Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research

Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research

GARY KING *Harvard University*
CHRISTOPHER J. L. MURRAY *World Health Organization*
JOSHUA A. SALOMON *Harvard University*
AJAY TANDON *World Health Organization*

The version of this article published in the November 2003 issue contained printing errors. This corrected version replaces it.

We address two long-standing survey research problems: measuring complicated concepts, such as political freedom and efficacy, that researchers define best with reference to examples; and what to do when respondents interpret identical questions in different ways. Scholars have long addressed these problems with approaches to reduce incomparability, such as writing more concrete questions—with uneven success. Our alternative is to measure directly response category incomparability and to correct for it. We measure incomparability via respondents' assessments, on the same scale as the self-assessments to be corrected, of hypothetical individuals described in short vignettes. Because the actual (but not necessarily reported) levels of the vignettes are invariant over respondents, variability in vignette answers reveals incomparability. Our corrections require either simple recodes or a statistical model designed to save survey administration costs. With analysis, simulations, and cross-national surveys, we show how response incomparability can drastically mislead survey researchers and how our approach can alleviate this problem.

The discipline of political science is built on *theory*, including a rough agreement on normative theories preferring freedom, democracy, and political equality, among others, and the development of positive theories focused on understanding the causes and consequences of these variables. Empirical political science, in turn, is devoted in large part to making causal inferences about these same variables. Undergirding this superstructure of theory and causality is *measurement*, the detailed mapping of the levels of these basic variables. Although it may not seem as exciting as causal inquiry, better measurement obviously has the potential to affect our understanding of the extent of any problems that may need addressing and the estimates of any causal effects. Indeed, achieving the theoretical and causal goals of our field and all other empirical fields "would seem to be virtually impossible unless its variables can be measured adequately" (Torgerson, 1958).

We address two long-standing problems with measurement using sample surveys (a data collection device used in about a quarter of all articles and about half of all quantitative articles published in major political science journals [King et al. 2001, fn 1]). The first is how to measure concepts researchers know how to define most clearly only with reference to examples—freedom, political efficacy, pornography, health, etc. The advice methodologists usually give when hearing "you know it when you see it" is to find a better, more precise theory and then measurement will be straightforward. This is the right advice, but it leads to a well-known problem in that highly concrete questions about big concepts like these often produce more reliable measurements but not more valid ones.

The second problem we address occurs because "individuals understand the 'same' question in vastly different ways" (Brady 1985). For example, Sen (2002) writes that

> the state of Kerala has the highest levels of literacy . . . and longevity . . . in India. But it also has, by a very wide margin, the highest rate of reported morbidity among all Indian

You may be interested in our Anchoring Vignettes Web site, which, as a companion to this paper, provides software to implement the methods here, answers to frequently asked questions, example vignettes, and other materials (see http://gking.harvard.edu/vign/). Our names on this paper are ordered alphabetically. Our thanks go to John Aldrich, Jim Alt, Larry Bartels, Neal Beck, David Cutler, Federico Girosi, Dan Ho, Kosuke Imai, Stanley Feldman, Michael Herron, Mel Hinich, Simon Jackman, Orit Kedar, Jeff Lewis, Jeffrey Liu, John Londregan, Joe Newhouse, Keith Poole, Sid Verba, Jonathan Wand, and Chris Winship for helpful discussions; Ken Benoit, Debbie Javeline, and Karen Ferree for help in writing vignettes; three anonymous referees and the editor for exceptionally helpful suggestions (One of our reviewers, who we now know is Henry Brady, wrote an extraordinary 20 page single-spaced review that greatly improved our work.); and NIA/NIH (Grant P01 AG17625-01), NSF (Grants SES-0112072 and IIS-9874747), WHO, the Center for Basic Research in the Social Sciences, and the Weatherhead Center for International Affairs for research support.

Gary King is David Florence Professor of Government, Harvard University, Center for Basic Research in the Social Sciences, Cambridge MA 02138. (http://GKing.Harvard.Edu, King@Harvard.Edu). Christopher J. L. Murray is Executive Director, Evidence and Information for Policy, World Health Organization, Geneva, Switzerland (MurrayC@WHO.int). Joshua A. Salomon is Assistant Professor of International Health, Harvard School of Public Health, Center for Population and Development Studies, Cambridge, MA (JSalomon@hsph.harvard.edu). Ajay Tandon is Health Economist, Evidence and Information for Policy, World Health Organization, Geneva, Switzerland (TandonA@WHO.int).

> states. . . . At the other extreme, states with low longevity, with woeful medical and educational facilities, such as Bihar, have the lowest rates of reported morbidity in India. Indeed, the lowness of reported morbidity runs almost fully in the opposite direction to life expectancy, in interstate comparisons. . . . In disease by disease comparison, while Kerala has much higher reported morbidity rates than the rest of India, the United States has even higher rates for the same illnesses. If we insist on relying on self-reported morbidity as the measure, we would have to conclude that the United States is the least healthy in this comparison, followed by Kerala, with ill provided Bihar enjoying the highest level of health. In other words, the most common measure of the health of populations is negatively correlated with actual health.

Studying why individuals have perceptions like these, so far out of line with empirical reality, "deserves attention" but measuring reality only by asking for respondents' perceptions in these situations can be "extremely misleading" (Sen 2002).

The literature on this problem has focused on developing ways of writing more concrete, objective, and standardized survey questions and developing methods to reduce incomparability. Despite a half-century of efforts, however, many important survey instruments are still not fully comparable (Suchman and Jordan 1990). Indeed, even though political scientists have been aware of the devastating consequences of ignoring the problem for almost two decades (Brady 1985), the lack of tools to deal with it has meant that the comparability of most of our survey questions has not even been studied.

We have designed a new approach to survey instrumentation that seems to partially ameliorate both problems. Our key idea, in addition to following the venerable tradition of trying to write clearer questions that are more comparable, is a method of directly measuring the incomparability of responses to survey questions, and then correcting for it. We ask respondents for self-assessments of the concept being measured along with assessments, on the same scale, of each of several hypothetical individuals described by short vignettes. We create interpersonally comparable measurements by using answers to the vignette assessments, which have actual (but not reported) levels of the variables that are the same for every respondent, to adjust the self-assessments. Our adjustments can be made with simple calculations (straightforward recode statements) or with a more sophisticated statistical model that has the advantage of lowering data collection costs. Easy-to-use software to implement our statistical methods, a library of examples of survey questions using our approach, and other related materials can be found at `http://GKing.Harvard.edu/vign/`.

PREVIOUS APPROACHES

The most widely used modern terminology for interpersonal incomparability is *differential item functioning* (DIF), which originated in the educational testing literature.[1] The search for methods of detecting or conquering DIF usually centers on the identification of common *anchors* that can be used to attach the answers of different individuals to the same standard scale.

The earliest and still the most common anchors involve giving the endpoints of the (or all) survey response categories concrete labels—"strongly disagree," "hawk," etc. This undoubtedly helps, but is often insufficient. An early and still used alternative is the "self-anchoring scale," where researchers ask respondents to identify the top- and bottommost extreme examples they can think of (e.g., the name of the respondent's most liberal friend and most conservative friend) and then to place themselves on the scale with endpoints defined by their own self-defined anchors (Cantril 1965). This approach is still used but, depending as it does on extremal statistics, it often lowers reliability, and it will not eliminate DIF if respondents possess different levels of knowledge about examples at the extreme values of the variable in question.

Researchers sometimes compare a survey response at issue to "designated anchors," which are questions that tap the same concept that experts believe have no DIF (Przeworski and Teune 1966–67; Thissen, Steinberg, and Wainer 1993). This is an important approach, but as the authors recognize, it begs the question of where knowledge of the anchors come from in the first place. Sometimes researchers evaluate each survey question in turn by comparing it with an average, or factor analyzed weighted average, of all the others that measure the same concept. As is also widely recognized, however, the assumption that all the other questions do not have DIF on average, as each question moves in and out of the "gold standard" comparison group, is internally inconsistent.

Although not widely known outside our field, the most satisfactory approaches to correcting for DIF in any field have been in the context of application-specific models built by political scientists. The first such model was Aldrich and McKelvey (1977), which estimated the positions of candidates and voters in a common issue space. The actual positions of candidates were assumed the same for all respondents and, so, could be used as anchors to adjust both candidate and voter issue positions. Since these actual positions are unobserved, Aldrich and McKelvey assume that voters have unbiased perceptions of candidate positions but that the reported positions are linearly distorted in an unknown, but estimable, way. Because of the constrained computational resources available at the time, they recognized

[1] In the educational testing literature, a test question is said to have DIF if equally able individuals have unequal probabilities of answering the question correctly. The analysis of reasons for the varying test performance of students in different racial groups has provided considerable impetus for the study of DIF. Indeed, the term DIF was chosen to replace the older "item bias" term as an effort to sidestep some of the politically charged issues involved (see Holland and Wainer 1993 for a review of the literature). Paradoxically, the method we introduce here would seem applicable to all fields where DIF is an issue except for educational testing.

but did not model several other features of the problem, such as the ordinal nature of the response categories.[2]

Using a similar logic, Groseclose, Levitt, and Snyder (1999) adjust interest group scores across time and houses of congress by using scores on the same legislator at different times (when serving in the same or different chambers) as anchors. Their model thus assumes that members have constant expected, but not measured, interest group scores. Poole and Rosenthal's (1991) widely used D-Nominate scores for scaling legislators and roll calls applies analogous ideas for anchors (see also Heckman and Snyder 1997 and Poole and Daniels 1985). Londregan (2000) uses similar anchoring in a model more amenable to small samples and resolves several identification problems by simultaneously modeling the agenda, while Clinton, Jackman, and Rivers (2002) present a fully Bayesian approach. Baum (1988) adjusts the scaling of the liberalness of Supreme Court decisions by assuming the stability of individual justices over time, and anchoring the court decisions to justices that serve in more than one "natural" court. See also Lewis 2001 for a similar approach to scaling voting behavior and for his review of other work in this area.

