Credible Comparisons Using Interpersonally Incomparable Data: Nonparametric Scales with Anchoring Vignettes

Jonathan Wand Stanford University

Comparisons of individuals based on their selections from an ordinal scale traditionally assume that all respondents interpret subjective scale categories in exactly the same way. Anchoring vignettes have been proposed as a method to replace this homogeneity assumption with individual-specific data about how each respondent uses the ordinal scale. However, improving interpersonal comparisons with anchoring vignettes also requires a new set of assumptions. In this article, I derive the assumptions needed to make credible nonparametric comparisons using anchoring vignettes, and propose a new nonparametric scale that does not assume homogeneity among respondents. I also provide methods for evaluating empirically whether a set of anchoring objects can produce credible nonparametric interpersonal comparisons. Two empirical studies illustrate the importance of accounting for differences in the use of ordinal scales by showing how our inferences about interpersonal comparisons may change as a function of the assumptions we accept.

Survey respondents are often asked to describe their own attitudes and attributes using ordinal scales. Researchers have questioned the validity of using these data to make interpersonal comparisons because respondents may differ in how they interpret scale categories.[1] For example, when measuring satisfaction on an ordinal scale, one respondent's use of a category labeled "moderately satisfied" may have the same effective meaning and behavioral consequences as another respondent's use of the category "slightly satisfied." The problem of interpersonally incomparable survey responses may exist anytime a question uses ordinal categories that are subjectively defined, but is particularly acute when comparing responses across countries and cultures.

Researchers have proposed using anchoring vignettes as a method for improving interpersonal comparisons (King et al. 2004; King and Wand 2007). Anchoring vignettes are short descriptions of hypothetical individuals with attributes that place them at different locations along the scale being evaluated.[2] By observing how each individual rates a common set of vignettes, a researcher may discern differences in scale use and and thereby adjust the meaning of each individual's self-evaluation. Anchoring information can also be obtained by asking respondents to

Jonathan Wand is Assistant Professor of Political Science, Encina Hall West, Stanford University, Stanford, CA 94305 (wand@stanford.edu).

Earlier versions of this article were presented at the Robert Wood Johnson Health Policy Scholars annual meeting, the Society for Political Methodology Summer Conference, the annual meeting of the American Political Science Association, and seminars at the University of Michigan and Stanford University. This work has benefited from the comments of the participants at these presentations. Thanks to Martha Bailey, Elizabeth Bruch, Vincent Hutchings, Gary King, Skip Lupia, Walter Mebane, Gerardo Munck, Maisy Samuelson, Jas Sekhon, and Chris Zorn for helpful comments and conversations. Two anonymous reviewers also made helpful comments that improved the presentation of this work. Thanks to the Robert Wood Johnson Foundation and Stanford's Institute for Research in the Social Sciences (IRiSS) for research support and to the Center for Political Studies and the Department of Health Management and Policy at the University of Michigan for their hospitality while this work was initially written. The data analysis uses the open-source R library *anchors* (Wand, King, and Lau 2011), which is available at the Comprehensive R Archive Network. The *anchors* library also contains the data and replication code for this study. I am solely responsible for any errors or omissions.

[1] See for examples Aldrich and McKelvey (1977); Brady (1989, 1990); Rossi, Gilula, and Allenby (2001); King et al. (2004); Javaras and Ripley (2007); King and Wand (2007); and Hopkins and King (2010).

[2] Examples of anchoring vignettes can be found in studies of personal health (King et al. 2004; Salomon and Murray 2004), studies of job disabilities (Kapteyn, Smith, and Soest 2007), studies of corruption and state effectiveness (Grzymala-Busse 2007), and many other topics. See also the recent symposium on anchoring vignettes in the JRSS-A (Chevalier and Fielding 2011; Van Soest et al. 2011).

American Journal of Political Science, Vol. 00, No. 0, xxx 2012, Pp. 1–14

DOI: 10.1111/j.1540-5907.2012.00597.x

rate commonly known individuals. Aldrich and McKelvey (1977) used ideological evaluations of presidential candidates to calibrate measures of individual ideology.

In this article, I derive the assumptions needed to credibly compare individuals using nonparametric anchor-based methods. I show that existing nonparametric anchor-based methods depend upon assumptions little different from the traditional homogeneity assumption and thus do not solve the problem they are meant to address. I propose an alternative scale that produces credible comparisons while requiring weaker assumptions. Finally, I provide empirical tests to evaluate whether the assumptions needed to produce credible comparisons are valid. These results provide new answers to the basic question of what can be learned from interpersonally incomparable survey responses.

In the next section, I describe the type of problems addressed here. The following section contains the main analytical results, including the assumptions needed to make credible nonparametric comparisons using anchoring objects. I then present results from two empirical studies. First, I revisit a study of self-reported political efficacy in Mexico and China that included anchoring vignettes (King et al. 2004). Second, I examine voting behavior in the 2004 presidential elections as a function of policy preferences, with major party candidates used as anchoring objects. Both of these studies illustrate the importance of accounting for differences in the subjective use of ordinal scales.

Comparisons Based on Ordinal Scales

For many questions of public opinion and political psychology, there is no quantitative or metric scale for measuring the attitude or attribute of an individual. For example, satisfaction with the quality of a political system or a consumer product cannot be directly measured in the same way as we can count the number of times individuals vote or purchase a product. An individual may nonetheless be more or less satisfied despite the problem of defining these values. A common approach to measuring such attitudes is to offer an individual a set of ordered categories with which to rate herself.

A running example in this article is the measurement of political efficacy solicited by asking respondents,

> How much say do you have in getting the government to address issues that interest you? *(1) No say at all, (2) Little say, (3) Some say, (4) A lot of say, (5) Unlimited say.*

The World Health Organization (WHO) asked this question in a 2002 cross-national survey, and the results were previously examined by King et al. (2004). Figure 1(a) shows the proportion of respondents in Mexico and China who selected each category for their self-evaluations. Mexico appears to have more respondents who are at the lowest end of the scale, with a majority of respondents selecting the category of “No say at all.”

Categories such as having “Little say” and “Some say” are not well defined and may cover a variety of situations. Even the category having “No say at all” could encompass different degrees of political efficacy depending on the standard of the individual. Statistical models commonly model ordinal scales as discrete representations of an underlying one-dimensional continuum, where categories are defined as a mutually exclusive and exhaustive set of intervals that divide up the continuum. This modeling assumption motivates ordered probit models and polytomous Mokken scaling. The underlying continuum is standardly referred to as the “latent” scale since a researcher does not observe the location of an individual on the continuum, but only the category of the survey response.

