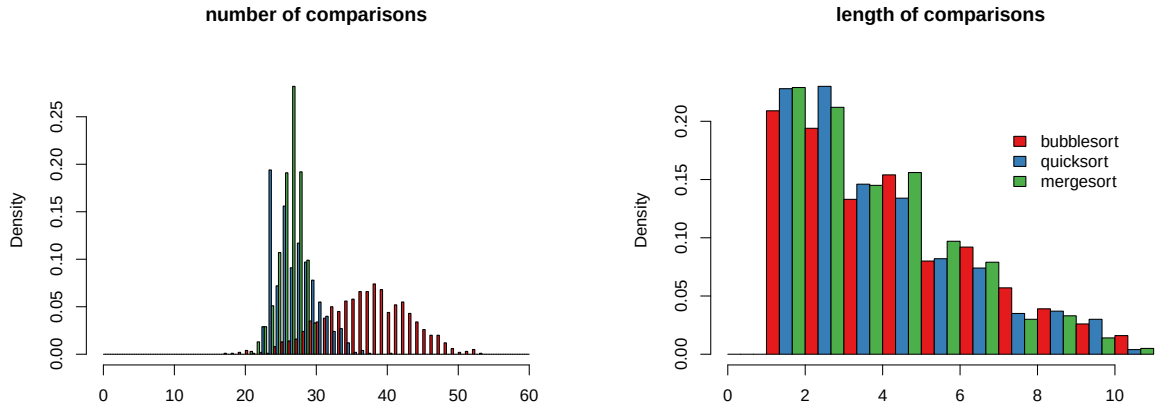

Figure 14: *Number of pairwise comparisons, and their distance in the gold standard data, for the ordering of latent democracy.*

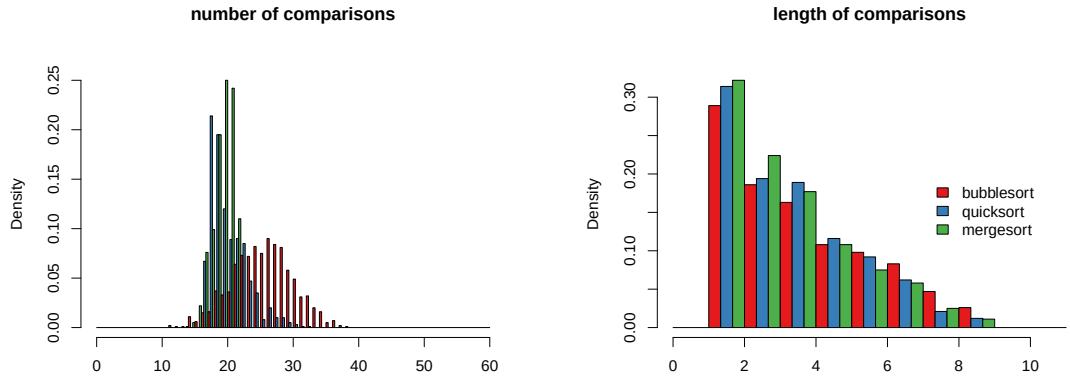

Figure 15: *Number of pairwise comparisons, and their distance in the gold standard data, for the ordering of TANF generosity.*

To the right of each figure are histograms showing the distance, as measured by the gold standard data, between each pair of objects that underwent comparison in all the runs of all the shufflings of the data. It is likely that comparisons between objects that are distant in the latent space (Quatar and Hungary for example) are less infor-

mationally useful that between countries that are close, especially if the orderings from the sorting algorithm are to be refined in a second stage by a statistical model on all observed comparisons. All three algorithms have similar distributions across the distances between compared pairs, with the majority of comparisons taking place between close observations (which would also be true of completely random pairings–there are only a limited number of observations that are able to be far apart from each other).