The anchors used in most political science applications are far better than the unadjusted values (and better than most anchors available in other fields), but as is fully recognized by the authors, the strategies employed by political actors mean that the anchors are not completely free of DIF. For example, a reasonable characterization of much of the partisan process of writing legislation is to create DIF—to make the choice harder for opposition legislators than members of one's own party. Similarly, if candidates succeed in being even in part "all things to all people," the use of voter perceptions of candidate positions as anchors could be biased.

Most current efforts at dealing with DIF in other fields try to identify questions with DIF and delete them or collapse categories to avoid the problem (Holland and Wainer 1993). Some model DIF in unidimensional scales as additional unobserved dimensions (Caroll and Chang 1970; Shealy and Stout 1993). Others use Rasch models, a special case of item response theory, which come with a variety of statistical tests and graphical diagnostics (see Piquero and Macintosh 2002). The multidimensional scaling literature has also paid considerable attention to DIF, which they call "interpersonal incomparability" (Brady 1989) or "individual differences scaling" (Alt, Sarlvik, and Crewe 1976; Clarkson 2000; Mead 1992). Others parse DIF into components like "acquiescence response set," the differential propensity of respondents to agree with any question, no matter how posed; "extreme response set," the differential propensity of respondents to use extreme choices offered, independent of the question; and many others (Cheung and Rensvold 2000; Johnson 1998; Stewart and Napoles-Springer 2000). DIF potentially affects most survey-based research throughout political science and in a wide variety of other fields.

[2] Palfrey and Poole (1987) show that the Aldrich and McKelvey procedure recovers candidate locations well, even if errors (contrary to the model) are heteroskedastic over candidates, but voter positions are biased toward the mean, especially for poorly informed voters. Poole (1998) generalizes Aldrich and McKelvey 1977 to multiple dimensions and to handle missing data.

SURVEY INSTRUMENTATION: ANCHORING VIGNETTES

The usual procedure for measuring sophisticated concepts with surveys is to gather a large number of examples and design a concrete question that covers as many of the examples as possible. Our idea is, in addition to this approach, to use the examples themselves in survey questions to estimate each person's unique DIF, and to correct for it. Examples presented in vignettes to respondents have a long history of use for other purposes in survey research (e.g., Kahneman, Schkade, and Sunstein 1998; Martin, Campanelli, and Fay 1991; Rossi and Nock 1983). We use an adapted version of vignettes that generalize the ideas in application-specific DIF-related research in political science.

We ask survey respondents in almost the same language for a self-assessment and for an assessment of several (usually five to seven) hypothetical persons described by written vignettes. For example, the anchoring vignettes for one particular domain of political efficacy might be as follows.

1. "[Alison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative."
2. "[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."
3. "[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win."
4. "[Toshiro] lacks clean drinking water. There is a group of local leaders who could do something about the problem, but they have said that industrial development is the most important policy right now instead of clean water."
5. "[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."

(We view these vignettes as falling on an ordered scale, from most to least efficacy; our empirical analyses, below, support this interpretation.) The following often-used question is then read to the respondent for each vignette and for a self-assessment:

> How much say [does 'name'/do you] have in getting the government to address issues that interest [him/her/you]?

For the self-assessment and each of the vignette question, respondents are given the same set of ordinal response categories, for example, "(1) No say at all, (2) Little say, (3) Some say, (4) A lot of say, (5) Unlimited say." Answers to this self-assessment question are normally referred to as "political efficacy," and we use this shorthand too. But what we are measuring in fact is no more or less than the concept defined by the vignette definitions, which is at best only one specific dimension of political efficacy. Other dimensions could be tapped with separate sets of vignettes.

We recommend asking the self-assessment first, followed by the vignettes randomly ordered. We also often randomly shuffle vignettes from two domains together. When feasible, we change the names on each vignette to match match each respondent's culture and sex.

MEASUREMENT ASSUMPTIONS

Our approach requires two key measurement assumptions. First, *response consistency* is the assumption that each individual uses the response categories for a particular survey question in the same way when providing a self-assessment as when assessing each of the hypothetical people in the vignettes. Respondents may have DIF in their use of survey response categories for both a self-assessment and the corresponding vignettes, but the type of DIF must be approximately the same across the two types of questions for each respondent. In other words, the type of DIF may vary across respondents, and also for a single respondent across survey questions (each with its own self-assessment and corresponding set of vignettes), but not within the self-assessment and vignette questions answered by any one respondent about a single survey question. This assumption would be violated if respondents who feel inferior to hypothetical individuals set a higher threshold for what counts as their having "a lot of say" in government than they set for the people described in the vignettes.

Second, *vignette equivalence* is the assumption that the level of the variable represented in any one vignette is perceived by all respondents in the same way and on the same unidimensional scale, apart from random measurement error. In other words, respondents may differ with each other in how they perceive the level of the variable portrayed in each vignette, but any differences must be random and hence independent of the characteristic being measured. (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.) This assumption would be violated if one set of respondents saw the vignettes above as referring to say in government through elections, as we intended, and the other interpreted our choice of words in one vignette to be referring to say in government through one's personal connections.

Thus, although we allow and ultimately correct for DIF in using survey response categories, assuming unidimensionality means that we assume the absence of DIF in the "stem question." It seems reasonable to focus on response-category DIF alone because the vignettes describe objective behaviors, for which traditional survey design advice to avoid DIF (such as writing items concretely and using pretesting and cognitive debriefing, etc.) is likely to work reasonably well. In contrast, response categories describe subjective feelings and attitudes, and so should be harder to lay out in concrete ways and avoid DIF without our methods. Whether our response-category DIF correction is sufficient is of course an empirical question. Future researchers may wish to try to generalize our methods to deal with both types of incomparability.

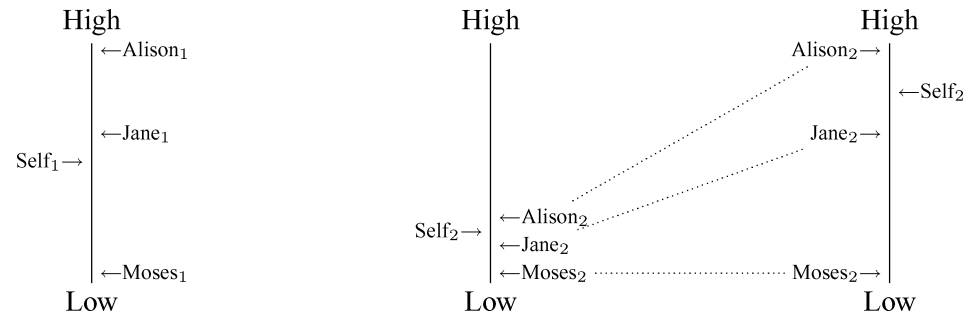
Even more basic than vignette equivalence, but implied by it, is the assumption that the variable being measured actually exists and has some logically coherent and consistent meaning in different cultures. For variables and cultures where the extreme version of the area studies critique is correct, so that different regions are truly unique and variables take on completely different meanings, then any procedure, including this one, will fail to produce comparable measures.

How do response consistency and vignette equivalence help correct for DIF? The problem with self-assessment questions is that answers to them differ across respondents according to *both* the actual level and DIF (along with random measurement error). In contrast, answers to the vignettes differ across respondents *only* because of DIF (and random measurement error). Since the actual level of political efficacy of the people described in the vignettes is the same for all respondents, we are able to use variation in answers to the vignettes to estimate DIF directly. We then "subtract off" this estimated DIF from the self-assessment question to produce our desired DIF-free (or DIF-reduced) measure.

The key goal of survey design under this approach, then, is not to design DIF-free vignette questions, which would be as difficult as for self-assessment questions, but rather to achieve response consistency and vignette equivalence. Thus, vignettes should be written to describe, in clear and concrete language, only the actual level of political efficacy of the person described, with all other language in the vignette geared to encourage respondents to think the person described is someone just like themselves in all other ways. In that way, the respondent would find it easier to use the response categories in the same way for the vignette as for the self-assessment.

The methods described below include some tests of aspects of these assumptions, but for the most part they require iterating among concept definition, question development, pretesting, and cognitive debriefing. Unlike purely observational research, the veracity of the assumptions here is under the active control of the investigator in designing the research—as in political science laboratory (Kinder and Palfrey 1993), field (Green and Gerber 2001), and survey experiments (Sniderman and Grob 1996)—but of course having control does not guarantee its proper use.

FIGURE 1. Comparing Preferences



Note: Respondent 1, on the left, reported a higher self-assessment of political efficacy than respondent 2, in the middle. On the right, Respondent 2's reported scale is deformed into one comparable to 1's scale: Now 2's vignette assessments match those for Respondent 1, revealing that Respondent 2 has a higher actual level of political efficacy than Respondent 1.

A SIMPLE (NONPARAMETRIC) APPROACH

We now combine our survey instrumentation and measurement assumptions to show how to correct DIF without sophisticated statistical techniques. The simplicity of this approach is also helpful in illustrating the key concepts and in clarifying the source of the new information.

This method can easily be used, and we use it below, but it also has two important disadvantages: First, it requires the vignette questions be asked of all the same respondents as the self-assessments, and so it can be expensive to administer. Second, as with many nonparametric methods it is statistically inefficient in some circumstances, which means that by foregoing assumptions some information is wasted. Our parametric approach, described in the section that follows, avoids these problems. However, since the nonparametric approach makes none of the parametric models' statistical assumptions and requires no explanatory variables, it makes possible several diagnostic tests of the parametric model's assumptions.

Figure 1 portrays one self-assessment and three vignette assessments for each of two individual survey respondents (labeled 1, on the left, and 2, in the middle). The self-assessed level of political efficacy is higher for Respondent 1 (and they agree on the ordinal ranking of the vignettes). However, the fact that Alison's (or Jane's or Moses's) *actual* level of political efficacy is the same no matter which respondent is being asked about her makes it possible to make the two comparable by stretching Respondent 2's scale so that the vignette assessments for the two respondents match. We do this in the scale on the right in Figure 1. With this adjustment, we can see that in fact Respondent 2 has a higher level of actual political efficacy than Respondent 1. This comes from the fact that Respondent 1's rates herself lower than Jane, whereas Respondent 2 rates herself higher than Jane.