Figure 2 illustrates the logic of mapping values defined on a continuous latent scale into ordinal categories. The attribute of individual i is located at a point $\tilde{y}_i$ on the latent continuum. Individuals know their own value $\tilde{y}_i$ and could be ordered by these values if they were observed, but respondents are not able (or are not asked) to describe this precise value. When posed with a question that requires a response on an ordinal scale, each respondent must first define the meaning of the categories by choosing the locations of the cutpoints that divide the latent scale into intervals. The cutpoint separating category k and $k+1$ is labeled τ_{ik}.[3] The observed category choice y_i is thus defined as

$$y_i = k \quad \Leftrightarrow \quad \tau_{i,k-1} \leq \tilde{y}_i < \tau_{ik}. \tag{1}$$

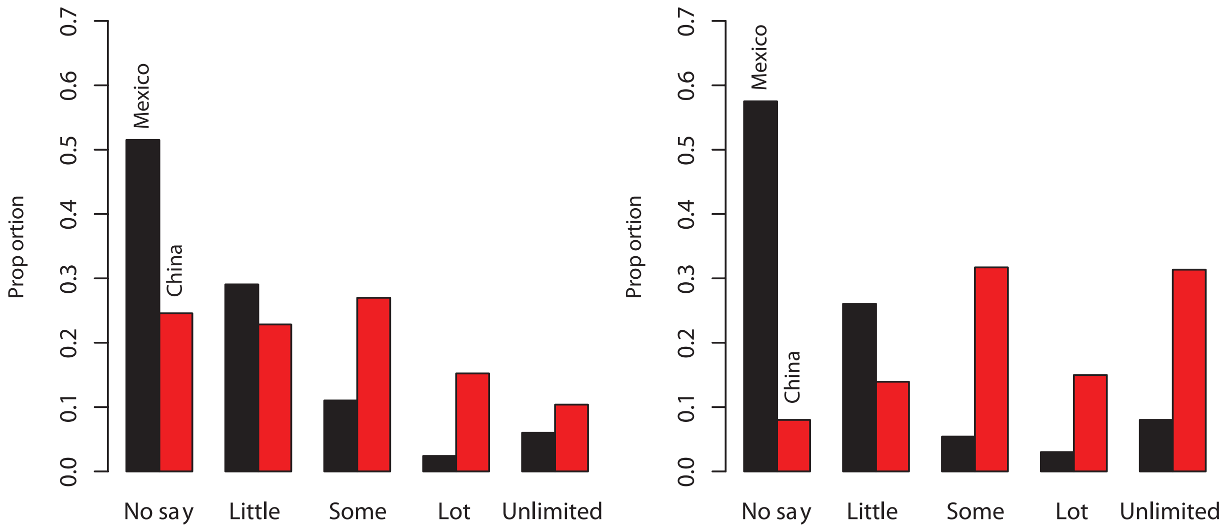
I refer to y_i as the “self-rating.” In Figure 2, an individual with an attribute located at $\tilde{y}_i$ would rate herself as having “Little say.”

Along with the self-evaluation question, respondents in the WHO survey were subsequently asked to rate five related vignettes using the same ordinal scale.[4] Each hypothetical individual in the vignettes lacked

[3] Given $K+1$ categories, define cutpoints τ_0 and τ_{K+1} at the lower and upper bounds of the latent space. If latent space is the real number line, then $\tau_0 = -\infty$ and $\tau_{K+1} = \infty$.

[4] The surveys were completed in June 2002, with N = 430 in China and N = 551 in Mexico. These data and the wording for additional anchoring vignettes are described by King et al. (2004).

FIGURE 1 Distribution of Ratings for Political Efficacy Questions in Mexico and China

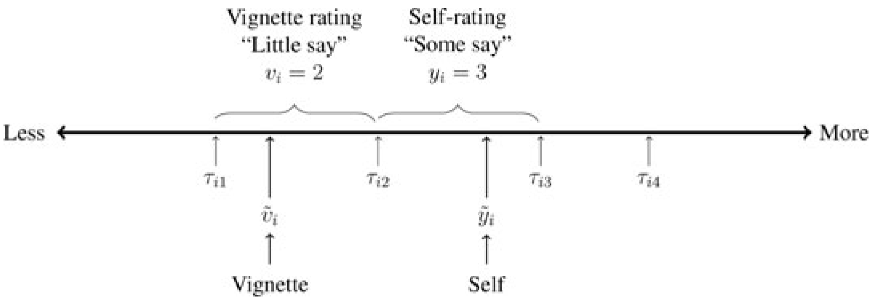


(a) Self-rating

(b) Rating of Suffering vignette

Source: 2002 WHO surveys.

FIGURE 2 Mapping Latent Values of "Say in Government" into Ordinal Categories



clean drinking water, but differed in terms of the stated belief about his or her capacity to have the government improve the situation. For example,

> [John][5] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future. How much say does [John] have in getting the government to address issues that interest [him]?

This vignette, which I will simply refer to as the "suffering" vignette, is on average considered by respondents to be the worst situation among the five vignettes. The distribution of ratings of this vignette in each country is shown in Figure 1(b).

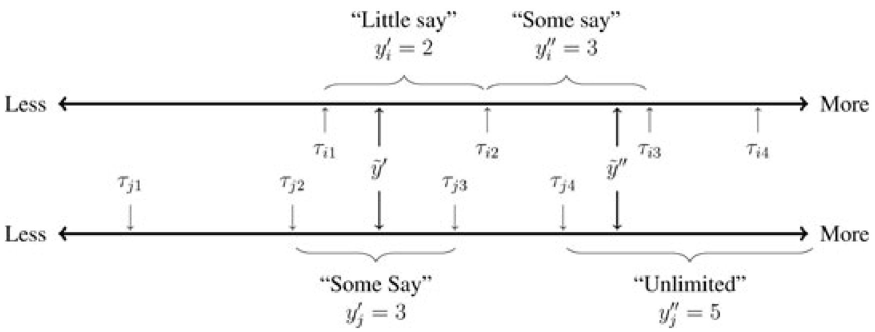
Anchoring vignettes are designed to be evaluated on the same latent scale using the same definitions of the ordinal categories for the self-ratings. The observed rating v_{im} for vignette m is defined in the same way as the self-rating,

$$v_{im} = k \quad \Leftrightarrow \quad \tau_{i,k-1} \leq \tilde{v}_{im} < \tau_{ik}, \qquad (2)$$

where $\tilde{v}_{im} \in \mathbb{R}$ is the respondent's judgment about the location of vignette m along the latent scale. In Figure 2, the individual would rate a vignette located at $\tilde{v}_i$ as having "Some say."

[5] The name in each vignette (I inserted "John" as an example) was replaced with a name intended to coincide with the culture and gender of the survey respondent.

FIGURE 3 Mapping Latent Values into Ordinal Categories for Individuals *i* and *j*, Subject to DIF



Standard approaches to comparing self-ratings measured on an ordinal scale rely upon an assumption that respondents share a common understanding of where to place each cutpoint, $\tau_k = \tau_{ik} = \tau_{i'k}$ for all i, i', and k. If one believed that all individuals in both countries agreed on the definition of the categories, and in particular what it means to have "No say at all" in government, then the logical conclusion would be that the Mexican respondents are worse off than the Chinese respondents.

The assumption of homogeneity of scale definitions across individuals is at least debatable in the context of subjective or loosely defined ordinal scales. One respondent's notion of having "Little say" could reasonably be another respondent's notion of having "Some say." Figure 3 illustrates an example of two types of respondents who interpret the same category differently in the upper and lower scales. If an individual of each type were located at the same latent location $\tilde{y}$ they would declare self-ratings of "Little say" and "Some say," respectively. Alternatively, if an individual of each type were located at $\tilde{y}$ they would declare self-ratings of "Some say" and "Unlimited say," respectively.

