Advance Access publication December 11, 2006

Political Analysis (2007) 15:46–66
doi:10.1093/pan/mpl011

Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes

Gary King
Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138
e-mail: king@harvard.edu (corresponding author)

Jonathan Wand
Department of Political Science, Encina Hall, Room 308 West, Stanford University, Stanford, CA 94305-6044
e-mail: wand@stanford.edu

When respondents use the ordinal response categories of standard survey questions in different ways, the validity of analyses based on the resulting data can be biased. Anchoring vignettes is a survey design technique, introduced by King et al. (2004, Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review* 94 [February]: 191–205), intended to correct for some of these problems. We develop new methods both for evaluating and choosing anchoring vignettes and for analyzing the resulting data. With surveys on a diverse range of topics in a range of countries, we illustrate how our proposed methods can improve the ability of anchoring vignettes to extract information from survey data, as well as saving in survey administration costs.

1 Introduction

Researchers have tried to ameliorate the problems of interpersonal and cross-cultural incomparability in survey research with careful question wording, translation (and back translation), focus groups, cognitive debriefing, and other techniques, most of which are designed to improve the survey question. In contrast, *anchoring vignettes* is a technique (developed by King et al. 2004) designed to ameliorate problems that occur when different groups of respondents understand and use ordinal response categories—such as (1) strongly disagree, (2) disagree, (3) neutral, (4) agree, or (5) strongly agree—in different ways. When one group of respondents happen to have comparatively higher standards for what constitutes the definition of "strongly agree," for example, they will report systematically lower levels of agreement than another group. Yet, some people obviously differ in

Authors' note: Our thanks go to Amos Golan, George Judge, Doug Miller, Chris Murray, Olivia Lau, and Josh Salomon for many helpful suggestions; Emmanuela Gakidou and Ajay Tandon for data; Dan Hopkins for insightful research assistance; and National Institute on Aging/National Institutes of Health (for grant P01 AG17625-01 to King) and the Robert Wood Johnson Scholars in Health Policy Research program at the University of Michigan (to Wand) for generous research support.

optimism, agreeability, mood, propensity to use extreme categories, and other characteristics, and so doing something about this "response-category differential item functioning" (or DIF) should be a high priority for researchers.

The methodology of anchoring vignettes attack DIF with new types of supplemental survey questions that make it possible to construct a common scale of measurement across respondents, along with specially designed statistical methods for analyzing the resulting data. Anchoring vignettes have now been used to measure numerous concepts and have been implemented in surveys in over 80 countries by a growing list of survey organizations and researchers in a variety of academic fields.[1] In this paper, we describe this approach and develop improved statistical methods for analyzing, evaluating, and selecting anchoring vignettes that require fewer assumptions, can extract more information from the same survey questions, and should save in research costs.

Previous applications of anchoring vignettes have used as many as 12 vignettes per self-assessment question. However, adding this many additional questions for each self-assessment, or even the five used by King et al. (2004), may be prohibitively expensive in some surveys and, we show, are often unnecessary to correct DIF. In some cases, the necessary correction may be achieved with a single vignette. In other cases, more vignettes may be informative. In all cases, the methods we introduce to evaluate the efficacy of each vignette to improve interpersonal incomparability should be of direct practical use to researchers.

We begin in Section 2 by summarizing the anchoring vignettes approach. Section 3 then provides a more general definition of this approach, a new formalization, a more generally applicable analytical method, and an illustration of the methods offered. Section 4 then develops new ways of evaluating the information available in a set of anchoring vignettes and detecting vignettes that may be empirically unnecessary or less useful and those that may violate key assumptions of the technique. Section 5 gives several examples of evaluating vignettes in practice, and Section 6 concludes. The new information our methods reveal greatly increases the efficacy of the nonparametric estimator, making it a powerful alternative to the parametric approach and one that requires considerably less stringent assumptions.

2 The Technique of Anchoring Vignettes

A variant of a question asked in numerous surveys seeks to measure what political scientists call *political efficacy*:

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

For this question, as most others, political scientists typically theorize that each respondent has an *actual* level of efficacy that may differ from the *reported* level due to measurement, respondent "considerations," the survey interview setting, positivity bias, or other aspects of DIF. Political scientists typically view the actual level of efficacy as a relatively objective psychological state: Respondents who genuinely feel politically efficacious are more likely to participate in politics, write letters to public officials, contribute to political campaigns, debate policy with their friends, and feel more generally part of the political system. The difference between true underlying perceived political efficacy and the reported level may differ due to a variety of measurement factors including idiosyncratic

[1] A library of anchoring vignette examples used in these and other surveys, and other materials, can be found at http://gking.harvard.edu/vign/. Our software for analyzing anchoring vignettes is at http://wand.stanford.edu; see Wand, King, and Lau (forthcoming).

considerations of the respondent or variations in the survey interview setting. There may be more systematic biases within groups, such as positivity bias or different standards for the categories on the scale. We address the issue of these systematic differences in the use of scale within groups, referred to generically as DIF.

King et al. (2004) measured political efficacy with this question in surveys from China and Mexico. Remarkably, the raw responses indicate that the citizens of (democratic) Mexico judge themselves to have substantially lower levels of political efficacy than citizens of (nondemocratic, communist) China judge themselves. For example, more than 50% of Mexicans but fewer than 30% of Chinese reported in these surveys having no say in the government. The massively divergent levels of actual freedom and democracy in the two countries strongly suggest exactly the opposite conclusion and thus a potential problem with the survey question or how it is understood. King et al. (2004) argue that both the levels of efficacy and the standards for any particular level (i.e., DIF) vary between the countries, and as such, the reported self-responses are incomparable between countries.

In the present example, King et al. (2004) explain the apparent paradox with the Chinese in fact having lower actual levels of political efficacy than the Mexicans. However, the Chinese report higher *levels* of say in government because they have lower *standards* for what counts as satisfying the level described by any given response category.

Reported survey responses alone cannot be used to address the issue of comparability of scales across groups. Addressing DIF requires some measure or benchmark for the actual unobserved level of the variable that the survey question is intended to measure. To provide a common reference point for people applying different standards for the same scale, one approach is to ask each respondent an additional anchoring vignette question after the self-assessment such as

> [Moses] 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 [Moses] have in getting the government to address issues that interest him?

with the same response categories as the self-assessment. (To increase the likelihood that respondents think of the vignette as describing someone like themselves except for the content of the vignette, the hypothetical individuals are given names appropriate to the language and culture and when possible also indicating the same sex as the respondent.)

Because the reported answer to the self-assessment question includes both the actual level of efficacy and DIF, we cannot separate the two without further information. The anchoring vignette question provides that additional information, since Moses has the same actual level of efficacy no matter which respondent in which country is queried. Thus, any systematic variation in answers about the Moses question can only be due to DIF. By assuming only that the DIF that a respondent applies to his or her self-assessment is the same as the DIF that this person applies to the vignette question, we can "subtract off" the DIF from the self-assessment to yield a DIF-free estimate of the actual level of political efficacy. In the simplest (nonparametric) method of analysis, we can correct for DIF by recoding the self-assessment response *relative* to the vignette response as (1) less than, (2) equal to, or (3) greater than the vignette response. In fact, although the raw responses had the Chinese judging themselves to have considerably more efficacy than the Mexicans judged themselves to have, the DIF-corrected responses indicates the reverse: whereas only about 12% of Mexicans judged themselves to have less political efficacy than "Moses" in the vignette above, 40% of the Chinese judged themselves to have less efficacy than Moses who suffers in silence (see King et al. 2004, 196).