Analyzing anchoring vignettes data by literally marking and stretching rubber bands to match Figure 1 would work fine, but we also offer an even simpler method. The idea is to recode the categorical self-assessment *relative* to the set of vignettes. Suppose that all respondents order the vignettes in the same way. Then for the vignettes in Figure 1, assign the recoded variable 1 if the self-assessment is below Moses, 2 if equal to Moses, 3 if between Moses and Jane, 4 if equal to Jane, 5 if between Jane and Alison, 6 if equal to Alison, and 7 if above Alison. (By this coding, the first respondent in Figure 1 is coded 3 and the second is coded 5.) The resulting variable is DIF-free, has easily interpretable units, and can be analyzed like any other ordinal variable (e.g., with histograms, contingency tables, or ordered probit). This method assumes response consistency and vignette equivalence, but no additional assumptions or models are required.

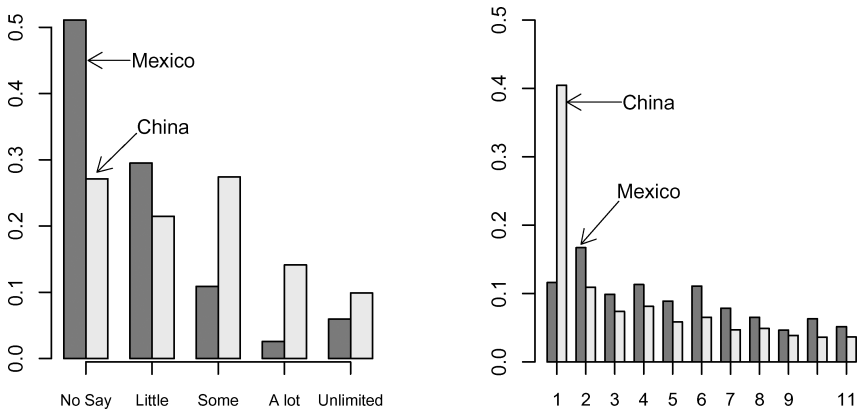
To define this idea more generally, let y_i be the categorical survey self-assessment for respondent i ($i = 1, \ldots, n$) and z_{ij} be the categorical survey response for respondent i on vignette j ($j = 1, \ldots, J$). Then for respondents with identical ordinal rankings on all vignettes ($z_{i,j-1} < z_{ij}$, for all i, j), the DIF-corrected variable is

$$C_i = \begin{cases} 1 & \text{if} \quad y_i < z_{i1}, \\ 2 & \text{if} \quad y_i = z_{i1}, \\ 3 & \text{if} \quad z_{i1} < y_{is} < z_{i2}, \\ \vdots & \vdots \\ 2J+1 & \text{if} \quad y_i > z_{iJ}. \end{cases}$$

Respondents with ties in the vignette answers would reduce our knowledge of C_i to a set of values rather than just one value. Inconsistencies in the ordinal ranking are grouped and treated as ties. When few survey response categories exist with which to distinguish among the categories of C, additional collapsing may occur. The inefficiencies in this method come precisely from the information lost due to these ties and ranking inconsistencies. (In contrast, our parametric method, described below, recognizes that some of these will be due to the random error always present in survey responses, and so it can extract more information from the data.)

To study this method, we included the questions on the electoral dimension of political efficacy described above on a sample survey of two provinces in China (with $n = 430$ respondents) and three in Mexico

FIGURE 2. Nonparametric Estimates of an Electoral Dimension of Political Efficacy



Note: The left graph is a histogram of the observed categorical self-assessments. The right graph is a histogram of C, our nonparametric DIF-corrected estimate of the same distribution.

($n = 551$). The surveys were completed in June of 2002 for the World Health Organization. Since these surveys were designed as pretests for subsequent nationally representative samples, each province surveyed was chosen to be roughly representative of the entire country. In our experience, pretests such as these usually turn out similar to the results from our subsequent nationwide surveys, but obviously this analysis should only be considered a comparison of the provinces or people surveyed.

Despite the absence of a gold standard measurement, the difference between these countries on political efficacy could hardly be more stark. The citizens of Mexico recently voted out of office the ruling PRI party in an election closely observed by the international community and widely declared to be free and fair. After a peaceful transition of power, the former opposition party took control of (and still controls) the reins of power. Despite the existence of limited forms of local democracy, nothing resembling this has occurred in China. Levels of political efficacy presumably also vary a good deal within each country, with, for example, political elites in China having high levels and prisoners in Mexico having low levels, but the average differences would seem to be unambiguous.

If we did not know these facts, and instead used standard survey research techniques, we would have been seriously misled. The left graph in Figure 2 plots histograms of the observed self-assessment responses, and quite remarkably, it shows that the Mexicans think that they have less say in government than the Chinese think that they have. The right graph plots C, our nonparametric DIF-corrected estimate of the same distribution. 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 fifth ("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.

Thus, the vignettes take the same logical place as the candidate position questions in Aldrich and McKelvey 1977, except that vignette questions are under control of the investigator and applicable to a wider range of substantive problems. In addition to political efficacy, we have written survey questions with corresponding vignettes for political freedom, responsiveness of the political system in some areas of policy, and separate domains of health (mobility, vision, etc.). We have tested subsets of these questions and our method in surveys we designed in 60 countries. The full battery of questions is now being used in the World Health Survey, which is presently in the field in about 80 countries. Other similar efforts are being used or considered by other survey organizations in several disciplines. We hope this paper will make it possible to apply the idea in other contexts.[3]

A PARAMETRIC APPROACH: MODELING THRESHOLDS

As a complement to our nonparametric approach, we now develop a parametric statistical model. This model enables researchers to save resources by asking vignettes of only a random sample (or subsample) from the same population as the self-assessments. For example, researchers could include the vignettes only on the pretest survey; alternatively, for each self-assessment on the main survey they could add, say, one additional item composed of four vignettes asked of one-quarter

[3] We have also tried a series of other ways of using these vignettes that we hoped would be even simpler, such as asking respondents to choose the vignette closest to their own level on the variable in question, but (in part because of the difficulty respondents have remembering and assessing all the vignettes at once) we have found no direct measurement alternatives that do as well as the approach we describe here.

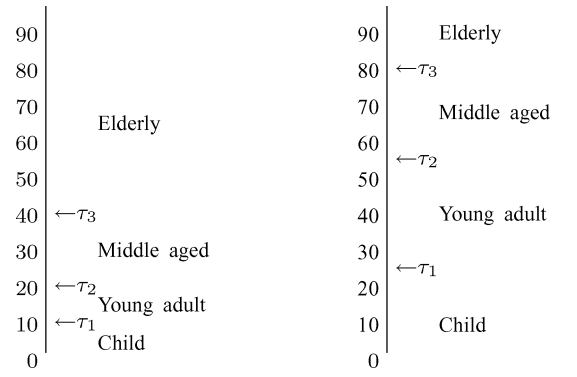
of the respondents each. For panel studies or those with a series of independent cross sections, researchers could include the vignettes on only some of the waves. This model avoids the inefficiencies of the nonparametric approach by recognizing that the variable being measured is perceived with random measurement error and, as we show below, is modeled with a normal error term. We further increase efficiency by allowing researchers to include multiple self-assessment questions for the same underlying concept (in a single factor analysis-type setup). We accomplish all these tasks by letting the thresholds (which turn the unobserved continuous variable to be measured into an observed categorical response) vary over individuals as a function of measured explanatory variables.

In broad outline, our model can be thought of as a generalization of the commonly used ordered probit model, where we model DIF via threshold variation, with the vignettes providing the key identifying information.[4] Given the importance of thresholds in this model-based method, we first illustrate their role with an alternative simplified view of DIF using a variable measured in almost every survey, *age.* Age also has an expository advantage since its perceived value is typically indistinguishable from the actual age. Then, instead of asking survey respondents for their date of birth (which obviously would be preferable), we imagine trying to make inferences if the survey question only asked whether respondents described themselves as (A) elderly, (B) middle-aged, (C) a young adult, or (D) a child.

Figure 3 considers interpretations two individuals might use to map their years of age into the available survey response categories. The age scale is broken at the threshold values τ_1, τ_2, and τ_3, but the two individuals have different values of these quantities. The scale on the left with lower threshold values (and hence, e.g., "elderly" defined as over 40 years of age) is what individuals might use in a country with a low life expectancy; the scale on the right is probably a better description of a developed country like the United States. If we knew only the response category chosen, we would not know much about that person's actual age since, for example, "middle-aged" could mean completely different things to different people. Without knowing the threshold differences, we could easily get the age rankings of the countries wrong.

If we somehow knew the threshold values, the only issue in understanding a persons' age would be grouping error, which is straightforward to deal with statistically (i.e., using an "interval regression model," which is an ordered probit with known thresholds). The key to our approach, then, is that the vignettes enable us to estimate the threshold values, and with this information we correct the self-assessment responses.

FIGURE 3. Categorizing Years of Age



Note: The graph portrays possible mappings from actual age to an individual's choice among the four survey response categories, possibly for individuals in low (on the left) and high (on the right) life expectancy countries. The τ's are thresholds between the categories.

Our model contains, for each respondent and survey question (continuous and unobserved), *actual* and *perceived* and (*ordered categorical* and *observed*) *reported* levels of the variable being measured. Respondents perceive their actual levels correctly on average but with noise (i.e., equal to the actual levels plus random measurement error), but when they turn their perceived values into an answer to the survey question, different types of people use systematically different threshold values. Hence, actual values are unobserved but comparable. Perceived values are comparable only on average due to random error, and are in any event unobserved. Raw survey responses are observed, but they are incomparable. The following two parts of this section define the self-assessment and vignette components of the model, respectively; the third part then provides a substantive interpretation of the model (Appendix A derives the likelihood function, and Appendix B shows how to compute quantities of interest from it). To help keep track of our notation, Figure 4 provides a graphic summary of the model and all its elements.