When any cutpoint location differs across individuals, i.e., $\tau_{ik} \neq \tau_{i'k}$ for any k, then the response of i and i' are said to be subject to *differential item functioning* (DIF; King et al. 2004).[6] There is an empirical basis for worrying that respondents in Mexico and China are using the ordinal scale in different ways. Observing in Figure 1(b) that respondents differ in their rating of the same vignette raises the concern that individuals have different standards for interpreting the scale categories. Moreover, difference in the use of scales appears to vary systematically across cultures, with Mexico being more prone to use the lower end of the scale in evaluating the vignette.

A comparison will be said to be *credible* if the claim that two respondents are strictly ordered implies that the respondents have the same strict ordering on the latent scale. It is possible to make credible comparisons based on the observed self-ratings (y_i) under the assumption that respondents have a common understanding of cutpoint locations, i.e., $\tau_k = \tau_{ik}$ for all i, k. In this case, $y_i = k < y_{i'}$ would imply the same ordering of the individuals i and i' on the latent scale, $\tilde{y}_i < \tau_k < \tilde{y}_{i'}$. Comparisons based on an ordinal scale are not credible in the presence of DIF because we could observe an ordering of individual responses $y_{i'} < y_i$ even if the true ordering of the individuals is the opposite, $\tilde{y}_{i'} > \tilde{y}_i$. For example, $\tau_{i',k} < \tilde{y}_{i'} < \tilde{y}_i < \tau_{i,k}$ yields $y_{i'} < y_i$.

In the next section, I show when and how anchoring vignettes provide the ability to make credible comparisons across individuals in the presence of DIF. In the context of comparing levels of political efficacy across countries, the availability of anchoring objects will enable us to determine whether one country has lower levels of perceived political efficacy or whether respondents differ in the interpretation of the scale categories.

Credible Interpersonal Comparisons with Anchors

Nonparametric anchor-based methods for interpersonal comparisons replace the potentially incomparable self-ratings with new measurements purged of DIF. An individual's rating of anchoring objects serves as a reference point against which her self-ratings are rescaled. While the general logic of using anchoring vignettes has intuitive

[6] Though people may differ in where their cutpoints are located, it is standard to assume that there exists a common ordering of cutpoints among all individuals: $\tau_{ik} < \tau_{ik'}$ for all $k < k'$. The ordering of cutpoints is usually implied in surveys by the order in which the categories are presented, even if the category labels are not intrinsically informative about the ordering.

Table 1 Calculations for the Nonparametric Scales *B* and *C*, Given a Self-Rating, y_i, and Ratings of Two Anchoring Vignettes, $v_{i1} < v_{i2}$

Relative Order of Ratings					Notation	*C*-scale	*B*-scale
Self	$<$	Vignette 1	$\leq$	Vignette 2	$y_i < v_{i1} \leq v_{i2}$	1	1
Self	$=$	Vignette 1	$<$	Vignette 2	$y_i = v_{i1} < v_{i2}$	2	{ 1,2 }
Vignette 1	$<$	Self	$<$	Vignette 2	$v_{i1} < y_i < v_{i2}$	3	2
Vignette 1	$<$	Self	$=$	Vignette 2	$v_{i1} < y_i = v_{i2}$	4	{ 2,3 }
Vignette 1	$\leq$	Vignette 2	$<$	Self	$v_{i1} \leq v_{i2} < y_i$	5	3

appeal, I show that some nonparametric methods using anchoring objects can require unappealing assumptions in order to produce credible comparisons. I also provide a new nonparametric scale that produces credible comparisons under weak assumptions.

The *C*-Scale

The *C*-scale was proposed as a method to make credible comparisons even in the presence of DIF (King et al. 2004; King and Wand 2007). The *C*-scale value for an individual can be defined in terms of the relative ordering of observable quantities.[7]

$$C_i = \begin{cases} 2m+1 & \text{if} \quad v_{im} < y_i < v_{i,m+1} \\ 2m & \text{if} \quad y_i = v_{im} \end{cases} \tag{3}$$

where $m \in \{1, \ldots, M\}$ is the index for the vignettes and $v_{i0} = -\infty$ and $v_{i,M+1} = \infty$. The task of measuring the attribute of an individual is thus reformulated to one of establishing whether a respondent rates herself higher, lower, or the same as one or more anchoring objects. It is important to note that the *C*-scale is nonetheless a function of the cutpoint locations $(\tau_{i1}, \ldots, \tau_{iK}), (\tau_{i1}, \ldots, \tau_{iK})$. This can be seen more clearly by restating the calculation of C_i using the definitions of y_i and v_{im},

$$C_i = \begin{cases} 2m+1 & \text{if} \quad \tilde{v}_{im} < \tau_{ik'} < \tilde{y}_i < \tau_{ik} < \tilde{v}_{i,m+1} \quad \text{for some } k' < k \\ 2m & \text{if} \quad \tau_{i,k-1} < \tilde{y}_i, \tilde{v}_{im} < \tau_{ik}. \end{cases}$$

Table 1 illustrates the construction of the *C*-scale. The table shows the five weak orderings of a self-rating (y_i) relative to the ratings of two anchoring objects $(v_{i1} < v_{i2})$, along with the associated value of the *C*-scale. The logic of interpersonal comparisons using the *C*-scale values is also the same as with self-ratings alone. Individuals who have higher values on the *C*-scale are interpreted as being higher on the latent scale.

The proposal of the original *C*-scale by King et al. (2004) was accompanied by two new assumptions. One assumption is *response consistency*, which requires that a respondent use the same cutpoint locations for evaluating the anchoring items and her self-evaluation question. This assumption is fundamental to all methods of analysis based on anchoring objects because it connects an individual's ratings of the anchoring objects to her self-rating. This assumption was implicit in the example of Figure 2, where the same set of cutpoints was used both in categorizing the latent value of the individual $\tilde{y}_i$ and in categorizing the perceived location of the anchor $\tilde{v}_i$.

The other assumption is *vignette equivalence*, which requires that every individual perceive a vignette at the same location on the latent scale, $\tilde{v}_m = \tilde{v}_{ij}$ for all *i* and all *m*. In the presence of DIF the observed rating of a vignette may differ across individuals even with vignette equivalence. In Figure 4, the two hypothetical individuals *i* and *j* agree on the location of the vignettes ($\tilde{v}_1 = \tilde{v}_{i1} = \tilde{v}_{j1}$ and $\tilde{v}_2 = \tilde{v}_{i2} = \tilde{v}_{j2}$), but not on the location of cutpoints. As such, a person *i* would rate the vignettes as "No say" and "Little say," while the person *j* would rate the same vignettes as "Little say" and "Some say," respectively.

It is quite a strong assumption that everyone would perceive each anchoring object in exactly the same way. Although respondents receive the same vignette description, respondents may differ in their judgment as to the location of a hypothetical individual on the latent scale. This problem is particularly relevant in the case of anchoring objects that are political candidates. Individuals are thought to overestimate the distance they perceive between themselves and less preferred candidates and minimize the perceived distance from friendly candidates (Brady and Sniderman 1985). The assumption of vignette equivalence may not be any more empirically plausible than the assumption of the absence of DIF in the self-ratings.

Even with these two assumptions, however, anchoring vignettes do not ensure credible comparisons. Figure 4

[7] A more general definition of *C* which includes the accommodation of ties between the self-rating and multiple vignettes is provided by King and Wand (2007).