To be more explicit, two assumptions enable us to regard this trichotomous recoded answer as DIF free and freely compared across different groups of people. The first is *response consistency*, which is the assumption that each respondent uses the survey response categories in the same way to answer the anchoring vignette and self-assessment questions. Different people may have different types of DIF, but any one person must apply the same DIF in approximately the same way across the two types of questions. Second is *vignette equivalence*, which is the assumption that the level of the variable represented in the vignette is understood by all respondents in the same way apart from random measurement error. Of course, even when respondents understand vignettes in the same way on average, different respondents may apply their own unique DIFs in choosing response categories.

Thus, unlike almost all existing survey research, anchoring vignettes allow and ultimately correct for the DIF that may exist when survey respondents choose among the response categories, but they assume like most previous research the absence of DIF in the "stem question." It seems reasonable to focus on response-category DIF as the main source of the problem because the vignettes describe behaviors or psychological states intended to be more objective and for which traditional survey design advice to avoid DIF (such as pretesting, cognitive debriefing, and writing questions concretely) is likely to work better (although the methods described here may also help to identify problematic vignettes). In contrast, response categories describe much more subjective feelings and attitudes attached to words or phrases that are by their nature imprecise; the responses should therefore be harder to lay out in concrete ways and avoid DIF without anchoring vignettes. This point also seems consistent with the finding in the literature that traditional survey design advice can work especially badly for Likert-type response scales (Schwarz 1999).

One issue with the DIF correction in the example thus far is that the five-category political efficacy response is reduced to only a three-category DIF-corrected variable, and so information may have been lost. As it turns out, however, we can recover additional information by adding more vignettes. For example, King et al. (2004) also use this vignette:

> [Imelda] lacks clean drinking water. She and her neighbors are drawing attention to the issue by collecting signatures on a petition. They plan to present the petition to each of the political parties before the upcoming election.

with the same survey question and response categories. If survey respondents rank the two vignettes in the same order, we can create a DIF-free variable by recoding the self-assessment into five categories: (1) less than Moses, (2) equal to Moses, (3) between Moses and Imelda, (4) equal to Imelda, and (5) greater than Imelda. Of course, we can obtain considerably more discriminatory power by adding more vignettes. In the political efficacy example, King et al. (2004) use five vignettes presented to the respondent in random order.

More vignettes of course also come with an additional assumption: that all the respondents understand the vignettes as falling along the same unidimensional scale, even if they do not use the scale and response categories in the same way. In the vignettes above, for example, Imelda presumably has more political efficacy than Moses, but if many respondents indicated otherwise, this might be an indication of multiple dimensions being tapped by the vignettes. Even if unidimensionality holds as assumed, using multiple vignettes gives rise to other potential challenges. For example, random error in perceptions or responses may produce inconsistencies in vignette rankings. Other respondents may not perceive the difference between some vignettes and may give them tied responses. We deal with these issues in this study.

Finally, we note that asking vignettes may seem like an expensive technique since it requires adding multiple questions to a survey to correct for each self-assessment question. In fact, however, King et al. (2004) develop a statistical technique that enables one to ask anchoring vignettes of only a small random subsample and to still statistically correct for DIF using parametric assumptions; the same technique can also be applied to respondents who were not asked all questions. Alternatively, one can include the vignettes on the pretest and not the full survey or include only a subset on the main survey. Or one can add, for each self-assessment, only one additional item, where each quarter of the respondents are asked a different vignette (although this presumes that the assumptions underlying the use of the vignettes to anchor the self-responses have already been verified with other data). It is also possible to choose vignettes adaptively, so that we give each respondent different vignettes depending on his or her response to previous questions. A parametric method is available that enables one to save survey administration costs in these and other ways. The mostly nonparametric methods we introduce here minimize statistical modeling assumptions and so provide a useful complement to parametric modeling.

3 Estimation Strategy

Our estimation strategy first extracts all known information via a fully nonparametric approach, without making any additional assumptions. We then supplement this nonparametric information with a parametric approach that, by making some additional assumptions, extracts additional information. The parametric supplement also makes it easy to analyze the nonparametric data to compute predictions, causal inferences, and counterfactual questions.

3.1 *The Nonparametric Estimator*

We now offer a simple generalization of the nonparametric approach, described as simple recodes in Section 2 and King et al. (2004). Let y be the self-assessment response and $z_1, \ldots, z_J$ be the J vignette responses, for a single respondent. The same discrete ordinal response choices (e.g., unlimited say, a lot of say, some say, etc.) are offered to the respondent for each of the questions. For respondents with consistently ordered rankings on all vignettes ($z_{j-1} < z_j$, for $j = 2, \ldots, J$), we create the DIF-corrected self-assessment by the recodes described in Section 2, which in mathematical notation is

$$C = \begin{cases} 1 & \text{if } y < z_1 \\ 2 & \text{if } y = z_1 \\ 3 & \text{if } z_1 < y < z_2 \\ \vdots & \vdots \\ 2J+1 & \text{if } y > z_J. \end{cases} \tag{1}$$

Since this section considers one survey respondent at a time, we omit the subscript denoting the respondent.[2]

The only remaining issue is how to generalize equation (1) to allow for respondents who give tied or inconsistently ordered vignette responses. We do this by first checking

[2]The ordering of vignettes is normally chosen by the researchers, but it is also possible to draw upon a consensus ordering by the respondents, so long as only one ordering is used for all respondents for the analysis. Differences between hypothesized ordering of the researchers and the consensus ordering may fruitfully be used for diagnosing problems in the survey instruments, particularly when translating the questions for use in different languages.

Table 1 All examples with two vignettes: this table gives calculations for the nonparametric estimator C for all possible examples (sans nonresponse) with two vignette responses, z_1 and z_2 (intended to be ordered as $z_1 < z_2$), and a self-assessment, y

Example	*Survey responses*	*1* $y < z_1$	*2* $y = z_1$	*3* $z_1 < y < z_2$	*4* $y = z_2$	*5* $y > z_2$	C
1	$y < z_1 < z_2$	1	0	0	0	0	{1}
2	$y = z_1 < z_2$	0	1	0	0	0	{2}
3	$z_1 < y < z_2$	0	0	1	0	0	{3}
4	$z_1 < y = z_2$	0	0	0	1	0	{4}
5	$z_1 < z_2 < y$	0	0	0	0	1	{5}
6	$y < z_1 = z_2$	1	0	0	0	0	{1}
7	$y = z_1 = z_2$	0	1	0	1	0	{2, 3, 4}
8	$z_1 = z_2 < y$	0	0	0	0	1	{5}
9	$y < z_2 < z_1$	1	0	0	0	0	{1}
10	$y = z_2 < z_1$	1	0	0	1	0	{1, 2, 3, 4}
11	$z_2 < y < z_1$	1	0	0	0	1	{1, 2, 3, 4, 5}
12	$z_2 < y = z_1$	0	1	0	0	1	{2, 3, 4, 5}
13	$z_2 < z_1 < y$	0	0	0	0	1	{5}

which of the conditions on the right side of equation (1) are true and then summarize C with the vector of responses that range from the minimum to maximum values among all the conditions that hold true. Values of C that are intervals (or vector valued), rather than scalar, represent the set of inequalities over which the analyst cannot distinguish without further assumption; we refer informally to cases that have interval values as being censored observations.

Table 1 gives all 13 examples that can result from two vignette responses and a self-assessment. Examples 1–5 have both vignettes correctly ordered and not tied, with the result for C being a scalar. The vignette responses are tied in examples 6–8, which produces a censored value for C only if the self-assessment is equal to them. Examples 9–13 are for survey responses that incorrectly order the vignettes. Within each set, the examples are ordered by moving y from left to right.