Self-Assessment Component

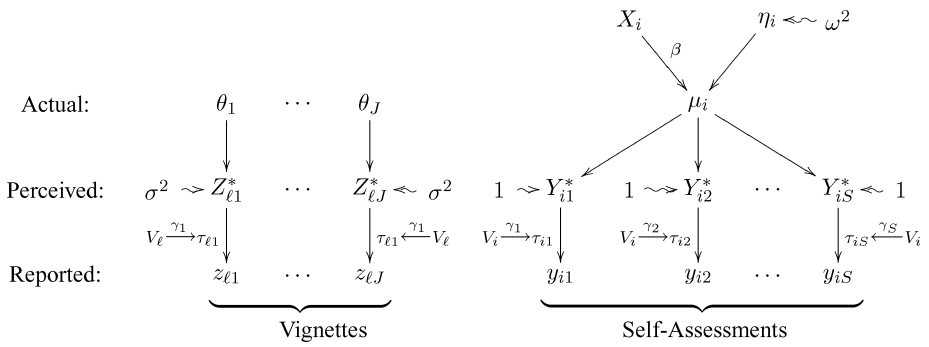
Denote the *actual* level of respondent i as μ_i ($i = 1, \ldots, n$) on a continuous, unbounded, and unidimensional scale (with higher values indicating more freedom, political efficacy, etc.). Respondent i perceives μ_i only with random (standard normal) error, as in ordered probit, so that for $s$elf-assessment question s ($s = 1, \ldots, S$),

$$Y_{is}^* \sim N(\mu_i, 1) \quad \textbf{(1)}$$

represents respondent i's unobserved *perceived* level. The actual level varies over i as a linear function of observed covariates X_i and an independent normal

[4] We have also experimented with many alternative versions, including models that generalize the "graded response" or "partial credit" frameworks more common in the psychometrics literature (Linden and Hambleton 1997). We find that the empirical results across the range of alternative models tend to be quite similar. The version we present here has the advantage of building on components that are more familiar to political scientists, but we emphasize that the particular parameterization chosen is less important than the idea of using anchoring vignettes to measure DIF directly in some way.

FIGURE 4. Parametric Model Summary



Note: Vignette questions are on the left, with perceived and reported but not actual levels varying over observations ℓ. Self-assessment questions are on the right, with all levels varying over observations i. The first self-assessment question (see Y^*_{i1}) is tied to the vignettes by the same coefficient on the variables predicting the thresholds, γ_1, and to the remaining self-assessment questions by person i's actual value, μ_i. Each solid arrow denotes a deterministic effect; a squiggly arrow denotes the addition of normal random error, with variance indicated at the arrow's source.

random effect η_i,

$$\mu_i = X_i\beta + \eta_i, \tag{2}$$

with parameter vector β (and no constant term, for identification), so that

$$\eta_i \sim N(0, \omega^2) \tag{3}$$

is modeled as independent of X_i. When $S = 1$, we drop η_i since ω is then not identified.

We elicit a reported answer for respondent i to self-assessment question s with K_s ordinal response categories (higher values indicating more political efficacy, freedom, etc.). Thus, respondent i turns the continuous perceived levels Y^*_{is} into the *reported* category y_{is} via this observation mechanism:

$$y_{is} = k \qquad \text{if} \qquad \tau^{k-1}_{is} < Y^*_{is} < \tau^k_{is}, \tag{4}$$

with a vector of thresholds τ_{is} (where $\tau^0_{is} = -\infty$, $\tau^{K_s}_{is} = \infty$, and $\tau^{k-1}_{is} < \tau^k_{is}$, with indices for categories $k = 1, \ldots, K_s$ and self-assessment questions $s = 1, \ldots, S$) that vary over the observations as a function of a vector of covariates, V_i (which may overlap X_i), and a vector of unknown parameter vectors, γ_s (with elements the vector γ^k_s):

$$\begin{aligned} \tau^1_{is} &= \gamma^1_s V_i, \\ \tau^k_{is} &= \tau^{k-1}_{is} + e^{\gamma^k_s V_i} \qquad (k = 2, \ldots, K_s - 1) \end{aligned} \tag{5}$$

(cf. Groot and van den Brink 1999 and Wolfe and Firth 2002).

Vignette Component

Denote the actual level for the hypothetical person described in vignette j as θ_j (for $j = 1, \ldots, J$), measured on a continuous and unbounded scale (higher values indicating more efficacy, freedom, etc.). The assumption of vignette equivalence is formalized by θ_j not being subscripted by (and thus assumed the same for every) respondent.

We index respondents in the sample of people asked vignettes by ℓ ($\ell = 1, \ldots, N$). (To allow vignettes to be asked of separate samples, i and ℓ may index different individuals.) Respondent ℓ perceives θ_j only with random (normal) error so that

$$Z^*_{\ell j} \sim N(\theta_j, \sigma^2) \tag{6}$$

represents respondent ℓ's unobserved real-valued perception of the level of the variable being measured described in vignette j (elements of which are assumed independent over j conditional on θ_j). (Although we avoid complicating the notation here, we also often let σ^2 vary over vignettes, since their estimates are convenient indicators of one aspect of how well each vignette is understood.)

The perception of respondent ℓ about the level of the person described in vignette j is elicited by the investigator via a survey question with the same K_1 ordinal categories as the first self-assessment question. Our software also allows other self-assessment questions, each with its own corresponding set of vignettes, but these notational complications are unnecessary for present purposes, since one set of vignettes corresponding to only one self-assessment question is sufficient to correct multiple self-assessments.

Thus, the respondent turns the continuous $Z^*_{\ell j}$ into a categorical answer to the survey question $z_{\ell j}$ via this observation mechanism:

$$z_{\ell j} = k \qquad \text{if} \qquad \tau^{k-1}_{\ell 1} \le Z^*_{\ell j} < \tau^k_{\ell 1}, \tag{7}$$

with thresholds determined by the same γ_1 coefficients as in (5) for y_{i1}, and the same explanatory variables but with values measured for units ℓ, V_ℓ:

$$\begin{aligned} \tau^1_{\ell 1} &= \gamma^1_1 V_\ell \\ \tau^k_{\ell 1} &= \tau^{k-1}_{\ell 1} + e^{\gamma^k_1 V_\ell} \qquad (k = 2, \ldots, K_1 - 1). \end{aligned} \tag{8}$$

Response consistency is thus formalized by γ_1 being

the same in both the self-assessment and the vignette components of the model.

Model Interpretation

Identification for DIF Correction. Response-category DIF appears in the model as threshold variation (τ_{is} and $\tau_{\ell s}$ varying over respondents i and ℓ) and requires at least one vignette for strong identification. We can see the essential role of vignettes by what happens if we try to estimate the self-assessment component separately and, also, set the explanatory variables X affecting the actual level to be the same as those V affecting the thresholds. In this case, β (the effect of X) and γ (the effect of V) would be dubiously identified only from the nonlinearities in the threshold model (5). This generlizes the well-known result in ordered probit that the thresholds are not all separately identified from the constant term (Johnson and Albert 1999, ch. 5).

For another view of how vignettes correct for DIF consider this simpler model based on an analogue to Aldrich and McKelvey (1977). Suppose that a single self-assessment response y_i and two vignette responses z_{ij} (for $j = 1, 2$) are continuous, perceptual error is nonexistent (i.e., the variances in Eqs. [1] and [6] are zero), and vignettes and self-assessments are asked of the same people ($i = \ell$). Then we could specify the self-assessment response (contrary to, but in the spirit of, the model) to be a linear function of the actual level with parameters that vary over respondents, $y_i = \tau_i^1 + \tau_i^2 \mu_i$, and the same for the two vignettes, $z_{ij} = \tau_i^1 + \tau_i^2 \theta_j$ (for $j = 1, 2$ and $z_{i1} < z_{i2}$). Since their values are arbitrary, we make the identifying restrictions $\theta_1 = 0$ and $\theta_2 = 1$. Finally, we solve: $\mu_i = (y_i - z_{i1})/(z_{i2} - z_{i1})$. This equation shows that the actual level is equal to the observed y_i distorted by the values on the two vignette questions. Clearly, without the vignettes, y_i would be of little use in estimating μ_i. Our model has a variety of useful features not in this simple model, but the intuition is closely analogous.

Specifying the Substantive Model. Explanatory variables X in the substantive model (Eq. [2]) must be correctly specified, just as in linear regression or ordered probit. Conditional on the model, β is interpreted as a vector of effect parameters and μ_i as the actual level (see Appendix B for details). The added random effect η_i is a strict improvement over the standard specification (when $\omega^2 > 0$), in that it recognizes that we are unlikely to be able to measure and include in X all reasons why actual levels differ across individuals. The random effect can greatly improve estimation of the actual level μ_i and, of course, makes estimates less sensitive to specification decisions about X (due to the result in the last section in Appendix B). However, it can only provide this added benefit for the portions of unmeasured explanatory variables that are unrelated to X (and it is only possible to use when multiple self-assessment questions are available). If variables omitted from and correlated with X have an effect on μ_i, we could have omitted variable bias just as in linear regression.

Specifying the Measurement Model. The explanatory variables V that predict threshold variation in the measurement model (Eqs. [5] and [8]) must also be correctly specified, but according to one of two different standards depending on the purposes for which they will be used. For our main goal of estimating the actual level μ or the effect parameters on the actual level β, V only need include enough information so that the Y and Z are independent given V (i.e., so that the product can be taken in the likelihood function in Eq. [12]). In fact, we can test this assumption nonparametrically when multiple observations are available for each unique vector of values of V_i. The test is to check that the crosstabulation of the values of Y and Z for observations that fall within each unique strata of V are unrelated. If not, then additional variables must be included. We can also perform parametric tests of this assumption by checking that elements of γ are significantly different from zero.

Measurement model specification decisions must meet higher standards if the goal is to study why different individuals have different thresholds. Then we must avoid omitted variable bias according to rules analogous to those for linear regression. The measurement model includes no random effect (and including one would be computationally complex and would make it impossible to ask vignettes and self-assessments of separate samples) and so we are not protected in the same way as with the substantive model from omitted variables unrelated to V_i.