FIGURE 4 Model of Individuals i and j with Response Consistency and Vignette Equivalence, Yet an Incorrect Ordering on the C-scale: $C_i = 4 > C_j = 2$ While the True Ordering Is $\tilde{y}_i > \tilde{y}_j$

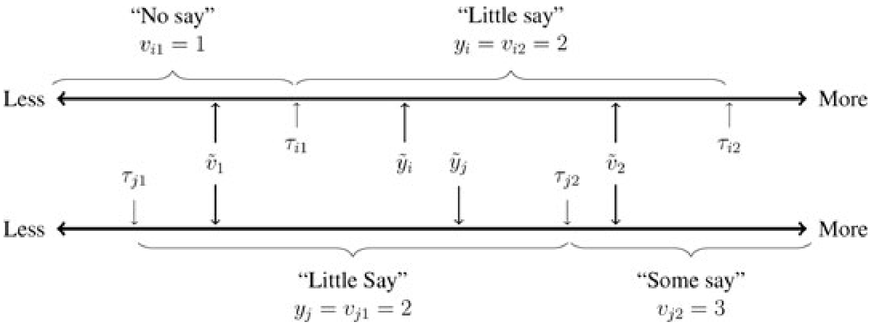
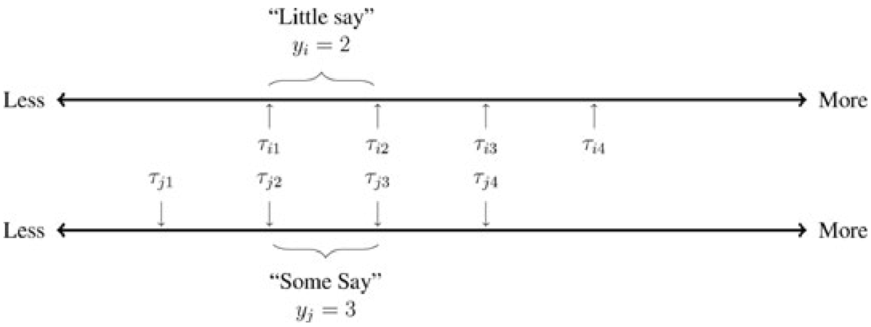


FIGURE 5 DIF with Shared Cutpoint Locations



illustrates the problem. As already noted, the two hypothetical individuals conform to response consistency and vignette equivalence, and yet based on the C-scale, we would infer the incorrect order. The C-scale indicates that individual i has more say in government than individual j ($C_i = 4 > C_j = 2$) while the opposite is true ($\tilde{y}_i < \tilde{y}_j$).

Ensuring credible comparisons using the C-scale requires two additional assumptions beyond vignette equivalence and response consistency. First, all individuals must agree on where cutpoints may be placed, even if they disagree on which cutpoint to use at a particular location. An example of this *interval equivalence* assumption is shown in Figure 5. The key feature of this figure is that the hypothetical individuals differ on which label to apply to particular intervals, but agree on the locations of the intervals shared in common. This assumption is needed in order for the C-scale to avoid the type of inferential failure illustrated in Figure 4.

The strong form of the second additional assumption is that an anchoring vignette cannot be rated as being in either the top (e.g., "unlimited say") or bottom category (e.g., "No say at all") of the ordinal scale by any individual. If this *moderate vignette* restriction is not fulfilled, then in Figure 5 we could have $C_i = 2 > C_j = 1$ despite $\tilde{y}_i < \tilde{y}_j$ by means of $\tilde{y}_i < \tilde{y}_j < \tau_{j1} < \tilde{v}_1 < \tau_{i1} = \tau_{i2}$. A weaker form of this assumption also suffices to ensure credible comparisons, wherein no anchoring vignette and self-rating may be rated both in the same extreme category by the same individual.

Since a researcher controls both the design and the inclusion of each anchoring vignette, the prohibition against extreme vignettes could in principle be fulfilled either through pilot studies or by excluding from the analysis anchoring objects that are rated in an extreme response category. In practice, however, it may be that no anchoring object satisfies this requirement. For example, substantial shares of respondents from both Mexico and China rated each of the five vignettes in either the top or bottom categories.

With this set of four assumptions, the C-scale produces credible comparisons, such that $C_i < C_{i'}$

implies $\tilde{y}_i < \tilde{y}_{i'}$. However, the cost of achieving credible comparisons is high, and the gains are somewhat modest. The requirement of vignette equivalence adds a new, unappealing form of strict homogeneity to the analysis. For this price, the C-scale only accommodates a restricted form of DIF defined by the interval equivalence assumption. The notion that respondents share a common understanding of how to divide the latent dimension into intervals has much in common with the traditional homogeneity assumption that there is no DIF, and thus C does not fully solve the problem for which it was intended.

An Alternative, the B-Scale

It is possible to make credible comparisons across individuals in the presence of DIF using an alternative scale, the B-scale, without requiring interval equivalence or vignette equivalence. The B-scale is built using the same materials as the C-scale, and they are often in agreement on the ordering of individuals produced by using the relative ranks of self-rating and anchor ratings. Consider, for example, the values of the B-scale shown in the final column of Table 1. The B-scale and the C-scale agree that an individual in the first row ($C = 1$, $B = 1$) would be in the lowest possible category, and an individual in the last row ($C = 5$, $B = 3$) would be in the highest possible category. The scales also agree that an individual in the third row ($C = 3$, $B = 2$) is strictly in between the lowest and highest categories.

The difference between the values of the B-scale and C-scale lies in the information that is claimed to exist when a self-rating is tied with the rating of an anchoring object, $y_i = v_{im}$. The C-scale claims that there is information for making strict comparisons with adjacent rank orderings in such cases. However, as just noted, the ability to make this claim relies on the assumptions of vignette equivalence and interval equivalence. The B-scale claims less information in the occurrence of a tie, represented as a set of B-scale values rather than a single index value. If $y_i = v_{i1}$ then $B_i = \{1, 2\}$, if $y_i = v_{i2}$ then $B_i = \{2, 3\}$, and so forth. By using both scalar and set values, the calculation of each individual's B-scale values indicates which comparisons provide strict orderings of individuals and which comparisons are ambiguous. An individual with $B_i = \{1, 2\}$ can be strictly ordered as lower than individuals with $B_j = 3$, $B_j = \{3, 4\}$, $B_j = 4$, and so on. However, this same individual could not be strictly ordered relative to those with $B_j = 1$, $B_j = 2$, or $B_j = \{2, 3\}$.

Formally, the B-scale represents the location of each individual relative to the average perceived location of each vignette,

$$B_i = m \quad \Leftrightarrow \quad \tilde{v}_{0,m-1} \leq \tilde{y}_i < \tilde{v}_{0m}, \tag{4}$$

where $\tilde{v}_{0m}$ can be thought of as either the true latent location or the average perceived location of anchoring vignette m, $\tilde{v}_{0m} = E(\tilde{v}_{im})$.[8] Again, like the definition of cutpoints, let $\tilde{v}_{i0} = -\infty$ and $\tilde{v}_{i,m+1} = \infty$. Although neither $\tilde{y}_i$ nor any of the v_{0m} are observed, survey responses bound their relative locations.