This generalized definition for C clarifies the impact of ties (as in examples 6 and 8) and inconsistencies (as in examples 9 and 13) among the vignettes that occur in a group strictly greater than or less than y. Note that all four of these examples in the table have scalar (uncensored) values for C. This is appropriate since we might reasonably expect respondents to be more likely to give some tied or inconsistent answers among vignettes that are far from their own self-assessment even when they correctly rank the vignettes that matter near their own value. For example, if we are measuring height and a respondent knew his or her height to within an inch, he or she still might have difficulty correctly ranking the heights of two trees 200 and 206 feet tall, swaying in the breeze. Yet, the same respondent would presumably have no difficulty understanding that both trees are taller than himself or herself. We thus regard the respondent's misordering of vignettes in this way to not have any censoring effect on the estimate.

3.2 *A Parametric Supplement*

The nonparametric estimator discussed in Section 3.1 recodes the vignettes and self-assessment questions into a single DIF-free variable C. Since this unusual dependent

variable has a scalar value for some observations and multiple values for others, we now develop a way to analyze such variables. In fact, the method described here would also apply to survey questions that permitted respondents to choose among any range of response categories, rather than a single one. Such questions do not appear to be used frequently, but they may be useful in tapping into respondents' central tendencies as well as uncertainties in a single question.

Consider the simple example of drawing a histogram of the results of C. If C were entirely scalar valued, we would do what we always do: Sort the values of C into each category j ($j = 1, \ldots, 2J + 1$), compute the proportion p_j in each, and plot one bar for each category with size proportional to p_j.

The key issue is what to do when C is a range (or vector valued) instead of a scalar. One simple possibility would be to discard the vector-valued observations. This would obviously waste information, at best, and of course, it may introduce selection bias as well. Another simple approach would be to take the vector values and spread the area evenly across all the members of the vector-valued set. This is what King et al. (2004) did, but the allocation rule they employed is obviously an assumption that should be considered carefully rather than used automatically. If the assumption is wrong, it will bias results toward a uniform density (a flat histogram) and so may cause one to obliterate features of the true frequency distribution.

In this section, we move beyond these simplistic approaches and seek to allocate the vector-valued responses to categories as best as possible so that we simultaneously use information about both the scalar and vector values of the variable C. Our approach for displaying C in a single-dimensional representation, such as a histogram, is to distribute each vector-valued response according to the proportion of "similar" respondents who chose the categories spanned by the vector. We thus estimate the proportion in each of the spanned categories using a generalization of the ordered probit model that allows "censored" values corresponding to our vector values. In this way, we obtain an estimate of the proportion of the sample in each category of C, which can be used to construct a histogram or for other analyses.

To define our approach, we begin with the classic ordered probit and then generalize in two ways. First denote Y_i (for respondent $i = 1, \ldots, n$) as a continuous unobserved dependent variable and x_i as a vector of explanatory variables (and for identification, with no constant term). If this is used as a predictive rather than causal model, as would be appropriate if the goal were to draw a histogram or for other descriptive purposes, whatever variables that might be associated with the outcome should be included. Covariate selection for causal purposes would follow the same rules as for a traditional ordered probit model. Then, we model Y_i as conditionally normal with mean $x_i\beta$ and variance 1. If Y_i were observed, the maximum likelihood estimate of β would simply be the coefficient from a linear regression of Y_i on x_i.

However, instead of observing Y_i, we instead see C_i through a specific observation mechanism. Thus, for scalar values, the mechanism is

$$C_i = c \quad \text{if } \tau_{c-1} \leq Y_i < \tau_c \tag{2}$$

with thresholds τ_c (where $\tau_0 = -\infty$, $\tau_{2J+1} = \infty$, and $\tau_{c-1} < \tau_c$ and for $c = 1, \ldots, 2J + 1$). Under this model, the probability of observing an outcome in category c is simply

$$\Pr(C = c \mid x_0) = \int_{\tau_{c-1}}^{\tau_c} N(y \mid x_0\beta, 1)\mathrm{d}y \tag{3}$$

for a given vector of values of the explanatory variables, x_0. This of course is exactly the ordered probit model.

We now generalize this model in the first of two ways by adding notation for vector values of C, which we do by altering the observation mechanism in equation (2) to

$$C_i = c \quad \text{if } \tau_{\min(c)-1} \leq Y_i < \tau_{\max(c)}. \tag{4}$$

This *censored ordered probit* model can also be used as a means of estimating causal effects under the maintained assumptions. One would merely estimate the model and interpret β exactly as under ordered probit (such as by using the procedures in King, Tomz, and Wittenberg 2000). Researchers could estimate the probability of a respondent answering in any individual category by using equation (3).[3]

To compute histograms, and for occasional other purposes, we are able to add considerable robustness by introducing a second generalization of the classic ordered probit model that conditions the calculation of the probability of being in a specific (single) category c on the observed vector or scalar c_i (using a technique analogous to that in King [1997] and King et al. [2004], Appendix B). This *conditional* calculation is simply

$$\Pr(C = c \mid x_0, c_i) = \begin{cases} \frac{\Pr(C=c|x_0)}{\sum_{a \in c_i} \Pr(C=a|x_0)}, & \text{for} \quad c \in c_i, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where the usual *unconditional* probability $\Pr(C = c|x_0)$ is defined in equation (3). The new expression in equation (5) conditions on c_i by normalizing the probability to sum to one within the set c_i and zero outside that set. For scalar values of c_i, this expression simply returns the observed category: $\Pr(C = c|x_i, c_i) = 1$ for category c and 0 otherwise. For vector-valued c_i, equation (5) puts a probability density over the categories within c_i, which in total sum to one.

A virtue of this method is that predictions of scalar-valued observations are fixed at their observed values, independent of the model, and thus are always correct no matter how misspecified the model. In addition, predictions of vector-valued observations are restricted to within their observed range, also with certainty and independent of modeling choices. The method uses all available information in the self-assessment, vignettes, and explanatory variables to estimate the distribution of frequencies for the vector-valued observations rather than merely assuming an arbitrary distribution ex ante. The robustness of this conditional approach is directly related to size of the range. Analogous to the influence of the proportion of missing data in multiple imputation, a smaller range of vector values offers relatively greater confidence in the resulting histogram independent of the modeling assumptions, since the result is bounded by the known information in the data (Heitjan and Rubin 1990; King et al. 2001).

3.3 *Empirical Illustration*

King et al. (2004) analyze political efficacy data in China and Mexico and show that, with an anchoring vignettes DIF correction, the ranking of the two countries switches compared

[3] If two response categories are observed as part of the vector-valued C for the same set of observations and for no others, some of the threshold parameters are not identified. The nonidentification in this unusual special case is partial in the sense that the likelihood still indicates the mass in the sum of the two categories. The problem can be easily addressed by combining the two categories or by using information to construct a prior for the τ's.