8 Conclusion

In quantitative research projects, we often face the need to measure large, complex, vague but important concepts from qualitative sources. The steps necessary to create such a measure through traditional measurement methods can be extremely demanding of resources and expert knowledge. We demonstrate the surprisingly powerful performance of non-expert human intelligence, when given sufficiently small structured tasks, set out as pairwise comparisons, and show how well-understood sorting algorithms can take these human judgments, select the comparisons to make, and uncover the latent ordering of the qualitative sources without any identification or quantification of the attributes of the objects. The results are very cheap, incredibly fast measures that correlate with gold standard statistical methods as well as alternate statistical models scale with each other.

References

Astrachan, Owen. (2003). "Bubble Sort: An Archaeological Algorithmic Analysis." *SIGCSE '03 Proceedings of the 34th SIGCSE technical symposium on Computer science education* 1-5.

Allard, Scott W. (2006). "A Starting Foul in the Study of Race to the Bottom: A Comment on 'Measuring State TANF Policy Variations and Change After Reform' " *Social Science Quarterly* 86:782-790.

Berkman, Michael, James Honaker, Christopher Ojeda, Eric Plutzer. (2013). "Measuring State TANF Policy" *Paper presented at the State Politics and Policy Annual Meeting.* Iowa City; May.

Benoit, Kenneth, Drew Conway, Michael Laver and Slava Mikhaylov. (2013). "Crowd-sourced Data Coding for the Social Sciences: Massive Non-expert Human Coding of Political Texts" *Paper presented at the Annual conference of the Midwest Political Sciece Association.* Chicago; April.

Berinsky, Huber, Lenz. (2012). "Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research." *Forthcoming, Political Analysis.*

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). "Amazonffs Mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*.

Casper, Gretchen and Cladiu Tufis. 2003. "Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets." *Political Analysis* 11(2):196ff203

Callison-Burch, Dredze. (2010) "Creating Speech and Language Data With Amazon's Mechanical Turk." *Proceedings of the NAACL HLT 2010 Workshop, Association for Computational Linguistics*

Cadena, Brian, Sheldon Danziger ad Kristin Seefeldt. (2006). "Measuring State Welfare Policy Changes: Why Don't They Explain Caseload and Employment Outcomes?" *Social Science Quarterly* 86:808-817.

Cover, Thomas M. and Joy A. Thomas. (1991) *Elements of Information Theory*. Wiley: New York.

De Jong, Gordon F., Deborah R. Graefe, and Shelley K. Irving, and Tanja St. Pierre. (2006) "Measuring State TANF Policy Variation and Change After Reform." *Social Science Quarterly* 86:755-781.

D'Orazio, Vito. (2013) "Big Data and the Art of Measurement". *Working Paper.*

Fellowes, Matthew C., and Gretchen Rowe. (2004). "Politics and the New American Welfare State." *American Journal of Political Science* 48(April): 362ff73.

H.H. Goldstine and J. von Neumann. (1948) "Planning and coding of problems for an electronic computing instrument", Part II, Volume 2, reprinted in John yon Neumann Collected Works, Volume V: Design of Computers, Theory of Automata and Numerical Analysis, Pergamon Press, Oxford, England, 1963, pp. 152-214.

Hoare, CAR (1962). "Quicksort" *The Computer Journal* 5 (1): 10-16.

Honaker, James and Joseph Wright. "The Structure of Autocratic Rule." *Paper presented at the Meetings of the European Political Science Association.* Barcelona; June.

Horton, J.J., Rand, D.G., & Zeckhauser, R.J. (2011). "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics*, 4,399-42

Knuth, Donald. (1998).*Art of Computer Programming, Volume 3: Sorting and Searching (2nd Edition).* Addison-Wesley Professional.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making.*

Pemstein, Daniel, Stephen A. Meserve and James Melton. (2010). "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426ff449.

Snow, O'Connor, Jurafsky & Ng (2008). "Cheap and fast - but is it good" *EMNLP Proceedings*

Soss, Joe, Meghan Condon, Matthew Holleque and Amber Wichowsky "The Illusion of Technique: How Method-Driven Research Leads Welfare Scholarship Astray." *Social Science Quarterly* 86:798-807.