The definition of the B-scale does not rely on cutpoint locations and, as a result, provides credible comparisons without the requirement of interval equivalence or vignette equivalence. For an arbitrary form of DIF, it is sufficient that respondents simply perceive vignettes to be on the same side as the true location of the vignette relative to the respondent's own location. For example, if a vignette describes a hypothetical individual who is worse off than the respondent, the respondent cannot perceive the hypothetical individual to be better off. Conversely, a hypothetical individual who is better off than the respondent cannot be perceived as worse off. This assumption only precludes $\tilde{v}_{0m} < \tilde{y}_i < \tilde{v}_{im}$ and $> \tilde{v}_{im} < \tilde{y}_i < \tilde{v}_{0m}$. I refer to this assumption as "Order Preserving Imperfect Anchors" (OPIA). OPIA has the appealing feature that individuals may disagree on the precise location of anchoring objects and encompasses the phenomena of perceptual bias investigated by Brady and Sniderman (1985). The B-scale produces credible comparisons assuming response consistency and OPIA.

Even with the assignment of set values to individuals, the distribution of B-scale values can be estimated. For example, consider the goal of calculating the proportion of respondents who are worse off than the lowest anchoring vignette. This is equivalent to

$$P(B = 1) \quad \Leftrightarrow \quad P(\tilde{y}_i < \tilde{v}_1). \tag{5}$$

This condition is satisfied by those wherein $y_i < v_{i1}$ but also by some fraction of cases where $y_i = v_{i1}$.[9] The share of respondents with $B = 1$ can be stated in terms of the decomposition,

$$\begin{aligned} P(B_i = 1) = P(y_i < v_{i1}) P(\tilde{y}_i < \tilde{v}_1 \mid y_i < v_{i1}) \\ + P(y_i = v_{i1}) P(\tilde{y}_i < \tilde{v}_1 \mid y_i < v_{i1}), \end{aligned} \tag{6}$$

where we know the values of three of the four quantities on the right-hand side. The value $P(\tilde{y}_i < \tilde{v}_1 \mid y < v_{i1}) = 1$

[8] Again, like the definition of cutpoints, let $\tilde{v}_0 = -\infty$ and $\tilde{v}_{J+1} = \infty$.

[9] The case of $\tau_{i,k-1} < \tilde{y}_i < \tilde{v}_1 < \tau_{i,k}$ satisfies $B = 1$, but $\tau_{i,k-1} < \tilde{v}_1 < \tilde{y}_i < \tau_{i,k}$ does not.

follows from response consistency and OPIA. The survey sample provides estimates of the ordering of ratings $P(y_i < v_{i1})$ and $P(y_i = v_{i1})$. However, the value of $\lambda_1 = P(\tilde{y}_i < \tilde{v}_1 \mid y_i = v_{i1})$ cannot be known without an assumption about the joint distribution of unobserved parts of the model $(\tau, \tilde{v}, \tilde{y}_i)$. Nonetheless, we know the logical range of the probability: $\lambda_1 \in (0, 1)$. Calculating the function at the two extremes ($\lambda_1 = 0$ and $\lambda_1 = 1$) yields bounds for $P(\tilde{y}_i < \tilde{v}_1)$ in terms of estimable quantities,

$$P(y_i < v_{i1}) \leq P(\tilde{y}_i < \tilde{v}_1) \leq P(y_i < v_{i1}) + P(y_i = v_{i1}). \tag{7}$$

The bounds of this equation are sharp per the definition of Manski (1995, 25), since there is no additional information in the observed data (v_{i1}, y_i) to shrink the interval around $P(\tilde{y}_i < \tilde{v}_1)$. The same logic can be applied to other values of B. Taking into account that the range of proportions are themselves subject to sampling variability, we also can construct confidence intervals for these bounds (cf. Manski 1995, 20).

Testable Implications of Nonparametric Scales

With multiple anchoring vignettes available, it is possible to evaluate their properties and their compliance with the identification assumptions of the nonparametric scales. Since a goal of using anchoring items is to overcome metaphysical arguments in the measurement of public opinion, it behooves researchers to employ multiple vignettes even if this is not thought to be needed for identifying a quantity of interest, such as the worst-off members of a population.

A testable implication of the B-scale assumptions is that there should not be a misordering of vignette evaluations that straddle the self-evaluation. Response consistency and OPIA precludes $v_{i,m'} \leq y_i < v_{im}$ if $m < m'$. It is not a problem for inference if respondents change the order of vignettes that are distant from their own circumstances so long as they do not straddle an individual's self-rating value. In practice, any disagreement among respondents on the ordering of vignettes should be investigated as an indication of a design problem.

For the C-scale, respondents need to agree on the location of the anchoring objects and the location of intervals (but not necessarily their labels). Moreover, no vignette can be rated in an extreme category along with a self-rating. If these assumptions hold, then the rating of each vignette should be shifted (if at all) by the same amount when comparing a pair of respondents. As such, the difference in observed ratings for a pair of anchoring items should be the same for all respondents. Formally, $v_{im} - v_{im'} = v_{i'm} - v_{i'm'}$ for all pairs of vignettes $m \neq m'$ and all pairs of individual $i \neq i'$.

Comparing Political Efficacy in Mexico and China

Continuing the study of political efficacy introduced earlier, I examine the comparison of survey responses in Mexico and China using the C- and B-scales. King et al. (2004) previously analyzed these data using the C-scale with five anchoring vignettes, and the distribution of C values for each country is replicated in Figure 6. Based on the differences between these distributions, King et al. (2004) argued that Chinese respondents revealed less of a sense of political efficacy than Mexican respondents, and

> The correction exactly switches the conclusion about which country has more political efficacy, and makes it in line with what we know. Indeed, the spike at $C = 1$ is particularly striking: 40% of Chinese respondents judge themselves to have less political efficacy then they think the person described in the [. . .] ("suffering in silence") vignette has. This result, which we never would have known using standard survey methods, calls into question research claims about the advances in local elections in China, even in the limited scope to which such elections are intended. (196)

To focus attention on the key differences in the survey responses across countries and a comparison of the inferences of B and C where it matters most, Table 2 presents a simplified summary of the distribution of B and C values calculated using just the "suffering" vignette. The same conclusion drawn by King et al. (2004) is also drawn in this simplified C-scale analysis,[10] with China having a greater proportion of people in the lowest category ($P(C = 1) = 0.56$) than Mexico ($P(C = 1) = 0.25$).

In contrast to either the self-ratings alone or the C-scale, the B-scale indicates that there is not enough

[10] In this simpler comparison, the proportion of cases in the lowest category ($C = 1$) increases relative to the analysis performed by King et al. based on five vignettes; the differences are primarily due to the treatment of intransitivities in observed vignette order, and also in part due to reclaiming some cases that were dropped by King et al. due to missing responses on other vignettes. The number of vignettes does not affect the key features of the inferences discussed here.