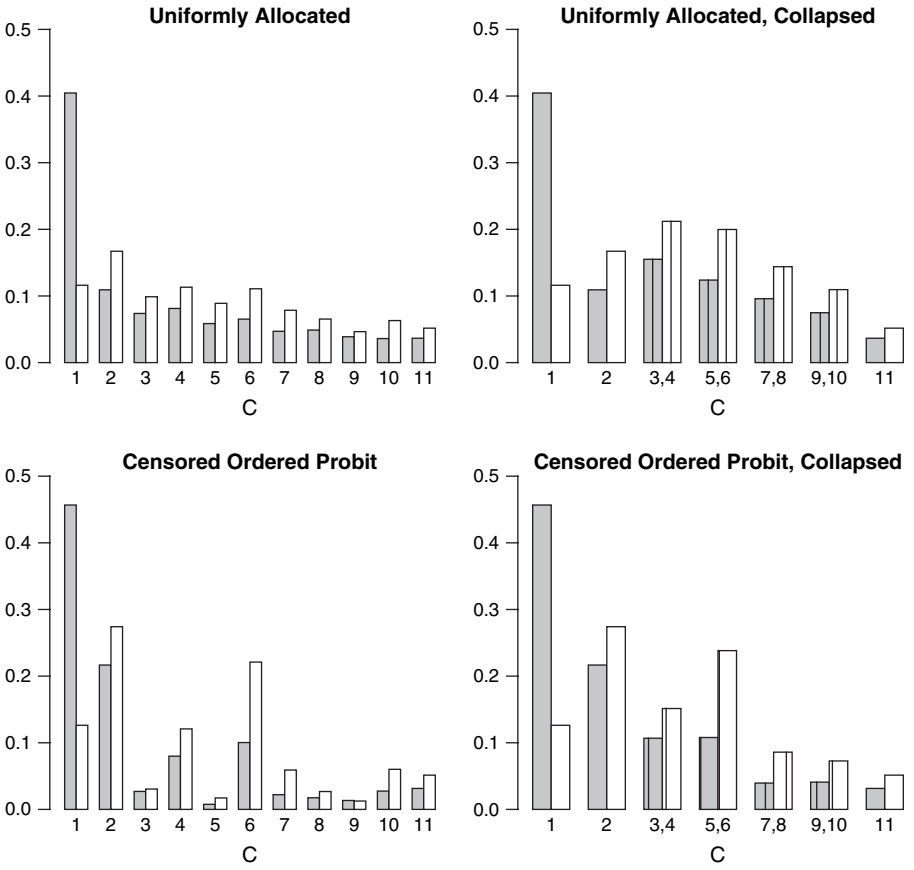

Fig. 1 Two methods of analyzing a DIF-free measure of political efficacy. The horizontal axis on each graph ranges from a low level of say in government on the left to a high level on the right. The top two graphs allocate vector values of C uniformly over the range, whereas in the bottom two, they are allocated according to the censored ordered probit model. For the right two graphs, categories have been combined for visual clarity; the fraction in each of the combined categories is denoted by a vertical line dividing the relevant bar on the graph in proportion to the sizes of the two categories. The shaded bars represent Chinese data, the clear bars Mexican.

to the raw survey responses, with the result being much more in line with reality. However, their analysis deals with ties and inconsistencies by spreading them uniformly over vector-valued C observations, which is not a plausible approach in general. The top left graph in Fig. 1 reproduces the figure from their article, and the tendency toward a uniform histogram is obvious, although of course we will need to go beyond the nearly uniform observed histograms to show evidence of bias.

The bottom two graphs in Fig. 1 present reanalyses of these same data using the conditional calculations from the censored ordered probit model.[4] Because of some very slightly populated categories, we combined categories in the right bottom graph from the same model (and the right top graph from the uniform model for comparison). That is,

[4]The plotted proportions are the conditional fitted probabilities from a censored probit model with covariates age, years of education, male dummy, and a China dummy. The cutpoints for the censored ordered probit are assumed constant for all observations.

because the survey question has only five response categories, and since most responses are clustered at the low end of efficacy, some odd-numbered values of C have very small numbers of respondents. The result is not a methodological problem but merely an esthetic one, since comparing histograms with sawtooth patterns can be visually complicated. Combining selected adjacent categories solves this problem. So that no information is lost by this procedure, we add a vertical line that splits any bar representing two categories with area proportional to the size of the two categories.

If the uniform assumption of the top graphs were correct, the censored ordered probit methods would pick up that information and the top and bottom rows of histograms would look fairly similar. In fact, the results clearly indicate that the censored ordered probit methods are picking up information not in the uniform allocation approach. Whereas the main difference between Mexico and China in the uniform graphs is the spike in category $C = 1$ for China, the ordered probit methods reveal considerably more texture. In fact, the preferred conditional approach in the bottom right graph illustrates very different patterns for the two countries. China shows a relatively smooth exponential decline, whereas the Mexico histogram indicates that the bulk of respondents have markedly higher levels of political efficacy. The result is much more consistent with what we know about the two countries' divergent levels of efficacy than the more modest differences revealed by the uniform assumption graph.

4 Evaluating Vignettes

Asking a random sample of Americans how often they run two marathons in a week obviously would not yield much information about how much the respondents exercise. Similarly, different vignettes for a given self-assessment question vary in the information they yield about respondents. In this section, we develop a useful and relatively natural measure of the information revealed by the set of vignettes used. This measure can be used to assist in choosing a subset of vignettes for more detailed analyses or subsequent surveys.

We begin by conceptualizing each respondent's actual level as falling on an unobserved unidimensional continuous scale broken up into the $2J + 1$ categories defining C. The set of vignettes used to create C gives meaning to each of its categories. Our task is to choose a set of vignettes to assign the most useful set of meanings to the categories of C, which we shall do by developing a measure of the *discriminatory power* of or *information* in the set of vignettes. We explain this first for the simpler case of scalar values of C and subsequently for scalar and vector values of C.

4.1 *Information in Scalar Values of C*

For simplicity, we assume in this section that C contains a single value for each observation (which occurs in the absence of ties or inconsistencies among the vignettes). We denote by p_j the proportion of observations in category j of C (for $j = 1, \ldots, 2J + 1$). The resulting set of proportions in the categories, $\{p_1, \ldots, p_{2J+1}\}$, define a frequency distribution of responses, such as might be portrayed via a histogram (with p_j proportional to the height of bar j).

As in the case of a survey question about running two marathons in a week, when a set of vignettes defines C such that all respondents fall into only one of its categories, C conveys the least possible information about the underlying continuous scale. It is true that we get a lot of information about the fitness of an individual who responds that he or she does run two marathons weekly, but we will not likely see many individuals of this type and so the expected amount of information is near zero. In contrast, C is best able to discriminate

among respondents on the underlying continuous scale when vignettes are chosen to define categories of C such that respondents are sorted equally across all the categories:

$$p_1 = p_2 = \cdots = p_{2J+1} = \frac{1}{2J+1}. \tag{6}$$

Our immediate goal, then, is to define a measure of the informativeness or discriminatory power of a set of vignettes for a given self-assessment question. In other words, we need to define a continuous function $H(\cdot)$ that takes the set of frequencies, $p_1, \ldots, p_{2J+1}$, as arguments and returns a real number indicating the information in the frequencies and hence the information in the vignettes that define C. This real number should be at a minimum, which we will denote as zero, when one category contains all respondents and at a maximum when respondents are spread across the categories equally. To avoid hard-to-justify application-specific assumptions about the relative importance of each of the categories of C, we require the arguments of H to be symmetric in that if we reordered the p's, H would return the same number. The symmetry requirement could easily be dropped if such information were available for a particular application. Of course, many functions satisfy these simple criteria, but we can narrow down the search to a unique choice by adding two more requirements, both needed to cope with the possibility of considering and comparing this measure with different numbers of vignettes.

First, since more categories of C (which result from using more vignettes) should never yield less discriminatory power, we require in the case of no ties or inconsistencies that H be a monotonically increasing function of the number of vignettes J and hence the number of categories, $2J + 1$. Second, when adding a new vignette and decomposing a category of C into smaller bins, the amount of expected information in the union of these smaller bins should remain the same as the original undecomposed bin and the expected information in the other unaffected bins should remain unchanged by the addition of the new vignette.