Tests for Vignette Equivalence. Our self-assessment questions are all assumed to measure the same unidimensional actual level. If the concept is actually multidimensional, then separate questions and vignettes should be used for each dimension. Unidimensionality is best verified via standard survey techniques of extensive pretesting and cognitive debriefing. Our approach does not mean that a researcher can ignore any of the advice on writing good survey questions learned over the last half-century. We still need to be careful of question wording, question order, accurate translation of the meaning of different items, sampling design, interview length, social background of the interviewer and respondent, etc.[5]

Under our parametric model, researchers can test to a degree for vignette equivalence by checking whether the θ values are ordered as expected. The extent of ranking inconsistencies in our nonparametric model can also be indicative of multidimensionality, although care must be used in interpretation since the same "inconsistencies" can also result under our parametric

[5] Working in different languages and cultures is of course particularly difficult. For example, in our research we considered asking variants of how healthy a person is who can run 20 km. With some question wordings and translations, however, some of our pretest subjects in sub-Saharan Africa revealed in in-depth cognitive interviews that they thought anyone who would even consider running that far must be peculiar, if not mentally ill, and so would clearly be judged less healthy than someone who could only run, say, 5 km! Missing cultural differences like these would obviously threaten our approach.

model from unidimensionality and large random measurement error. The key in detecting multidimensionality is searching for inconsistencies that are systematically related to any measured variable.

Number and Type of Vignettes Needed. The optimal number of vignettes to ask (or whether to ask more vignettes or to ask the same vignettes of more respondents), in terms of the right trade-off in bias reduction vs. survey costs, depends on the nature of DIF and what information the investigator needs. For example, in some of our experiments with these methods, we were most interested in having higher resolution in measurement near the top of the scale and so we included more vignettes near that end. In general, only one vignette is needed to identify our parametric model, but we normally advise including more. In the nonparametric model, the amount of information about the actual self-assessments increases with $2J + 1$ (the number of categories of the nonparametric estimate, C) in the number of vignettes J. In both methods, the vignettes only help when they divide up the distribution of self-assessment answers and so have discriminatory power. Since the vignettes identify γ, the perfect vignette for our model is one with θ that falls between the τ's predicted by categories of V. For example, if V includes a country dummy, then the optimal vignette is one with θ between the values of the thresholds of the two countries.

When possible, we recommend asking all respondents self-assessment and vignette questions during the pretest and then studying how much information is lost by examining the stability of the γ parameters when dropping subsets of vignettes and respondents. In our experience, much of the benefit of our approach is realized after including the first two or three vignettes if they are carefully chosen to be near the self-assessments, although in practice at this early stage in using this methodology we have typically used five to seven. Similarly, in the literature on scaling roll calls, the values of only one or two legislators are typically used as anchors (e.g., Clinton, Jackman, and Rivers 2002 and Londregan 2000).

Weights on Self-Assessment Questions. When multiple self-assessment questions to measure the same construct are available, the model estimates a single actual level for all the questions. Although the variance of the perceived level is the same for each, the variance of the reported answers can still differ because the model allows the thresholds to vary across self-assessment questions. (Letting the variance at the perceived level differ also would not be separately identified or needed.) As such, under the model, questions with less measurement or perceptual error, and those that are more highly correlated with the single dimension of the concept being measured, provide more discriminatory power and are effectively weighted more heavily in estimating μ. Thus, the model's reported level provides the equivalent of the item-discrimination parameter in item-response theory or factor loadings in scaling theory. The consequence is that the actual level, μ, and effect parameters in the substantive model, β, will be fairly robust to self-assessment questions of differing quality, but studies of how and why thresholds vary over respondents will be more model-dependent.

MONTE CARLO EVIDENCE

Our parametric model meets all the standard regularity conditions of maximum likelihood and so estimates from it are consistent and asymptotically efficient. In this section we offer Monte Carlo evidence that demonstrates it has similarly desirable small sample properties. We do this by drawing 1,000 data sets from the model with a fixed set of parameters and examining the properties of the estimates of these parameters.[6] The results reported in this section are therefore conditional on the model being correct and thus do not address issues such as robustness to misspecification.

We summarize the results in Table 1, which shows that the maximum likelihood estimates and asymptotic standard errors are unbiased (i.e., within Monte Carlo approximation error of zero). Similarly, the 95% asymptotic confidence intervals seem to cover the true value about 95% of the time.

We designed this Monte Carlo experiment to simulate the conditions for which the method was designed by shifting the actual level μ_i in one direction (with the coefficient on the country dummy in X, $\beta_2 = 1$) and shifting the threshold value for a country in the same direction (so that $\gamma_1^{12} = 1$). When DIF like this occurs, the absence of an anchor means that ordered probit will not detect the change in either the coefficient or the threshold, which we demonstrate in the top two graphs in Figure 5. These graphs plot a density estimate (a smooth version of a histogram) of the estimated values across the 1,000 simulated data sets for both ordered probit and our method. As expected, ordered probit finds no difference in the actual levels between the countries because it is not able to detect the threshold variation.

A similar result occurs when studying estimates of the actual level, μ, which we illustrate using the first data set drawn with our simulation algorithm. The bottom graph in Figure 5 gives the true variation in the actual level μ (with variation coming from the random effect) for a hypothetical 65-year-old respondent and compares it to the posterior density computed by ordered probit and our model. As with the coefficients,

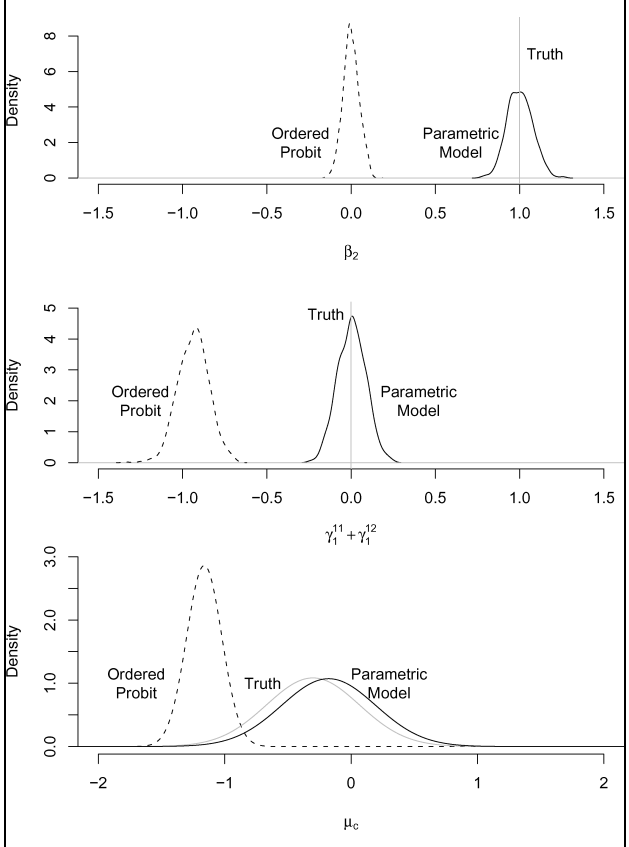
[6] To draw the 1,000 data sets, we follow this algorithm: (1) Set $\beta, \sigma^2, \omega^2, \theta$, and γ to the values in Table 1, as well as $n = N = 2{,}000$, $S = 2$, $K_s = 4$, and $J = 3$. (2) Set X to a variable corresponding to a country dummy and age, fixed across the two countries, and V to a constant and the country dummy, but (for simplicity and to save computational time) for τ^1 only. (3) Draw values for $\eta_i (i = 1, \ldots, n)$ from Eq. (3), orthogonalizing with respect to a constant and X for efficiency, and fix it for a set of simulations. (4) Finally, draw m data sets (y_{is} and $z_{\ell j}$) by repeating this algorithm m times: (a) Draw one y_{is} (for $i = 1, \ldots, n$, $s = 1, \ldots, S$) by calculating μ_i from Eq. (2); drawing Y_{is}^* from Eq. (1) for each $s (s = 1, \ldots, S)$, calculating the τ_{is}'s by Eq. (5), and calculating y_{is} from Y_{is}^* by using Eq. (4). (b) Draw one value of $z_{\ell j}$ (for $\ell = 1, \ldots, N$ and $j = 1, \ldots, J$) by drawing $Z_{\ell j}^*$ from Eq. (6) and turning $Z_{\ell j}^*$ into $z_{\ell j}$ with Eq. (7). We set $m = 100$ and then repeated the entire algorithm 10 times.

TABLE 1. Monte Carlo Analysis of Point Estimates, Standard Errors, and Confidence Interval Coverage

Parameter	True Value	Mean Bias		95% Coverage
		Point Estimate	SE	
θ_1	1	0.0042	0.000059	0.95
θ_2	−0.25	0.0027	−0.0034	0.96
θ_3	−0.7	0.0021	−0.0024	0.95
β_1: age	−0.02	0.000023	−0.000075	0.96
β_2: country	1	0.0034	−0.000097	0.95
ln(ω)	−1	−0.015	−0.0014	0.96
ln(σ)	0	0.00066	0.0019	0.95
γ_1^{11}	−1	0.001	−0.0031	0.96
γ_1^{12}: country	1	0.0029	0.00086	0.95
γ_1^{21}	−0.8	−0.0000056	0.0015	0.94
γ_1^{31}	−0.9	0.0018	0.0011	0.94
γ_2^{11}	−1.3	0.00031	−0.0045	0.97
γ_2^{12}: country	1	0.0028	0.0016	0.94
γ_2^{21}	−1	−0.0025	−0.00042	0.96
γ_2^{31}	−1	−0.0003	0.00058	0.95

Note: All estimates are given to two significant digits.

FIGURE 5. Monte Carlo Results



Note: The top two graphs display the sampling distribution of parameters across 1,000 Monte Carlo experiments, in comparison to the truth. The bottom graph compares the unconditional posterior for a hypothetical 65-year-old respondent in Country 1, based on one simulated data set.

ordered probit's inability to correct for DIF makes it miss most of the true density, while estimates from our model are on target.