FIGURE 6 Distribution of *C* Values Using All Five Vignettes for Political Efficacy in Mexico and China, by Country

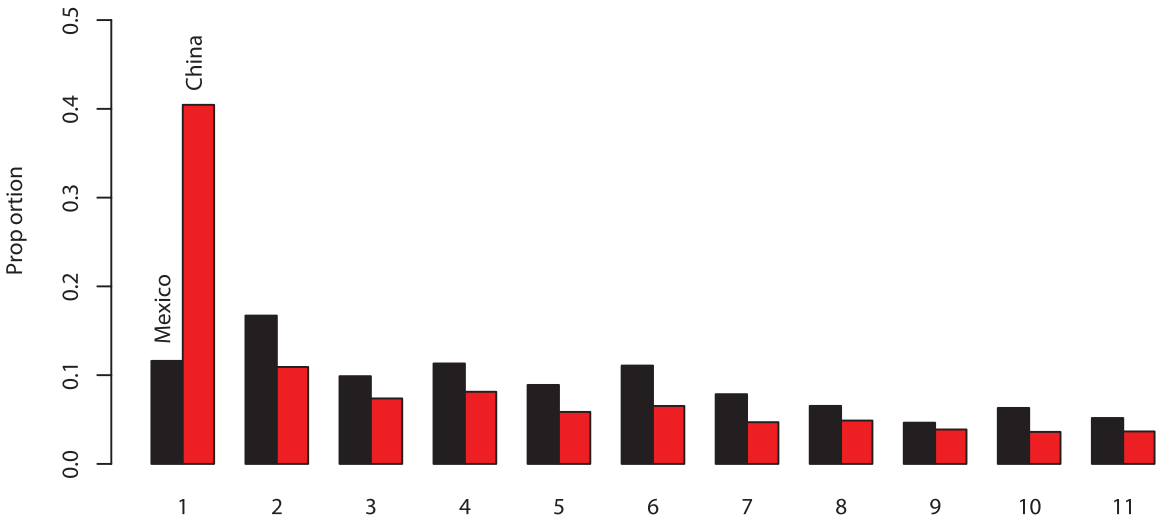


TABLE 2 Distribution of Relative Ranks of Political Efficacy Self-Rating and the Rating of the "Suffering" Vignette, by Country

	Rank Value		Mexico		China	
Survey Responses	***B***	***C***	**N**	**Pr**	**N**	**Pr**
Self < Suffering Vign. $(y_i < v_i)$	1	1	124	0.25	160	0.56
Self = Suffering Vign. $(y_i = v_i)$	{1,2}	2	226	0.45	73	0.25
Self > Suffering Vign. $(y_i > v_i)$	2	3	149	0.30	54	0.19

information to discern whether respondents in Mexico or China are worse off. The bounds on $P(B = 1)$ for the two countries overlap and thus support the possibility of either ordering. Mexico could be worse off, with $P(B = 1)$ as high as 0.70 (0.25+0.45) in Mexico and as low as .56 in China. Or, Chinese respondents could be disproportionately worse off with $P(B = 1)$ as low as 0.25 in Mexico and as high as 0.81 (0.56+0.25) in China.[11]

With three different conclusions from three different methods, it is clear that inferences about the relative political efficacy in Mexico and China depend upon which assumptions we believe. The availability of anchoring vignettes makes it possible to evaluate the plausibility of these assumptions.

To accept the conclusion that Mexico is worse off based on self-ratings alone, as shown in Figure 1(a), one must believe that there are no systematic differences in the use of the scale across countries. The striking differences in the distribution of vignette ratings between the countries, shown in Figure 1(b), cast doubt on the assumption that DIF is not present. China respondents appear to systematically use higher categories of the ordinal scale, with only 8% of respondents rating the "suffering" vignette as having "No say." In contrast, 58% of Mexican respondents rate the same vignette in the lowest category.

To accept the C-scale conclusion that China is worse off than Mexico requires accepting the stringent assumptions of this scale, including vignette and interval equivalence. However, based on all five vignettes used by King et al. (2004), only 8% of respondents give responses that are consistent with the restrictions needed by C to produce credible comparisons. In a more favorable analysis

[11] Taking into account that the ranges of these proportions are themselves subject to sampling variability, we can construct confidence intervals for these identification regions. The 95% confidence interval for Mexico's identification region is (0.20, 0.75), while the confidence interval for China is (0.51, 0.85). Using these intervals, we again come to the same conclusion that one cannot determine using these data whether China or Mexico is worse off. The relative similarity of the confidence intervals and identification regions reflects the fact that the dominant problem for inference is the width of the identification region due to ties in B, while sampling variation is a second-order problem.

using only two vignettes, the "suffering" vignette and the vignette describing the most say in government, the percentage of responses consistent with credible comparisons using C rises to only 34%.

The B-scale concludes that there is not enough information in the survey responses to clearly declare one country worse off than the other. In contrast to the implausibility of assuming no DIF or assuming all the requirements of the C-scale, 94% give responses that are consistent with the B-scale constructed using all five vignettes; the rate of consistency rises to 99% with only two vignettes. Going beyond the conclusion of the B-scale and claiming that more people in Mexico or China perceive themselves as worse off than the "suffering" vignette depends on defending assumptions that are not supported by the survey data.

In addition to testing the assumptions of each method, it is also useful to examine the substance of the ordering claimed by C and why C and B come to different conclusions. As noted above, the key difference between the C- and B-scales is their treatment of cases where respondents assigned the same rating to themselves and the "suffering" vignette. With nearly half of all Mexico respondents giving responses where $y_i = v_i$, our inference about the ordering of countries hinges on whether to think of these cases as better off than those where $y_i < v_i$ as in the case of the C-scale or whether to attribute ambiguity to them as in the B-scale.

Treating $C_i = 2$ as greater than $C_i = 1$ is less plausible when respondents do not have room on the ordinal scale to rate themselves worse than a vignette. In particular, for respondents who would rate a vignette in the lowest category ($v_{im} = 1$), it is not possible for respondents to also rate themselves lower than this vignette. In such cases, observing $C_i = 1$ is not possible, and the plausibility of the strict ordering between $C_i = 1$ and $C_i = 2$ is particularly dubious.

Among Mexican respondents, not only are 45% of cases coded as $C = 2$, but these ties occur with respondents rating both themselves and the suffering vignettes as having "No say" ($y_i = v_i = 1$). If tied cases at the bottom of the scale are not treated as greater than $C = 1$, we would not find that China is worse off in terms of $P(C = 1)$. To see this, reassign cases to $y_i = v_i = 1$ to $C = 1$. With 33% of all Mexican respondents giving ties at the lowest category, we have $P(C_i = 1) = .25 + .33 = .58$ in Mexico. In China, only 3% of all respondents had $y_i = v_i = 1$, such that the counterfactual allocation is $P(C_i = 1) = .56 + .03 = .59$. Taking account of the possible floor effects in the scales, the country appears to be essentially the same in terms of low political efficacy.

Policy Preferences and Their Behavioral Consequences

If the B-scale provides a more accurate ordering of individuals according to the latent attitude or attribute, then the B-scale values should also better predict related behavior than the original self-rating. In this section, I consider the relationship between policy preferences and vote intentions in the the 2004 American National Election Study.

The measure of policy preferences is a 7-point ordinal scale described to respondents as follows:

> Some people think the government should provide fewer services even in areas such as health and education in order to reduce spending. Suppose these people are at one end of a scale, at point 1. Other people feel it is important for the government to provide many more services even if it means an increase in spending. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between, at points 2, 3, 4, 5, or 6.

Unlike the political efficacy scale, only the endpoints are defined in words, with the intermediary values left as integer values. Using this scale, each respondent was asked to rate herself and subsequently to rate the presidential candidates:

> Where would you place yourself on this scale, or haven't you thought much about this?
> Where would you place [George W. Bush/John Kerry] on this issue?