For example, suppose we begin with a single vignette ($J = 1$), so that C has $2J + 1 = 3$ categories, with proportions labeled p_1, p_2, and q, respectively, and we wish to compute $H(p_1, p_2, q)$. The three proportions refer, respectively, to outcomes where the self-assessment y is less than, equal to, and greater than z_1. Now consider adding an additional vignette with a value higher than the existing one: $z_1 < z_2$. This effectively breaks up the group of respondents with self-assessments that fall to the right of z_1 (previously the third event, with frequency q) into the self-assessment falling between z_1 and z_2, equal to z_2, and greater than z_2, and of course, the decomposition is logically consistent: $q = p_3 + p_4 + p_5$. Since the first two events, being less than z_1 and equal to z_1, have not changed with the introduction of z_2, the first two proportions, p_1 and p_2, are unchanged, and so we require our measure of the information that they convey to also be unchanged.

The second criterion, then, is that we ought to be able to compute identical values of the information in the vignettes by applying the function to the set of all five probabilities $H(p_1, p_2, p_3, p_4, p_5)$ *or* by computing a weighted average of the information in grouped categories $H(p_1, p_2, q)$ and the information in the components of the third category, $H(p_3, p_4, p_5)$. More formally, this amounts to adding a consistency requirement so that in this example

$$H(p_1, p_2, p_3, p_4, p_5) = H(p_1, p_2, q) + qH(p_3, p_4, p_5), \tag{7}$$

where the second term quantifies separately how much information the second vignette adds. Thus, we shall require a generalization of this, which states that when multiple ways of hierarchically decomposing $H(\cdot)$ exist, they shall all yield identical values.

Although numerous potential candidates for $H(\cdot)$ exist—such as any of the measures of income inequality, like the Gini index, the variance, and mean absolute deviations—the essential requirements given above rule out all but one. As proved by Shannon (1949) in the context of mathematical communications theory, the unique definition of this function is proportional to

$$H(p_1, \ldots, p_{2J+1}) = -\sum_{j=1}^{2J+1} p_j \ln(p_j), \tag{8}$$

where for convenience we define $-0 \ln(0) \equiv 0$ (since $\lim_{a \to 0^+} a \ln(1/a) = 0$). The function H is known as *entropy* and has found many applied uses (Golan, Judge, and Miller 1996).[5]

The measure of entropy is such that $H = 0$ if and only if all respondents fall in only one category of C, the situation of minimal information. For any given J, H is at a maximum and equal to $\ln(2J + 1)$, when $p_j = 1/(2J + 1)$ (for all j), which is a uniform frequency distribution. The maximum entropy (and informativeness) is thus higher as the number of vignettes J gets larger. Finally, any equalizing of the frequencies of two or more categories produces an increase in H, which implies that the measure is also logically consistent between the extremes.

The key insight, however, is that *no* other measure—not the variance of the p_j's, not mean absolute deviations, not the Gini coefficient, not anything but entropy—satisfies the essential criteria set out above.

4.2 *Information in Scalar and Vector Values of C*

When two or more vignettes are tied or inconsistently ranked by a survey respondent, C can be vector valued and so the standard definition of entropy in equation (8) cannot be applied directly. We now offer several ways forward when C includes censored values for some observations.

One simple approach is to estimate the full frequency distribution following the procedures described in Section 3.2 and then apply H to these estimates. This *estimated entropy* approach usually works well, and we recommend its use. However, we should understand that entropy computed this way is a measure of (a) the informativeness of the vignettes (b) as supplemented by the predictive information of the covariates included in the censored ordered probit and (c) assuming the probit specification is correct. This is a highly useful calculation, of course, especially since the assumptions are similar to those in the statistical models routinely used in social science data analyses. But we also need a pure non-model-based measure of only the information made available by the vignettes for certain. Indeed, the difference between these two measures, if we can create the second, would indicate how much information is contributed by the censored ordered probit estimation, conditional on its assumptions.

Thus, we now consider the general situation where we compute the information in the vignettes, without making additional assumptions. To do this, we think of the frequency distribution as partly known, due to the observations with scalar values of C. The rest of

[5]Paradoxically, entropy was developed as a measure of randomness or the *lack* of information in a communications signal and was used even earlier in physics as a measure of the disorder, or the amount of thermal energy not available to do work, in a closed system. In contrast, in our context, entropy is roughly the opposite, a measure of the *amount* of information in our survey responses. What unites the examples is that, in both cases, entropy is a measure of equality.

Table 2 Selected vignette orderings and possible values of C

Observed ranking of vignettes for respondent i	*Possible values of C_i (depending on y_i)*
$z_1 < z_2 < z_3$	1, 2, 3, 4, 5, 6, 7
$z_1 = z_2 < z_3$	1, {2, 3, 4}, 5, 6, 7
$z_1 = z_2 = z_3$	1, {2, 3, 4, 5, 6}, 7
$z_2 < z_1 = z_3$	1, {2, 3, 4, 5, 6}, {1, 2, 3, 4, 5}, {1, 2, 3, 4}, 7
$z_2 < z_1 < z_3$	1, {1, 2, 3, 4, 5}, {2, 3, 4, 5}, {1, 2, 3, 4}, 5, 6, 7

Note. Braces in the right column denote vector-valued responses for C.

the frequency distribution, from the vector-valued observations, are partially unknown, and so we shall estimate them. However, as we shall see, estimation in this context does not involve any added uncertainty or assumptions. Instead, the "estimation" process in this context involves calculating the information we are certain the vignettes and self-assessment provide.

To fix ideas, Table 2 shows what happens with five selected orderings of three vignettes. For each, it lists all possible values of C (depending on the self-assessment response). Thus, the first row is the canonical case with the vignettes uniquely ranked in the correct order. (Vignette values are intended to be ordered by the number in their subscript, and items in braces are either tied or inconsistent.) This produces only scalar values for C. The second ordering can generate four possible scalar values of C and one vector-valued response, and so on.

To compute a full frequency distribution, we follow five steps. First, sort all the scalar values of C into their appropriate bins. Second, parameterize frequency distributions of responses for each unique combination of the vector values. For example, for all the vignette responses that follow the pattern in the second row in the table and where the self-assessment leads to a vector-valued set, $y \in [z_1, z_2]$, we have $C = \{2, 3, 4\}$. In this situation, we assign unknown frequencies q_2, q_3, and q_4 to the three categories, respectively. The values of these frequencies are unknown, but we know that they sum to one: $q_2 + q_3 + q_4 = 1$ (meaning that the probability of C taking on the values 1, 5, 6, or 7 is 0), and so the number of free parameters for this example is only two. We also make similar parameterizations for the vector-valued responses in the third and fourth rows of Table 2, yielding eight free parameters for the entire problem. Third, we estimate the q values, by a procedure we describe shortly, and add the estimated q's (weighted by the number of respondents for which C takes on the same value) into the appropriate bins and into which we have already put the scalar values. At this point, we have an estimate of the full frequency distribution, $p_1, p_2, \ldots, p_7$, and so as a final step, we compute the information in the vignettes by applying the entropy formula in equation (8).