EMPIRICAL EVIDENCE

To illustrate the difference our parametric approach can make compared to the most common method of analyzing ordinal dependent variables, ordered probit, we include here two very different empirical examples: a political variable, which is an extension of the political efficacy example introduced during our discussion of the nonparametric method above, and a policy outcome variable, the visual acuity dimension of health. Although many possible uses of our technology are within a single country, we choose two especially difficult examples, each requiring comparison across a pair of highly diverse countries. Since ordered probit and our model are scaled in the same way, the results from the two methods are directly comparable, although if DIF is present, only our approach would normally be comparable across cultures and people.[7]

Political Efficacy

As a baseline, we compare China and Mexico by running an ordered probit of the response to the self-assessment question on a dummy variable for country (1 for China, 0 for Mexico), controlling for years of age, sex (1 for female, 2 for male), and years of education. The results appear in the first numerical column of Table 2. The key result is the country dummy, which

[7] We estimate the model with a generic optimizer and, when multiple self-assessments are available, simple one-dimensional numerical integration.

TABLE 2. Comparing Political Efficacy in Mexico and China

		Ordered Probit		Our Method	
Eq.	Variable	Coeff.	(SE)	Coeff.	(SE)
μ	**China**	**0.670**	(**0.082**)	−**0.364**	(**0.090**)
	Age	0.004	(0.003)	0.006	(0.003)
	Male	0.087	(0.076)	0.114	(0.081)
	Education	0.020	(0.008)	0.020	(0.008)
τ^1	**China**			−**1.059**	(**0.059**)
	Age			0.002	(0.001)
	Male			0.044	(0.036)
	Education			−0.001	(0.004)
	Constant	0.425	(0.147)	0.431	(0.151)
τ^2	**China**			−**0.162**	(**0.071**)
	Age			−0.002	(0.002)
	Male			−0.059	(0.051)
	Education			0.001	(0.006)
	Constant	−0.320	(0.059)	−0.245	(0.114)
τ^3	**China**			**0.345**	(**0.053**)
	Age			−0.001	(0.002)
	Male			0.044	(0.047)
	Education			−0.003	(0.005)
	Constant	−0.449	(0.074)	−0.476	(0.105)
τ^4	**China**			**0.631**	(**0.083**)
	Age			0.004	(0.002)
	Male			−0.097	(0.072)
	Education			0.027	(0.007)
	Constant	−0.898	(0.119)	−1.621	(0.149)
Vignettes	θ_1			1.284	(0.161)
	θ_2			1.196	(0.160)
	θ_3			0.845	(0.159)
	θ_4			0.795	(0.159)
	θ_5			0.621	(0.159)
$\ln \sigma$				−0.239	(0.042)

Note: Ordered probit indicates counterintuitively and probably incorrectly that the Chinese have higher political efficacy than the Mexicans, whereas our approach reveals that this is because the Chinese have comparatively lower standards (τ's) for moving from one categorical response into the next highest category. The result is that although the Chinese give higher reported levels of political efficacy than the Mexicans, it is the Mexicans who are in fact more politically efficacious.

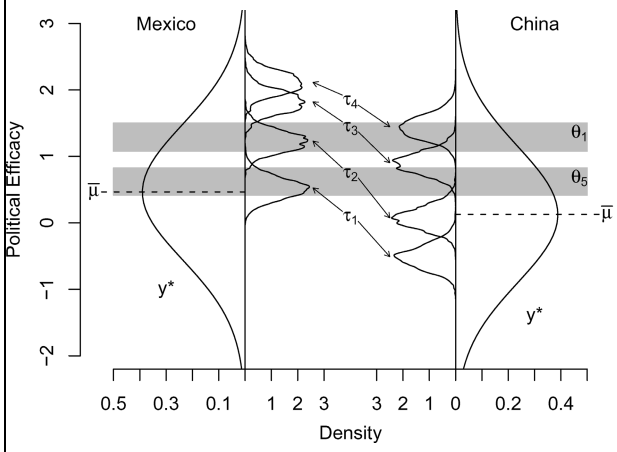
is in boldface. It shows the same remarkable result from Figure 2: Even though we have now also included controls, citizens of China choose response categories indicating higher levels of political efficacy than do citizens of Mexico. Since the underlying political efficacy scale being estimated is conditionally normal with standard deviation 1, the coefficient on the China dummy of 0.67 is quite large, and its standard error indicates that a researcher using the ordered probit model would conclude that they have a high degree of confidence in this counterintuitive conclusion.

We now use our parametric model to analyze the same example. (We include the same variables in the mean function as in the model for threshold variation. Our experiments indicate that the key results on the differences between the countries are not sensitive to many changes in these specifications.) Results appear in the last pair of columns in Table 2. The key conclusion to be drawn from our model is the opposite to that of ordered probit: the country dummy (in the top panel in boldface) has now switched signs. This means that once DIF is corrected we can see that Mexicans do indeed have higher levels of political efficacy than the Chinese. The effect (−0.364) is reasonably large and the standard error indicates considerable precision, conditional on this improved model. (Note also that the significant positive effect of education on this dimension of efficacy has not changed appreciably between the two models, which shows that correcting DIF only affects parameters related to it.)

The other parameters clarify the reason why estimates of the actual level switched and so provides some additional insight into how respondents are understanding these survey questions. To begin, note that the estimates of the actual values of the vignettes (at the bottom of Table 2) are not constrained by our model to be ordered, but they all turn out to be ordered in exactly the way we expected (as in the list above). This provides some evidence that the concept being measured is as we understood it, and thus, for example, is likely to be unidimensional.

Another important feature is the country dummy predicting each of the thresholds (given in boldface). The γ coefficient on the China dummy variable in the equation predicting τ^1 is the threshold between "no say" and "a little say." This large and significantly negative coefficient (−1.059) indicates that the same actual low level of political efficacy is considerably

FIGURE 6. Threshold Variation



Note: This figure gives a political efficacy scale from high (up) to low (down). Density estimates (a smooth version of a histogram) of the τ's are given for Mexico (on the left) and China (on the right). Also given is the distribution of Y^* and the 95% confidence interval for the actual values of the first and last vignettes (θ_1 and θ_5), in gray horizontal bars.

more likely to be judged by Chinese respondents than Mexican respondents to be a little rather than no say in government. Another way of saying this is that the Chinese have lower standards for what constitutes this particular level of efficacy. The parameterization in Eqs. (5) and (8) means that the other τ's are easier to interpret graphically, which we do in Figure 6. This figure plots the distribution of each τ across respondents, for Mexico (on the left) and China (on the right). All four of the τ distributions (pointed out in the middle of the graph) are all shifted substantially lower for China, indicating that they have lower standards for the level of efficacy in every category than the Mexicans.

Figure 6 also presents the distribution of Y^*, the unbiased self-perceptions, in each country, which shows how the τ's in each country divide up these perceived self-assessments. (The actual values, the μ_i, are not presented, but their average value, which is also the average of the Y^* distribution, does appear.) The figure also displays the 95% confidence interval for the actual value of θ_1 and θ_5 (the first and last vignette), which are constant across the two countries (see the two horizontal gray bars; the others are omitted to reduce graphical clutter). Since the power of the vignettes comes from breaking up the distribution of the thresholds, we can use the figure to evaluate the vignettes. It shows that the vignettes are best for identifying the coefficients in the τ^1 and τ^2 equations in Mexico and in τ^3 and τ^4 in China. The vignettes clearly provide much more information than necessary to identify the difference between the countries; indeed, to pick up the general direction of intercountry differences, one vignette would be sufficient. If instead we could afford to add vignettes to subsequent surveys, the extreme ends of the scale would be the most productive place to add them. Of course, other data sets need not follow this particular pattern.

So what is happening is that the Chinese respondents have lower actual levels of political efficacy than the Mexicans. But the Chinese also have even lower standards for what qualifies as any particular level of "say in government." The combination of these effects causes the Chinese to report higher levels of efficacy than those reported by the Mexicans. Thus, relying on the observed self-assessment responses for a measure of the political efficacy differences between China and Mexico would seriously mislead researchers. Using standard techniques like ordered probit to analyze these numbers would also produce badly biased results. Our parametric and nonparametric approaches reveal the problem with the self-assessments and fix it by using vignettes as anchors to generate interpersonally and interculturally comparable measures.

Although our main purpose is to design a method that makes it possible to correct for DIF to improve measurement, the reasons for these threshold differences seem well worth studying in and of themselves. This could be pursued by including other variables in the threshold portion of the model. If some of the underlying reasons for the intercountry differences were found and controlled, the coefficient on the country dummy would likely drop. We expect that research into these kinds of social–psychological questions would be a productive path to follow.

Visual Acuity

We included self-assessment and vignette questions to measure visual acuity, a fairly concrete policy outcome variable, on surveys for the World Health Organization in China ($n = 9{,}484$; completed February 2001) and Slovakia ($n = 1{,}183$; completed December 2000). Half of the respondents, randomly chosen, were asked vignette questions.

These surveys were useful because we were also able to include a "measured test" for vision—the Snellen Eye Chart test—for half of the respondents, randomly chosen. This is the familiar tumbling "E" eye chart test, with each row having smaller and smaller Es, and with respondents having to judge which direction each E is facing. Although this test is subject to measurement error, the errors should be less subject to cultural differences and so the test should provide a relatively DIF-free standard for comparison.

Our vision self-assessment question was, "In the last 30 days, how much difficulty did you have in seeing and recognizing a person you know across the road (i.e., from a distance of about 20 meters)?" with response categories (A) none, (B) mild, (C) moderate, (D) severe, (E) extreme/cannot do. We also included eight separate vignettes, such as "[Angela] needs glasses to read newsprint (and to thread a needle). She can recognize people's faces and pick out details in pictures from across 10 meters quite distinctly. She has no problems with seeing in dim light." We then followed our procedure of asking almost the same question about the people in the vignettes and with the same response categories as used in the self-assessments.

TABLE 3. Comparing Estimates of Vision in Slovakia and China Using the Snellen Eye Chart Test with Analyses of Survey Responses Using Ordered Probit and Our Approach

	Snellen Eye Chart		Ordered Probit		Our Method	
	Mean	(SE)	μ	(SE)	μ	(SE)
Slovakia	8.006	(0.272)	0.660	(0.127)	0.286	(0.129)
China	10.780	(0.148)	0.673	(0.073)	0.749	(0.081)
Difference	−2.774	(0.452)	−0.013	(0.053)	−0.463	(0.053)

Note: All numbers indicate the badness of vision, but the eye chart test is measured on a different scale than the statistical procedures.