In the following discussion, the candidates will serve as anchoring objects. Vote intentions were measured by offering three options: John Kerry, George W. Bush, and Ralph Nader, although I focus on the subset of respondents who had the intention of voting for one of the major party candidates.[12] I reverse the indexing of the ratings such that higher values imply a reduction in spending, which means the expected ordering of the Democratic (D) and Republican (R) candidate locations is $\tilde{v}_D < \tilde{v}_R$. This recoding allows for the traditional spatial representation of ideological debates, placing the liberal position

[12] Those with intentions to vote on Election Day were asked, "Who do you think you will vote for in the election for President?" Those who did not indicate an intention to vote were asked, "If you were going to vote, who do you think you would vote for in the election for President?" Those who declined to say were given a probe question.

TABLE 3 Percent with Democratic Vote Intention in 2004 Presidential Election for Each Combination of Self-Rating and *B*

	Position of Self-Rating Relative to Candidate Ratings					
	More Liberal Than D	Tied with D	Between D and R	Tied with R	More Conserv. Than R	Self-Rating Average
Self-rating (y_i):	$B = 1$	$B \in \{1, 2\}$	$B = 2$	$B \in \{2, 3\}$	$B = 3$	
1. More services	95	86	—	—	—	92
2.	95	94	78	—	—	93
3.	82	90	72	44	0	77
4.	100	85	57	22	7	49
5.	—	—	46	9	13	22
6.	—	—	50	0	6	10
7. Reduce spending	—	—	—	0	3	3
B-scale average	93	90	61	16	8	

Notes: $N = 621$. Blank cells have no respondents. (D)emocrat = Kerry, (R)epublican = Bush.
Source: 2004 ANES.

(more services) to the left of the spectrum and the conservative position (reduce spending) to the right end of the continuum. Ideological labels can also be applied to the B-scale. For example, a respondent with $B_i = 1$ ($y_i < v_{iD}$) can be described as being more liberal than the Democrat.

The margins of Table 3 show the percentage of respondents who intend to vote for the Democratic presidential candidate by self-rating (rows) and by B-scale values (columns).[13] The margins for the self-ratings y_i reveal no indication of a problem with the scale, with increasing support for more services producing the expected monotonic increase in support for the Democratic candidate. The margins for the B-scale show a similar monotonic relationship with vote intentions.

Looking at the vote intentions conditional on the combination of self-rating value and B- scale value shown within Table 3, however, reveals intransitivities along the self-rating scale.[14] For example, consider the two types of respondents illustrated in Figure 7. At the top are individuals of type i who rated themselves as $y_i = 5$ and between the presidential candidates ($B_i = 2$). At the bottom are individuals of type j who rated themselves as wanting more services, $y_j = 4$, but placed themselves as more conservative than both candidates ($B_j = 3$). If the self-rating scale were used in a comparable manner across individuals, then we would expect a lower level of support for the Democrat with a higher self-rating in order to reflect a preference for less spending. We see in Table 3 that these two types of individuals, i and j, have the opposite ordering of Democratic support than expected by the nominal self-ratings, y_i. Among those with $y_i = 5$ and $B_i = 2$, 46% support the Democrat, while only 7% of those with $y_j = 4$ and $B_j = 3$ do the same.[15] Note, in contrast, that the order of vote intentions is consistent with the B-scale values.

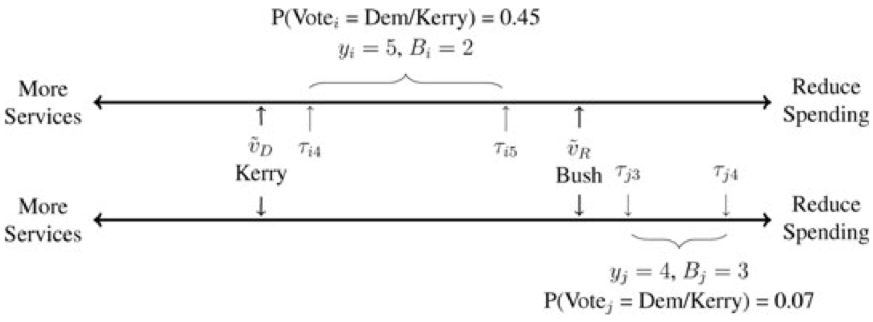
Further evidence that the B-scale is superior at ordering individuals is that an individual's self-rating contains no information predictive of vote intention net of its use in constructing the individual's B-scale value. Given an individual's B-scale value, there is no association between the raw scale responses and vote intentions. For example, for the set of respondents with $B = 1$ (column 1), there is no significant association across the raw scale responses and the probability of supporting Kerry, p = .32 as measured by a χ^2 goodness-of-fit statistic. Similarly, conditional on other columns, there is no significant association across the raw scale values, $B = \{1, 2\} : \text{p} = .69$ $B = 2 : \text{p} = .1$; $B = \{2, 3\} : \text{p} = .16$; $B = 3 : \text{p} = .53$. The opposite is not true: there is a significant monotonic association across values of the B-scale conditional on a

[13] The table uses data from the 2004 pre-election ANES survey and summarizes the behavior of respondents who rated themselves and both general election presidential candidates, ordered the Democrat closer to services than the Republicans, and had a vote intention for either the Republican or Democratic presidential candidate. Of respondents with a complete set of responses (self and both candidates), 96.5% give ratings of the 2004 presidential candidates that are consistent with the use of the B method. Only 19% of these same respondents give responses that are consistent with the use of C.

[14] The number of observations in each cell are as follows by row: $y_i = 1$, (55 , 21 , 0 , 0 , 0); $y_i = 2$, (39 , 36 , 9 , 0 , 0); $y_i = 3$, (17 , 51 , 64 , 9, 1); $y_i = 4$, (4 , 13 , 88 , 27, 14); $y_i = 5$, (1 , 0 , 26 , 23 , 39); $y_i = 6$, (0 , 0 , 6 , 12 , 34); $y_i = 7$, (0 , 0 , 0 , 3 , 29).

[15] The probability of the proportions being equal is less than 0.01.

FIGURE 7 **Comparing Vote Intentions by Values of the *B*-Scale and Self-Rating *y***



Note: Higher values of ***y*** and ***B*** indicate preference for fewer services and reduced spending.

particular self-rating value ($y = 4 : \text{p} < .001$; $y = 3 : \text{p} = .003$; $y = 2 : \text{p} = 0.02$).

The strong association between the B-scale and vote intentions has a theoretical foundation, since the B-scale and spatial models of voting share a similar logic. In a proximity-based, spatial voting model, voters are assumed to choose candidates closest to them, either as a deterministic or probabilistic function.[16] Defining $\Delta =\mid \tilde{y}_i - \tilde{v}_{iD} \mid - \mid \tilde{y}_i - \tilde{v}_{iR} \mid$ then, a voter is more likely to choose candidate D over candidate R if $\Delta < 0$. Conversely, a voter is more likely to choose candidate R if $\Delta > 0$. The B-scale extracts information about the spatial locations of voters relative to candidates without assuming that survey respondents agree on the meaning of scale categories or that categories represent interval values.[17] The nonparametric value $B = 1$ groups together respondents whose ideal point is less than the perceived location of the Democrat ($\tilde{y}_i < \tilde{v}_D$) and therefore on average finds Democrats more appealing, $\Delta < 0$.[18] Similarly, $B = 3$ groups together respondent with $\Delta > 0$: the Republican is more appealing. The value $B = 2$ does not illuminate which candidate is perceived to be close to the candidate, although the Democratic vote share should be between that of $B = 1$ and $B = 3$.[19]

Comparisons with Parametric Models Using Anchors

While it is beyond the scope of this article to fully review and compare the range of alternative models, I briefly note key differences between the nonparametric comparisons discussed here and two prominent parametric models of anchoring objects.