"Welfare Rules Database" (2011), http://hdl.handle.net/1902.1/12907 V2 [Version]

A Detail on sorting algorithms with provided comparator matrix

```
# Do bubblesort by while loop
bubblesort<-function(m){
    k<-nrow(m)
    order<-1:k
    flag<-FALSE
    count<-0
    while(!flag){
        count<-count1+
        cat(count,"")
        tempflag<-TRUE
        for(i in 1:(k-1)){
            if (m[ order[i] , order[i1]] ¿ 0.5)+
                hold<-order[i]
                order[i]<-order[i1]+
                order[i1]¡-hold+
                tempflag<-FALSE
            }
        }
        flag<-tempflag
    }
    cat("\n")
    return(order)
}

# Do quicksort entirely by recursion
quicksort<-function(names,m,p2s=FALSE){
    k<-nrow(m)
    if(p2s) cat("k",k,"\n")
    order<-1:k
    if(k>1){
    flag<-rep(FALSE,k)
    a<-sample(order,1)
    for(i in 1:k){
        if(i != a){  # Don't compare to self
            flag[i]<-(m[ order[i] , order[a]] > 0.5)
        }
    }
    if(p2s) print(flag)
    if(p2s) cat("a",a,"\n")
    lowerorder<-order[-a][!flag[-a]]
    upperorder<-order[flag]
        if(length(lowerorder)>1)  {
            lower<-quicksort(names=names[lowerorder],m=m[lowerorder,lowerorder])
        }else{
            lower<-names[lowerorder]
        }
        if(length(upperorder)>1) {
            upper<-quicksort(names=names[upperorder],m=m[upperorder,upperorder])
        }else{
            upper<-names[upperorder]
        }
    order<-c(lower,names[a],upper)
    }
    if(p2s) cat("order","l",lower,"a",names[a],"upper",upper,"\n")
    return(order)
}
```

B Detail on IRT model curves

Figure 16 provides a detailed view of all the response curves for all items in the IRT model presented in figure 7. color coded in the center plot, where letters correspond to an abbreviated codebook name and numbers further distinguish whether that rule applies to one of 8 possible classes individuals (subdividing education, age of children and whether the individual is a minor). Items that scale in the same direction are on the left, and the items on the right are the few (relatively flat) curves that scale in the incorrect direction. These are HSL:"High school attendance and or work towards a GED are allowable activities," JRP:"Job readiness activities are allowable activities," and JDP:"Job development and job placement are allowable activities."

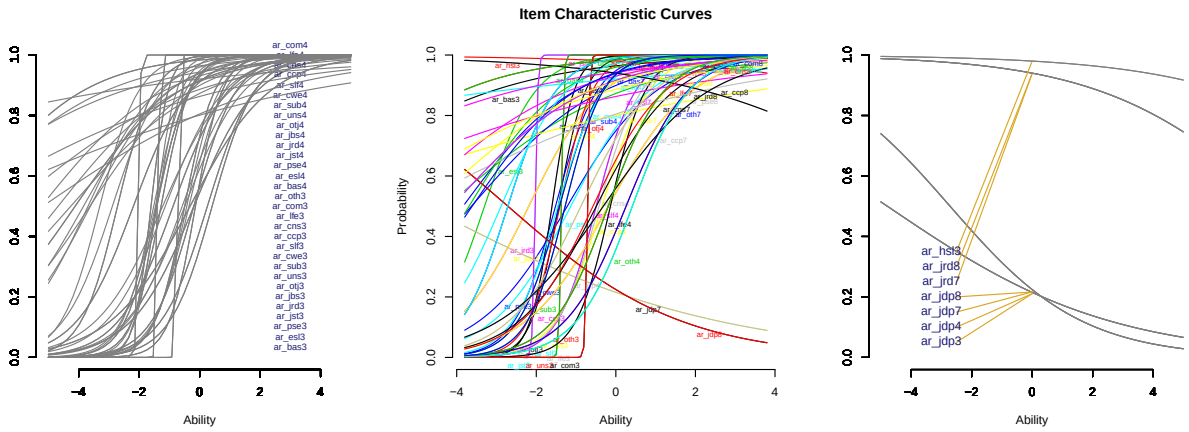

Figure 16: *Detailed view of all response curves for all items in the IRT model presented in figure 7, separated by direction.*