To save space, we give results here only for our quantities of interest (see Table 3). All numbers in the table are measures of how bad the respondent's vision is. The first column is the Snellen Eye Chart test, which is an estimate of the number of meters away from an object a person with "20/20 vision" would have to stand to have the same vision as the respondent at 6 m. So the larger the number is over six, the worse the respondent's vision. In part because glasses are not generally available, and in part due to inferior health care, the Chinese, as expected, have considerably worse vision than the Slovakians. In contrast, the ordered probit model is not able to detect a significant difference between the countries at all. The Slovakians have higher standards for their own vision, which translates into higher threshold values and hence more reported values in the worse vision categories.

In contrast to the implausible and apparently incorrect ordered probit results, our approach seems to correct appropriately, producing an answer in the same direction as the measured test. The scale of the our parametric model (and ordered probit) results is not the same as the eye chart test, but we find that the Chinese have substantially worse vision than the Slovakians (0.463 on a standard normal scale with a small standard error), as in the measured test.

Measured tests provide a useful standard of comparison here for judging the relative performance of ordered probit and our model. They would also be a general solution to the problem of DIF if they could always be used accurately in place of survey questions. Unfortunately, administering these tests is far more expensive, and maintaining quality control is much more difficult, than for traditional survey questions. Part of the problem is that interviewers are trained in soliciting attitudes, not conducting medical tests. But even when highly trained medical personnel are used, the difficulties of conducting these tests in extremely diverse environments can generate substantial measurement error. In some preliminary tests we have conducted of different types of measured tests for other policy outcomes, we have found that the error in some versions of these tests swamps the error that results even from unadjusted self-assessments. Although carefully administered measured tests can provide us with a clear gold standard to evaluate our methodology for some constructs, they are infeasible for most concepts survey researchers measure, such as freedom, political efficacy, and partisan identification.

CONCLUDING REMARKS AND EXTENSIONS

The approach offered here would seem to be applicable to measuring a wide range of concepts routinely appearing in survey research. These include concepts like partisan identification, ideology, tolerance, political efficacy, happiness, life satisfaction, postmaterialism, health, cognitive attributes, attitudes, and Likert scale items measuring most attitudes, preferences, and perceptions. We do not know which of the presently used survey questions have bias due to DIF and would thus benefit from our corrections, but without some approach to verifying that survey responses are indeed interpersonally comparable, the vast majority of survey research remains vulnerable to this long-standing criticism.

We have found our survey instrumentation and statistical methods useful even when DIF is not present, as they tend to make our survey measurements far more concrete. They also often lead us to discover, clarify, and define additional dimensions of complicated concepts, and they may ultimately help develop clearer concepts.

Vignettes could be used with a modification of our model for survey responses that are closer to continuous, such as income, wealth, and prices. Indeed, our general approach might also be used to improve nonsurvey measures like the Consumer Price Index, which is derived from overlapping market baskets of goods from different historical periods. A similar approach could be used to create comparable measures of income or exchange rates over time or across cultures where the market baskets of goods chosen would also change. In these applications, instead of trying to identify something New Yorkers and Ethiopians both routinely buy, we could use DVD players for the former and goats for the latter. That is, each anchor could be designed to span only a few years or countries, so long as the entire set of observations were linked at least pairwise since it would then be correctable in a chain by many anchors analyzed together.

Ideally, our basic theoretical concepts would be sufficiently well developed so that neither vignettes nor a statistical model would be necessary. Perhaps eventually we will improve our concepts and learn how to design survey questions that apply across cultures without risk of bias from DIF. Until then, we think that survey researchers should recognize that some approach, such as the one we introduce here, will be necessary. Anchors designed by the investigator, such

as with vignettes, do not solve all the problems, but they should have the potential to reduce bias, increase efficiency, and make measurements closer to interpersonally comparable than existing methods. Moreover, researchers who are confident that their survey questions are already clearly conceptualized, are well measured, and have no DIF now have the first real opportunity to verify empirically these normally implicit but highly consequential assumptions.

APPENDIX A: THE JOINT LIKELIHOOD FUNCTION

If the random effect term η_i were observed, the likelihood for observation i, for the self-assessment component, would take the form of an ordered probit with varying thresholds:

$$\mathrm{P}(y_i \mid \eta_i) = \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[F(\tau_{is}^{k} \mid \mu_i, 1) - F\left(\tau_{is}^{k-1} \mid \mu_i, 1\right) \right]^{\mathbf{I}(y_{is}=k)}, \quad \textbf{(9)}$$

where $\mathbf{I}(y_{is} = k)$ is one if $y_{is} = k$ and zero otherwise, and F is the normal CDF and where $y_i = \{y_{is}; s = 1, \ldots, S\}$. However, since η_i is unknown, the likelihood for the *s*elf-assessment component requires averaging over η_i, in addition to taking the product over i:

$$\begin{aligned} L_s(\beta, \omega^2, \gamma \mid y) \propto \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{s=1}^{S} \prod_{k=1}^{K_s} & \left[F\left(\tau_{is}^{k} \mid X_i\beta + \eta_i, 1\right) \right. \\ & \left. - F\left(\tau_{is}^{k-1} \mid X_i\beta + \eta_i, 1\right) \right]^{(\mathbf{I}_{y_{is}=k})} \\ & \times N(\eta_i \mid 0, \omega^2) d\eta_i. \end{aligned} \quad \textbf{(10)}$$

In the special case where $S = 1$, this simplifies to

$$\begin{aligned} L_s(\beta, \omega^2, \gamma_1 \mid y) \propto \prod_{i=1}^{n} \prod_{k=1}^{K_1} & \left[F\left(\tau_{i1}^{k} \mid X_i\beta, 1 + \omega^2\right) \right. \\ & \left. - F\left(\tau_{i1}^{k-1} \mid X_i\beta, 1 + \omega^2\right) \right]^{\mathbf{I}(y_{i1}=k)}, \end{aligned} \quad \textbf{(11)}$$

which is possible by writing out the definition of the normal CDF, invoking Fubini's theorem, and solving. Equation (11) is also useful because it clearly shows that the variance of the perceived value of the vignette's level (which is set to one in the model) and ω^2 would not be separately identified if this component were estimated alone. If $S > 1$, we evaluate (10) with one-dimensional numerical integration.

The likelihood for the *v*ignette component is a J-variate ordered probit with varying thresholds:

$$\begin{aligned} L_v(\theta, \gamma_1 \mid z) \propto \prod_{\ell=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K_1} & \left[F\left(\tau_{\ell 1}^{k} \mid \theta_j, \sigma^2\right) \right. \\ & \left. - F\left(\tau_{\ell 1}^{k-1} \mid \theta_j, \sigma^2\right) \right]^{\mathbf{I}(z_{\ell j}=k)}, \end{aligned}$$

where the product terms are over observations, vignettes, and survey response categories, respectively. The likelihoods from the two components share the parameter vector γ_1 and so should be estimated together. The complete likelihood is

$$L(\beta, \sigma^2, \omega^2, \theta, \gamma \mid y, z) = L_s(\beta, \omega^2, \gamma \mid y) L_v(\theta, \gamma_1, \sigma^2 \mid z). \quad \textbf{(12)}$$

APPENDIX B: COMPUTING QUANTITIES OF INTEREST

Several quantities are of interest from this model, which we describe here, along with computational algorithms.

Effect Parameters

The effect parameters β that indicate how actual levels μ_i depend on X_i can be interpreted as one would a linear regression of Y_{i1}^* on X_i, with a standard error of the regression of one, just as in ordered probit. For example, if X_{i1} is a researcher's key causal variable, and the model is correctly specified, then β_1 is the causal effect—the increase in actual levels of freedom, or political efficacy, etc., when X_{i1} increases by one unit. (Although we have scaled our model so that it is directly comparable to ordered probit, in applications we often scale μ [and β] relative to the most and least extreme vignettes, so that the results will be simpler to interpret.)

The other set of effect parameters γ show how the thresholds τ depend on explanatory variables V. They indicate how norms and expectations differ by cultures and types of people.

Actual Levels, without a Self-Assessment Response

Suppose that we are interested in the actual level for a (possibly hypothetical) person described by his or her values of the explanatory variables, which we denote X_c. Since we have no direct information with which to distinguish this person from anyone else with the same X_c, the posterior density of μ_c is similar to that in linear regression,

$$\mathrm{P}(\mu_c \mid y, z) = N(\mu_c \mid X_c\hat{\beta}, X_c\hat{V}(\hat{\beta})X_c' + \hat{\omega}^2), \quad \textbf{(13)}$$

where we are using the asymptotic normal approximation to the posterior density of β (with mean, the MLE $\hat{\beta}$, and variance matrix $\hat{V}[\hat{\beta}]$) and are conditioning on the MLE of the random effect variance, $\hat{\omega}^2$, and the full set of data y (although we do not observe y_c). Sampling from the exact posteriors of β and ω would be a theoretical improvement, but our Monte Carlos so far indicate that these complications are unnecessary.

We can compute quantities of interest from this posterior density analytically or via simulation. For example, the actual level for a person with characteristics X_c is $E(\mu_c \mid X_c) = X_c\beta$, and the point estimate is $X_c\hat{\beta}$. Since the thresholds adjust from person to person on the basis of how they respond differently to the same questions, estimates of μ_c for any two people are directly comparable (conditional on the model).

Actual Levels, with a Self-Assessment Response

We could use the algorithm in the previous section for people we have asked a self-assessment question, but such a procedure would be inefficient, as well as more sensitive to model misspecification than necessary. Their properties are also highly dependent on the correct specification. Thus, when we have self-assessment information y_c for person c, we shall estimate $P(\mu_c \mid y, z, y_c)$ rather than $P(\mu_c \mid y, z)$ (following a strategy analogous to that of Gelman and King [1994] and King [1997]).