The parametric model of Aldrich and McKelvey (1977) assumes that perceptions of anchoring objects are subject to random perturbations and that the ordinal scale categories have metric values that are subject to idiosyncratic linear transformations across individuals.[20] The model proposed by King et al. (2004) shares a similar notion of a random perception error of anchoring objects, but it models the ordinal nature of the data. In both models, the scales of respondents are bridged by assuming a common mean location of the anchoring objects for all respondents.

The parametric and nonparametric models differ in terms of the information they recover. First, both parametric models produce estimates of the average location of anchoring objects in the normalized latent space. In studies using candidates or other real anchors, these locations may be of interest. In the studies using anchoring vignettes, the locations of hypothetical individuals are rarely of inherent interest. Since nonparametric methods use only relative information, no information about anchor locations is recovered.

Second, the parametric models produce metric information about the expected location of each respondent. The accuracy of this additional information for

[16] For a recent review of the alternative spatial theories of how voters make voting decisions based on the stances of candidates, see Tomz and Van Houweling (2008).

[17] Approaches which address one or the other of these issues include those by Aldrich and McKelvey (1977) and Mebane (2000).

[18] If the vote model is probabilistic, the Democrat is more appealing in terms of the expectation of the function.

[19] It is possible to state that Δ is less negative (and potentially positive) on average for these respondents than for respondents who have $B = 1$ and that Δ is less positive (and potentially negative) than for respondents who have $B = 3$.

[20] The statistical model is a principal components estimator for the anchoring object location parameters, combined with a regression estimate of the respondent self-evaluation parameters.

comparing individuals should, however, be carefully examined in light of the potential failures of the modeling assumptions. While the Aldrich-McKelvey model has been shown to produce accurate estimates of the locations of anchors under a broad range of violations of the model's statistical assumptions (Palfrey and Poole 1987), the difficulties of accurately recovering the relative distance between individual respondents have been noted (Aldrich and McKelvey 1977; Palfrey and Poole 1987). For example, a bimodal distribution of attributes may appear unimodal and vice versa.

The parametric model proposed by King et al. (2004) is unique in estimating the location of the cutpoints of the ordinal scale as a function of respondent-specific covariates. Individuals with the same covariates, however, are assumed to use the scale in exactly the same way. As such, this model assumes that DIF may exist across different parameterized strata of respondents, but not within a stratum. In contrast, both the B-scale and Aldrich-McKelvey model allow each individual to use the scale differently. Adjudicating the trade-off between the costs of imposing such homogeneity assumptions and the benefits of allowing cutpoints to vary by covariates is an area in which we do not currently have much guidance.

Overall, the parametric models provide more information than the nonparametric approach by invoking assumptions. The credibility of the inferences depends on the ability to justify these assumptions in a given empirical study. The nonparametric method B-scale provides lower bounds both for the assumptions needed to make credible comparisons and the amount of information yielded by anchoring objects. Each additional bit of information regarding the relative ordering of individuals is bought in the currency of assumptions.

Conclusion

In this article, I show credible interpersonal comparisons can be made without the traditional assumption that all respondents in a survey agree on the use of an ordinal scale. A nonparametric anchor-based method can provide credible comparisons while requiring only weak assumptions about how respondents disagree on the location of vignettes. Two empirical studies illustrate the importance of accounting for differences in the use of an ordinal scale.

I also derive empirical tests of the assumptions needed to produce credible comparison using alternative nonparametric anchor-based methods. A key advantage of anchoring vignettes is that a researcher can empirically evaluate the suitability of a vignette for making comparisons. With this information, researchers can adapt their existing statistical analysis and adjust, if necessary, the design of the vignettes for use in a subsequent survey sample.

References

Aldrich, John H., and Richard D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1): 111–30.

Brady, Henry E. 1989. "Factor and Ideal Point Analysis for Interpersonally Incomparable Data." *Psychometrika* 54(2): 181–202.

Brady, Henry E. 1990. "Dimensional Analysis of Ranking Data." *American Journal of Political Science* 34(4): 1017–48.

Brady, Henry E., and Paul M. Sniderman. 1985. "Attitude Attribution: A Group Basis for Political Reasoning." *American Political Science Review* 79(4): 1061–78.

Chevalier, Arnaud, and Antony Fielding. 2011. "An Introduction to Anchoring Vignettes." *Journal of the Royal Statistical Society: Series A* 174(3): 569–74.

Grzymala-Busse, Anna M. 2007. *Redeeming the Communist Past*. Cambridge: Cambridge University Press.

Hopkins, Daniel J., and Gary King. 2010. "Improving Anchoring Vignettes." *Public Opinion Quarterly* 74(2): 201–22.

Javaras, Kristin N., and Brian D. Ripley. 2007. "An 'Unfolding' Latent Variable Model for Likert Attitude Data: Drawing Inferences Adjusted for Response Style." *Journal of the American Statistical Association* 102(478): 454–63.

Kapteyn, Arie, James P. Smith, and Arthur Soest. 2007. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *American Economic Review* 97(1): 461–73.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1): 191–207.

King, Gary, and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes." *Political Analysis* 15: 46–66.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mebane, Walter R., Jr. 2000. "Coordination, Moderation, and Institutional Balancing in American Presidential and House Elections." *American Political Science Review* 94(1): 37–57.

Palfrey, Thomas R., and Keith T. Poole. 1987. "The Relationship between Information, Ideology, and Voting Behavior." *American Journal of Political Science* 31(3): 511–30.

Rossi, Peter E., Zvi Gilula, and Greg M. Allenby. 2001. "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical

Approach." *Journal of the American Statistical Association* 96(453): 20–31.

Salomon, Joshua A., Ajay Tandon, and Christopher J. L. Murray. 2004. "Comparability of Self-Rated Health: Cross-Sectional Multi-Country Survey Using Anchoring Vignettes." *British Medical Journal* 328(7434): 258.

Tomz, Michael, and Robert P. Van Houweling. 2008. "Candidate Positioning and Voter Choice." *American Political Science Review* 102(3): 303–18.

Van Soest, Arthur, Liam Delaney, Colm Harmon, Arie Kapteyn, and James P. Smith. 2011. "Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjective Questions." *Journal of the Royal Statistical Society: Series A* 174(3): 575–95.

Wand, Jonathan, Gary King, and Olivia Lau. 2011. "anchors: Software for Anchoring Vignette Data." *Journal of Statistical Software* 42(6): 1–25.

Supporting Information

Additional Supporting Information may be found in the online version of this article:
Table 1: Monte Carlo parameters.
Table 2: Percentage of all comparisons that are misordered: 95th percentile intervals from 500 simulations, each with 1000 observations. Lower numbers are better, with zero indicating no errors.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.