The only remaining question, then, is to estimate the unknown q parameters. We do so by minimizing equation (8). If we only use the information in the vignettes and self-assessment responses, then the minimum entropy is exactly the information in C that we *know* exists in our data. Any other information that may exist would be estimated and hence would require potentially incorrect modeling or other statistical assumptions.[6] As it

[6]Although the optimization procedure produces estimates of the q's, they are ancillary parameters and are of no particular interest in and of themselves. Because our criteria indicate that we are indifferent among all histograms with the same entropy, the only relevant quantity produced by this procedure is the value of the minimum entropy. We would be interested in differences between two densities with the same level of entropy if, for example, we had preferences for measures that provided more precision at a certain portion of the scale or if we only wished to identify some specific percentile or fraction of respondents. In these situations, alternative formal criteria would probably lead to a measure of weighted or relative entropy, such as that computed from the Kullback and Leibler (1951) distance, $\mathrm{KL} = \sum_{j=1}^{2J+1} p_j \ln(p_j/q_j)$.

happens, minimizing equation (8) is easy since the unknown parameters at the minimum always take on the value 1 for one q and 0 for all the others. At the minimum, if we moved any one individual, we would still have no reason to move any other individual.[7]

4.3 *Detecting Unidimensionality Violations*

As should be clear from Section 3.1, we do not require that each respondent give unique answers to each vignette in the set or that all respondents rank the vignettes in the same order. We only need assume that respondents understand the vignettes on a common scale apart from random perceptual, response, and sampling error. The ties and inconsistencies that result from these types of errors violate no assumptions of our methodology.

The only necessary assumption for the valid application of the nonparametric estimator C is the absence of nonrandom error that might generate ties or inconsistencies that affect y. With censored values for C, the sum of random and nonrandom error is below detectable levels, and we can safely ignore any nonrandom error problem. What we must focus on, then, is the evidence that violations of the assumption exist among observations for censored values. Since either random or nonrandom error can cause the data to produce vector values, there exists no certain method of partialing out the two and uniquely detecting the nonrandom error. Nevertheless, we offer in this section an approach that is likely to be indicative of problems when they occur while still making minimal assumptions. The method is meant to supplement, not substitute for, traditional survey methods of detailed cognitive debriefing and pretesting.

At a general level, we regard any number of values of C greater than one to be undesirable, with undesirableness increasing as the reported range increases, although we cannot tell whether this is due to random error, which violates no assumption, or nonrandom error, which does. Before we parse which values are more likely to be generated by nonrandom error, we first distinguish among different patterns that lead to the identical values of C.

For example, consider survey responses such as these:

$$z_1 < \{y, z_2, z_3\} < z_4 < z_5, \tag{9}$$

where the braces identify a group that is tied or inconsistent so that y is not strictly less than or greater than them all and the subscripts on the z's indicate the assumed vignette ordering. Since this notation does not distinguish among the various possible orderings within this set, we now consider two possible orderings within the same set:

$$z_1 < y = z_2 = z_3 < z_4 < z_5, \tag{10a}$$

$$z_1 < z_3 < y < z_2 < z_4 < z_5, \tag{10b}$$

or, to clarify, we focus only on the responses that matter for this example:

$$y = z_2 = z_3, \tag{11a}$$

$$z_3 < y < z_2. \tag{11b}$$

[7]We wondered whether it might be possible to determine analytically the minimum entropy without going through the intermediate step of estimating the frequency distribution. We thus posed this question to some experts in the field and soon received very helpful suggestions from Amos Golan, George Judge, and Doug Miller for how to do this in several interesting special cases. We have even received a working paper on the subject generated by our question Grendar and Grendar (2003), and it seems that research on the question continues. Since our needs are more general than current results, we compute the minimum value of entropy, in the presence of a vector-valued C, via a genetic algorithm optimizer, GENOUD (Sekhon and Mebane 1998).

Table 3 Counting switch distance until C is scalar valued: two examples

Step	*1*	*2*	*3*	*4*	*5*	G_0
0		y, z_2, z_3				2
1		y, z_2	$\rightarrow z_3$			1
0	z_3	y	z_2			2
1		$\rightarrow z_3, y$	z_2			2
2		y	$\rightarrow z_3, z_2$			0

Note. Arrows identify items moved and the direction they moved, from the previous step on the line above.

Both of these inequalities are inconsistent and yet clearly equation (11a), where the self-assessment and the two vignettes are tied, seems a good deal less problematic than equation (11b), where the vignettes are out of order. The problem is that there exist many possible orderings of ties and inconsistencies that lead to the same range for C.

Our immediate goal, then, is a metric with which we can order the two examples and any others that arise. We do this by adapting ideas from the field of statistical genetics. Researchers in that field often need to compare two sequences of DNA, where each item in the sequence is composed of one of four possible letters (or base pairs). Since a natural ordering between any sequences does not exist, developing a metric based on the similarity of or "distance" between two sequences is needed. Researchers thus often use the *switch distance*, which is the number of switches in individual letters that it takes to turn one sequence into another (Lin et al. 2002).

By a roughly analogous logic, we compute the distance between the expressions in equations (11a) and (11b) by the *minimum number of single-unit moves* in vignette responses until C is scalar valued. To simplify this calculation, we first define G_0 as the set of vignettes and y that excludes all vignettes strictly greater than y and another set that is strictly less than y. Then, we add or subtract 1 from a chosen vignette response on each round (so that the new "response" is still between 1 and J) and compute the minimum number of such steps required until C has a scalar value; we label this distance M. The specific steps in a sequence of length M are not necessarily unique, but the key is that we are indifferent between all sequences that give the same value of M.

Table 3 computes M for each expression, one step at a time. It shows that although $\#G_0 = 2$ for both, $M = 1$ for equation (11a) but $M = 2$ for equation (11b).

With M, we have a measure of how far any individual survey response is from having a scalar-valued C. It indicates how much error—random or nonrandom—exists in the data. Other things equal (such as entropy), choosing vignettes that reduce the average M among observations in the data would seem to be a good idea.

In addition, we suggest several strategies to use M to provide hints about the possible existence of nonrandom error. First, we suggest examining a histogram of M, since random occurrences of ties or inconsistencies that implicate y should normally lead to a unimodal histogram, with the first mode near 0, but nonrandom violations may yield additional modes. This point should be checked for "edge effects," since different patterns of M may result when y is near 0 or J. To do this, we merely examine a separate histogram of M for respondents giving each value of y.

Finally, we recommend exploring M statistically by trying to predict it with available explanatory variables. It may be that we can identify a stratum of respondents who conceptualize the underlying scale differently than others. If so, we might be able to

rewrite the questions to correct for this problem, choose vignettes differently, or test the hypothesis further with the parametric model.

5 Choosing Vignettes in Practice

We now provide empirical applications of both the *known (minimum) entropy* and the *estimated entropy* statistics. The known entropy reveals the amount of information that we know to exist for certain in the self-assessment and vignette questions, that is, without making any new assumptions, whereas the veracity of the estimated entropy statistics (calculated from the histogram resulting from the conditional probability expression applied to estimates from our censored ordered probit model) depends on plausible modeling assumptions but ones that could be wrong.

In practice, researchers may include a large number of vignettes in a pretest survey and will need to evaluate them and decide which subset to include in the more expensive regular survey. We thus now offer several examples of this process, first using the political efficacy data described in Section 2 and subsequently with vignettes measuring several components of health.

5.1 *Political Efficacy*

The political efficacy data include five vignettes—from the lowest level of efficacy, which we label "1" to the highest level, labeled "5." (1 is Moses in the example in Section 2; all five are given in King et al. 2004).

Given five vignettes on a pretest survey, we must choose among 31 possible subsets of vignettes (where, using the labels above, these include subsets 1, 2, 3, 4, 5, 12, 13, 14, . . ., 12345), excluding of course the possibility of writing new vignettes. For each of the 31 choices, we computed the minimum entropy and the estimated entropy and plot the pair in a scatterplot in Fig. 2.