To see the advantage of this strategy, suppose that we are trying to measure the actual levels of Respondents 1 and 2, who have the same explanatory variable values, $X_1 = X_2$. By the unconditional method, these individuals will also have

the same posterior density, $P(\mu_1 \mid y, z) = P(\mu_2 \mid y, z)$. If they also have the same values of their explanatory variables on the thresholds, $V_1 = V_2$, and hence the same threshold values, they will have the same posterior distribution of probabilities across each of their K survey responses. But suppose also that Respondent 1 has chosen the self-assessment category y_1 with the highest posterior probability, but Respondent 2 chose the y_2 with the lowest posterior probability. In this situation, it would make sense to adjust the predictions of Respondent 2 (but not Respondent 1) in the direction of the observed value y_2, since we have this extra bit of information with which to distinguish the two cases. In other words, the observed y_2 looks like enough of an outlier to cause us to think that this person might not act like others with the same description and so should have an adjusted prediction for μ that differs from the others. (We would not wish to adjust the prediction all the way to y_2 because of interpersonal incomparability and higher variance of this realized value; i.e., there is an advantage to borrowing strength from all the other observations that are used in the predicted value.) If we had covariates with a very high discriminatory power (i.e., if ω^2 is small), very little adjustment would be necessary, whereas if our covariates did not predict well (i.e., when ω^2 is large), we would adjust more. This, of course, is classic Bayesian shrinkage, but instead of shrinking the observed value toward a global mean, we shrink toward the common interpersonally comparable adjusted value our model assigns to all people with the same values of X, μ_2.

To calculate $P(\mu_c \mid y, z, y_c)$, we start with $P(\mu_c \mid y, z)$ from Eq. (13), and use Bayes theorem to condition on y_c also, $P(\mu_c \mid y, z, y_c) \propto P(y_c \mid \mu_c, y, z) P(\mu_c \mid y, z)$, where $P(y_c \mid \mu_c, y, z)$ is Eq. (9) integrated over τ (and which we approximate by replacing τ in [9] with its MLE). Thus,

$$
\begin{aligned}
P(\mu_c \mid y, z, y_c) \propto & \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[F\left(\hat{\tau}_{cs}^{k} \mid \mu_c, 1\right) \right. \\
& \left. - F\left(\hat{\tau}_{cs}^{k-1} \mid \mu_c, 1\right) \right]^{\mathbf{I}(y_{cs}=k)} \\
& \times N\left(\mu_c \mid X_c \hat{\beta}, X_c \hat{V}(\hat{\beta}) X_c' + \hat{\omega}^2\right),
\end{aligned}
$$

which we could summarize with a histogram or a point estimate (such as a mean) and a (Bayesian) confidence interval.[8]

[8] We draw the univariate μ_c by discretization, with the inverse CDF method applied to trapezoidal approximations within each discrete area, which we find to be fast and accurate. If self-assessments and vignettes are asked of the same people, we can improve estimates even further by conditioning on both y_c and z_c:

$$
\begin{aligned}
P(\mu_c \mid y, z, y_c, z_c) \propto & \int \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[F\left(\tau_{cs}^{k} \mid \mu_c, 1\right) - F\left(\tau_{cs}^{k-1} \mid \mu_c, 1\right) \right]^{\mathbf{I}(y_{cs}=k)} \\
& \times \prod_{j=1}^{J} \prod_{k=1}^{K_s} \left[F\left(\tau_{cs}^{k} \mid \hat{\theta}_j, \hat{\sigma}^2\right) \right. \\
& \left. - F\left(\tau_{cs}^{k-1} \mid \hat{\theta}_j, \hat{\sigma}^2\right) \right]^{\mathbf{I}(z_{ij}=k)} N(\gamma \mid \hat{\gamma}, \hat{V}(\hat{\gamma})) d\gamma \\
& \times N(\mu_c \mid X_c \hat{\beta}, X_c \hat{V}(\hat{\beta}) X_c' + \hat{\omega}^2),
\end{aligned}
$$

where we assume before conditioning on y_c and z_c that β and γ are independent (which is closely approximated empirically), and we set θ, ω, and σ (which are constant over c) at their MLEs. This univariate density can be constructed by using the integral, which can be evaluated by averaging the expression for different simulations of γ, to scale the last normal at each of a grid of values on μ_c. The uncertainty in θ, ω, and σ can also be added here by drawing them from their posteriors during the simulation of the integral.

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 (March): 111–30.

Alt, James, Bo Sarlvik, and Ivor Crewe. 1976. "Individual Differences Scaling and Group Attitude Structures: British Party Imagery in 1974." *Quality and Quantity* 10 (October): 297–320.

Baum, Lawrence. 1988. "Measuring Policy Change in the U.S. Supreme Court." *American Political Science Review* 82 (September): 905–12.

Brady, Henry E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11 (June): 269–90.

Brady, Henry E. 1989. "Factor and Ideal Point Analysis for Interpersonally Incomparable Data." *Psychometrika* 542 (June): 181–202.

Cantril, Hadley. 1965. *The Pattern of Human Concerns*. New Brunswick, NJ: Rutgers University Press.

Caroll, J. D., and J. J. Chang. 1970. "Analysis of Individual Differences in Multidimensional Scaling." *Psychometrika* 35 (September): 283–319.

Cheung, Gordon W., and Roger B. Rensvold. 2000. "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling (with Comments)." *Journal of Cross-Cultural Psychology* 31 (March): 187–212.

Clarkson, Douglas B. 2000. "A Random Effects Individual Difference Multidimensional Scaling Model." *Computational Statistics and Data Analysis* 32 (January): 337–47.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2002. "The Statistical Analysis of Roll Call Data." Unpublished manuscript. Stanford University.

Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38 (June): 514–54.

Green, Donald P., and Alan Gerber. 2001. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: State of the Discipline, III*, ed. Helen Milner and Ira Katznelson. Washington, DC: APSA.

Groot, Wim, and Henriette Maassen van den Brink. 1999. "Job Satisfaction and Preference Drift." *Economics Letters* 63 (June): 363–67.

Groseclose, Tim, Steven D. Levitt, and James Snyder. 1999. "Comparing Interest Group Scores Across Time and Chambers: Adjusted ADA Scores for the U.S. Congress." *American Political Science Review* 93 (March): 33–50.

Heckman, James, and James Snyder. 1997. "Linear Probabilty Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *Rand Journal of Economics* 28 (Special Issue): 142–89.

Holland, Paul W., and Howard Wainer, eds. 1993. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Johnson, Timothy P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research." *ZUMA Nachrichten Spezial* 3: 1–40.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer.

Kahneman, Daniel, David Schkade, and Cass R. Sunstein. 1998. "Shared Outrage and Erratic Awards: The Psychology of Punitive Damages." *Journal of Risk and Uncertainty* 16 (April): 49–86.

Kinder, Donald R., and Thomas R. Palfrey, eds. 1993. *Experimental Foundations of Political Science*. Ann Arbor: University of Michigan Press.

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.

Lewis, Jeffrey B. 2001. "Estimating Voter Preference Distributions from Individual-Level Voting Data." *Political Analysis* 9 (Summer): 275–97.

Linden, Wim Van Der, and Ronald K. Hambleton, eds. 1997. *Handbook of Modern Item Response Theory*. New York: Springer.

Londregan, John. 2000. "Estimating Legislator's Preferred Points." *Political Analysis* 8 (Winter): 21–34.
Martin, Elizabeth A., Pamela C. Campanelli, and Robert E. Fay. 1991. "An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of 'Work' in the Current Population Survey." *The Statistician* 40 (September): 265–76.
Mead, A. 1992. "Review of the Development of Multidimensional Scaling Methods." *The Statistician* 41 (April): 27–39.
Palfrey, Thomas R., and Keith T. Poole. 1987. "The Relationship between Information, Ideology, and Voter Behavior." *American Journal of Political Science* 31 (September): 511–30.
Piquero, Alex R., and Randall Macintosh. 2002. "The Validity of a Self-Reported Delinquency Scale: Comparisons across Gender, Age, Race, and Place of Residence." *Sociological Methods and Research* 30 (May): 492–529.
Poole, Keith T. 1998. "Recovering a Basic Space from a Set of Issue Scales." *American Journal of Political Science* 42 (September): 954–93.
Poole, Keith, and R. Steven Daniels. 1985. "Ideology, Party, and Voting in the U.S. Congress, 1959–1980." *American Political Science Review* 79 (June): 373–99.
Poole, Keith, and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (February): 228–78.
Przeworski, Adam, and Henry Teune. 1966–67. "Equivalence in Cross-National Research." *Public Opinion Quarterly* 30 (Winter): 551–68.
Rossi, P. H., and S. L. Nock, eds. 1983. *Measuring Social Judgements: The Factorial Survey Approach*. Beverly Hills, CA: Sage.
Sen, Amartya. 2002. "Health: Perception versus Observation." *British Medical Journal* 324 (April 13): 860–61.
Shealy, R., and W. Stout. 1993. "A Model-Based Standardization Approach That Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DIF as Well as Item Bias/DIF." *Psychometrika* 58 (June): 159–94.
Sniderman, Paul M., and Douglas B. Grob. 1996. "Innovations in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22 (August): 377–99.
Stewart, Anita L., and Anna Napoles-Springer. 2000. "Health-Related Quality of Life Assessments in Diverse Population Groups in the United States." *Medical Care* 38 (September): II–102–II–124.
Suchman, L., and B. Jordan. 1990. "Interactional Troubles in Face to Face Survey Interviews (with Comments and Rejoinder)." *Journal of the American Statistical Association* 85 (March): 232–53.
Thissen, David, Lynn Steinberg, and Howard Wainer. 1993. "Detection of Differential Item Functioning Using the Parameters of the Item Response Models." In *Differential Item Functioning*. ed. Paul W. Holland and Howard Wainer. Hillsdale, NJ: Lawrence Erlbaum.
Torgerson, Warren S. 1958. *Theory and Methods of Scaling*. New York: Wiley and Sons.
Wolfe, Rory, and David Firth. 2002. "Modelling Subjective Use of an Ordinal Reponse Scale in a Many Period Crossover Experiment." *Applied Statistics* 51 (April): 245–55.