Figure 2 has a variety of informative features. Consider first solutions with a single vignette, which appear in the lower left corner of the figure. Each of these appears exactly on the 45° line where known and estimated entropy are identical. They are identical because the uncertainty in estimated entropy is entirely due to ties and inconsistencies, and these are not possible with a single vignette. Substantively, we find that in these data the vignettes with lower levels of efficacy are most informative. Indeed, in the data, 40% of the Chinese respondents placed themselves below vignette 1 (Moses, suffering in silence) and so are lumped together in a single category ($C = 1$), and all remaining vignettes break up the rest of the distribution. In other words, the vignettes closest to the largest number of people are most informative, and in the present sample, these are the low-efficacy vignettes. Adding additional vignettes, below Moses suffering in silence would be a good idea methodologically, although convincing countries to allow these questions on public surveys that they must approve may be infeasible.

All subsets with greater than one vignette have higher estimated than known entropy (and thus appear above the 45° line in Fig. 2) because of the presence of ties and inconsistencies. Overall, the figure demonstrates that adding additional vignettes always results in more estimated entropy. This relationship is a feature of the definition of entropy and will always hold, although in differing degrees. In this sample, the bonus in estimated entropy of going from one to two vignettes is a good deal larger (for all pairs other than 45) than moving from two to three, three to four, or four to five. Indeed, there appears to be a clear decreasing return to scale, as each additional vignette adds somewhat less than the

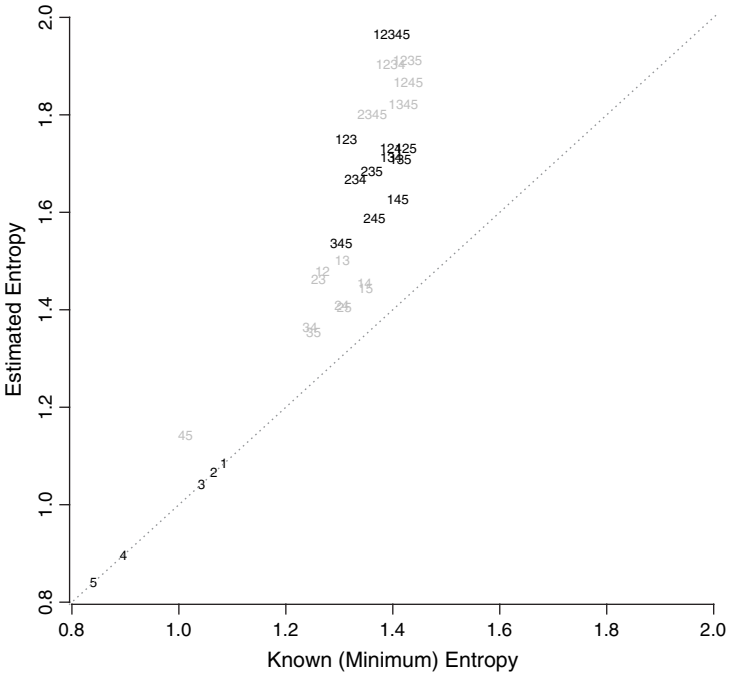

Fig. 2 Estimated by known (minimum) entropy for political efficacy. The vertical axis is the entropy calculated from the histogram estimated by our censored ordered probit. The horizontal axis is the amount of entropy known to exist without making any assumptions. The list of vignette numbers is plotted, with combinations of two and four vignettes appearing in gray to distinguish them, for clarity.

previous one. How this fact translates into the choice of vignettes depends on the survey costs of additional vignettes and the value of the additional information.

Although adding vignettes always increases estimated entropy, it does not always help with known (minimum) entropy and indeed can reduce the level. The intuition for this nonmonotonicity is that by adding a vignette and inducing ties, a distribution which was relatively flat previously now has the possibility, because of the tied cases, of becoming more spiked and hence less informative. As such, although adding vignettes that induce no ties is uniformly better—assuming as we do throughout this section that they satisfy response consistency and vignette equivalence—in practice, we have a trade-off since adding vignettes not only enables us to compute C with additional precision but also may induce more ties and inconsistencies. The great advantage of this measure of known entropy, then, is that it provides information with which to decide whether the benefits of adding a new vignette outweigh the costs, all on a single quantitative scale. This approach enables us to focus attention on a relatively small number of the most important comparisons by ruling out combinations of vignettes that are strictly dominated on both dimensions (and so on the graph fall below and to the left) by another subset.

Thus, we see by the position of the points relative to the horizontal axis in the figure that vignette combination 45 has a much lower level of entropy on either scale than any other pair of vignettes. If one preferred to make no statistical modeling assumptions, a good trade-off between known information in C gained and the cost of vignettes may be subset 125. Subset 1245 has slightly more known information (as it appears slightly to the right of 125), but the cost of additional items on most surveys would probably make the addition of

vignette 4 not worthwhile, unless one were willing to make statistical modeling assumptions and thus focus on estimated entropy and the vertical dimension of the figure. If we could afford to include only four vignettes, we would not use 2345, 1345, or 1245 since they are dominated by 1235, 1234, or both. Out of the 10 combinations of three vignettes, 123 or 125 would appear to be likely choices.

Whether to add vignettes 3 and 4 to the subset 125, then, depends on one's trust in the ordered probit modeling assumptions. These can be judged only in part by checking the fit of the ordered probit model to the untied observations. In addition, one should keep in mind that the amount gained by going from subset 125 to 12345 (vertically on the graph) is only about half the entropy gained by going from one vignette to three. So if one can afford three vignettes, 125 would appear to be a good choice according to these criteria. Whether to include more depends on how comfortable one is with the assumptions.

Finally, we note that the application we have been referring to here considers only the histogram of the entire sample. If the goal of the research is to compute the density of political efficacy in each country separately or of some causal effect, one should evaluate these quantities directly, in a fashion directly analogous to the analysis we performed here.

5.2 *Health*

In order to illustrate different features of entropy results, we now conduct the same analysis for two different variables. Both are components of health with self-assessments and vignettes asked as part of the 2002 World Health Survey (conducted by the World Health Organization) in a different survey in China.

We begin with an analysis of vignettes about the quality of sleep and restfulness during the day. For example, one of the five vignettes asked is

> [Noemi] falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night.

(the others can be found at the anchoring vignettes Web site), with survey question

> In the last 30 days, how much of a problem did you have [does "name" have] due to not feeling rested and refreshed during the day? (1) None, (2) Mild, (3) Moderate, (4) Severe, (5) Extreme/ Cannot Do.

Figure 3 analyzes these data with an entropy graph in a form directly parallel to that in Fig. 2.[8] A particularly interesting feature of this graph is that it gives several examples of a single strategically placed vignette providing more information than a set of *three* apparently less-well-placed vignettes. In particular, either vignette 1 or 2 alone provides more discriminatory power than vignettes 3, 4, and 5 do taken together. Indeed, this is true whether we measure discriminatory power by either estimated or known entropy. Similarly, a combination of two vignettes (1 and 2) provide as much or more information than two sets with four vignettes (1345 and 2345) and many with three vignettes.

Finally, we offer an analysis of the self-care, with a representative vignette:

> [Victor] usually requires no assistance with cleanliness, dressing and eating. He occasionally suffers from back pain and when this happens he needs help with bathing and dressing.

[8]The estimated entropy for the sleep questions are calculated from the conditional fitted probabilities of a censored probit model with covariates sex, age, weight, years of schooling, and marital status. The same holds for self-care ability, subsequently, except that height is excluded from the set of covariates since it's inclusion leads to an unidentified cutpoint. No country dummy is included because figures are for China only.

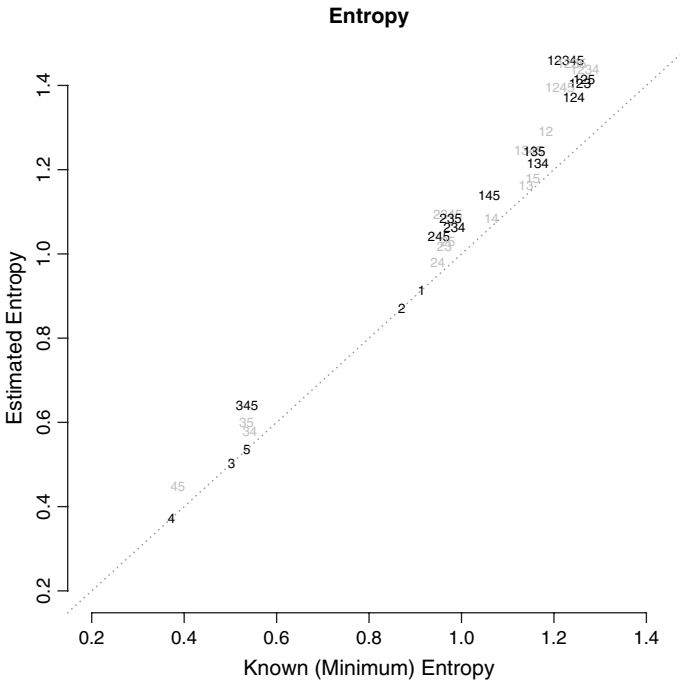

Fig. 3 Estimated by known (minimum) entropy for quality of sleep. The vertical axis is the entropy calculated from the histogram estimated by our censored ordered probit. The horizontal axis is the amount of entropy known to exist without making any assumptions. The list of vignette numbers is plotted, with combinations of one, three, and five vignettes appearing in gray for clarity.

and the survey question:

> Overall in the last 30 days, how much difficulty did [name of person/you] have with self-care, such as washing or dressing [yourself/himself/herself]? (1) None, (2) Mild, (3) Moderate, (4) Severe Extreme/Cannot Do.

Figure 4 provides the entropy analysis, using the same covariates as the previous example. As can be seen by either measure, these vignettes provide little discriminatory power. A key reason entropy is so low in these data is that the vast majority of people surveyed feel they have little trouble with self-care and so the vignettes are far from their self-assessments and hence are relatively uninformative.

However, the low levels of entropy produced by these vignettes might instead be due to the poor quality of the vignettes themselves. For example, the vignette above describes self-care with respect to "cleanliness," "dressing," "eating," and "bathing." The survey question omits all but one of these (dressing) and adds another not directly referenced in the vignette (washing). In addition, the vignette gives a reason for the level of self-care (namely, back pain), whereas the reason for any "difficulty" mentioned in the survey question is not given. It is even possible that some respondents may have interpreted "difficulty . . . with self-care, such as washing or dressing" as being unrelated to the health causes and instead referring to personal hygiene and style.

6 Concluding Remarks

The methods offered here are intended to evaluate and improve the information revealed by surveys using anchoring vignettes. The diverse types of data and survey questions we

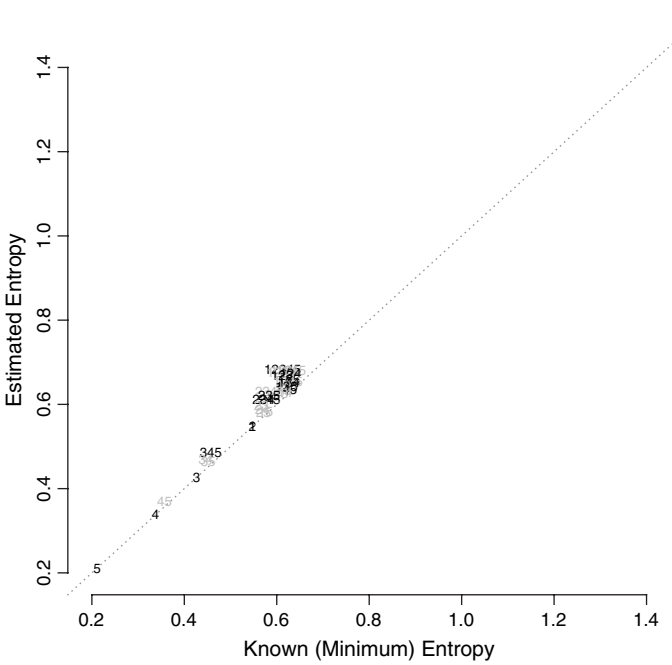

Fig. 4 Estimated by known (minimum) entropy for self-care ability. The vertical axis is the entropy calculated from the histogram estimated by our censored ordered probit. The horizontal axis is the amount of entropy known to exist without making any assumptions. The list of vignette numbers is plotted, with combinations of one, three, and five vignettes appearing in gray for clarity.

have thus far analyzed seem to indicate that these methods can reveal a considerable amount of information not otherwise available, and so their use would appear to be recommended.

Researchers should also keep in mind that anchoring vignettes were designed to correct for response-category DIF, not for all forms of DIF. If respondents understand the stem question in fundamentally different ways, then anchoring vignettes and the methods described here may not fix all inferential problems. Obviously, if we inadvertently asked half of our respondents the wrong survey question, we would get nonsense results. It should be no less obvious that if we ask a single question of everyone but half of the respondents interpret the question in a massively different way, methodological fixes of almost any kind would be unlikely to fix the problem.

In our view, most of those who conduct surveys and use survey data in their research need to start correcting for, or evaluating the presence of, DIF in response categories using anchoring vignettes and the methods described herein. Without such an evaluation, many survey results could be called into question. For survey researchers and methodologists, the remaining question is how also to address other aspects of interpersonal incomparability aside from that due to response-category DIF.

References

Golan, Amos, George Judge, and Doug Miller. 1996. *Maximum entropy econometrics: Robust estimation with limited data.* London: John Wiley and Sons.

Grendar, M., Jr., and M. Grendar. 2003. Maximum probability/entropy translating of contiguous categorical observations into frequencies. Working paper, Institute of Mathematics and Computer Science, Mathematical Institute of Slovak Academy of Sciences, Banska Bystrica.

Heitjan, Daniel F., and Donald Rubin. 1990. Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* 85:304–14.

King, Gary. 1997. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton, NJ: Princeton University Press.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95 (March): 49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

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 (February): 191–205. http://gking.harvard.edu/files/abs/vign-abs.shtml.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science* 44 (April): 341–55. http://gking.harvard.edu/files/abs/making-abs.shtml.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (March): 79–86.

Lin, Shin, David L. Cutler, Michael E. Zwick, and Aravinda Chakravarti. 2002. Haplotype inference in random population samples. *American Journal of Human Genetics* 71:1129–37.

Schwarz, Norbert. 1999. Self-reports: How the questions shape the answers. *American Psychologist* 54:93–105.

Sekhon, Jasjeet Singh, and Walter R. Mebane, Jr. 1998. Genetic optimization using derivatives: Theory and application to nonlinear model. *Political Analysis* 7:187–210.

Shannon, Claude E. 1949. *The mathematical theory of communication.* Urbana-Champaign, IL: University of Illinois Press.

Wand, Jonathan, Gary King, and Olivia Lau. Forthcoming. Anchors: Software for anchoring vignettes data. *Journal of Statistical Software*.