Evaluating the Credibility of Interpersonal Comparisons using Survey Experiments and Benchmarks*

Jonathan Wand†

April 28, 2008

*Research support from the Robert Wood Johnson Foundation as part of the Health Policy Scholars program is gratefully acknowledged. Thanks to the Center for Political Studies at the University of Michigan and the Victoria General Hospital for their hospitality while this work was written.

†Assistant Professor, Department of Political Science, Stanford University, and Robert Wood Johnson Health Policy Scholar, University of Michigan. Url: http://wand.stanford.edu
Email: wand(at)stanford.edu

1 Introduction

When analyzing survey data, a fundamental question must be answered: can the responses to a survey question be meaningfully compared across individuals. The comparison of responses across groups, and the pooling of responses to estimate the conditional distribution of attitudes or attributes within a group rests on the assumption that the individuals interpret and use the response categories in a comparable manner. To minimize the concern that individuals posed with the same question might not interpret it in the same manner, careful question wording in surveys is usually supplemented with validation methods including focus groups, cognitive debriefing, and translation (and back translation); see Smith (2003) for a review. However, even when people share a common understanding of a question, the responses measured by an ordinal response scale may still be interpersonally incomparable if individuals do not use the response categories in the same manner. Variability in the use of a scale or response sets could be due to a variety of causes, including differences in response styles (Hamilton, 1968; Cunningham, Cunningham, and Green, 1977) or differences in the perceived difficulty of answering the question (Cronbach, 1946, 1950).

In this paper I study how individuals use response scales when evaluating themselves in two domains of personal health, and provide an approach for evaluating methods seeking to improve the accuracy of interpersonal comparisons. In particular, I consider methods for adjusting the ordering of individuals based on their ratings of anchoring vignettes. Vignettes are increasingly used in surveys ranging from studies of political corruption to visual impairment.

This study answers (a) how well vignettes improve a researchers ability to determine an ordering of wellness of individuals; and (b) how individual self-evaluations of health status relate to validated measures of health. While the empirical inferences that can be drawn from this study are specific to the sample that is analyzed, this paper provides a set of tools that can be generally applied in other studies that seek to evaluate the interpersonal comparability of survey responses and improve the reliable measurement of attitudes

This study proposes a survey experiment design to test fundamental assumptions necessary for anchoring vignettes to improve our ability to order individuals. A key assumption of anchoring vignettes is that individuals use the same standards for evaluating themselves as they do all the vignettes in the same domain, and that each respondent perceives the vignette in the same manner. In the survey experiment, respondents were asked to evaluate the same stimuli but were randomly assigned either to receive the target stimuli first among a list of vignettes or after being asked to rate a series of other vignettes. This design enables a test of whether individuals have stable standards for evaluating questions. I show that the evaluation of the vignettes do indeed change depending on the order that the vignettes are evaluated, but find that the assumptions of one non-parametric method of using vignettes is consistent with the data.

I also consider how well alternative methods of ordering individuals perform compared to other validated measures of health. The usefulness of anchors to improve on ordering individuals as measured to objective benchmarks, has already demonstrated in other fields (King et al., 2004; Wand, 2007a). I provide an example wherein within a relatively homogeneous sample of individuals there is remarkable agreement on the use of the self-evaluation scales when compared to the Medical Outcomes Study Short Form 12 (SF-12) Physical Component Summary and Mental Component Summary (Ware, Kosinski, and Keller, 1996).

In the next section I briefly review previous research on the measurement and comparison of survey responses. I also provide an introduction to anchoring vignettes using examples from the Wisconsin Longitudinal Study (WLS) survey that I will subsequently analyze. In Section 3, I describe the design of the survey experiment. In Section 4, I review models for using anchoring vignettes to correct for DIF, and formulate the tests of assumptions that are made possible by the survey experiment. In Section 5, I present the empirical results. Finally, I conclude with a discussion of the implications of this study and future directions, including a research agenda for accounting for the challenge of differing scale use and changing standards.

2 Overview of DIF and a survey with vignettes

Survey researchers and psychologists have long been concerned with the issue that individuals posed with the same question might not interpret it in the same manner (e.g., Mosier, 1941; Jones and Thurstone, 1955). The efforts of careful question wording and validation methods are aimed at minimize disagreement in interpretation. With respect to survey design there is also an extensive literature on how respondents react to closed-ended questions as a function of what options are offered and how they are presented (e.g., Tourangeau, Couper, and Conrad, 2004). The research on the instrument effects of response options has focused on identifying the systematic biases in responses as a function of the particulars of the presentation, and demonstrates the variability of response behavior and the difficulty of making comparisons across different instrumentations. However this does not address whether we can compare individuals conditional on the same instrumentation being used.

Methods of analysis based on anchoring vignettes aim to infer how people use a scale by using each respondent's own evaluations of the common stimuli of vignettes as reference points for comparisons. The first use of common stimuli to improve the comparison of self-evaluations was by Aldrich and McKelvey (1977),[1] and the core idea is at the heart of research using explicit anchoring vignettes (Martin, Campanelli, and Fay, 1991; King et al., 2004; King and Wand, 2007; Wand, King, and Lau, Forthcoming). However, it is empirical question for any given analysis whether the assumptions necessary for using anchoring vignettes to calibrate each respondents use of the scale are more defensible than the assumptions needed to compare people using their responses on the original scale or by some other statistical model. The virtue of anchoring vignettes is that the assumptions can be empirically tested (Wand, 2007a), and this study contributes to this area of research.

The literatures within psychophysics and psychometrics have focused directly on hetero-

[1]The Item Response Theory (IRT) and related literatures also use common stimuli to measure the relative abilities of individuals but this is in general distinct from models of self-evaluations which account for differences in the interpretations of ordinal response scales.

geneity in how people use scales. Psychophysics has provided insight into the heterogeneity of how people use ordinal scales and how the use of a scale change as a result of the individuals being asked evaluate a series of stimuli (e.g., weights, colors); see McGarvey (1943) and Torgerson (1958, 78–82) for reviews of early studies.. The initial stimuli, and the rating that an individual attributes to it on the scale, has a pivotal role as a benchmark against which subsequent stimuli are rated. From psychophysics we also have evidence that the use of a scale is unlikely to be stable until the range of stimuli are observed by a respondent.[2] By bringing more data to bear on the analysis, it is hoped that it is possible to improve our ability to accurately compare individuals but any improvement relies on the assumption of stable standards across an individual's evaluations of herself and the vignettes.

The psychometric and statistical literature has primarily focused on scaling multiple measures of related self-evaluations through parametric models (e.g., Rossi, Gilula, and Allenby, 2001; Javaras and Ripley, 2007) or the transformation of ratings into rank information (e.g., Brady, 1989). A companion paper compares these multiple measure approaches to correcting for differences in scale use with the anchoring vignette based methods studied in this paper.

The standard design of the use of anchoring vignettes is as follows. All individuals are first asked to evaluate themselves on some dimension. For example, respondents in the 2004 survey of graduates in the Wisconsin Longitudinal Study (WLS) were asked,

> Overall in the last 30 days, how much of a problem did you have with moving around?

and given five possible response categories which were both numbered and labeled: (1) None; (2) Mild; (3) Moderate; (4) Severe; (5) Extreme.

[2]If a respondent is exposed repeatedly to a range of stimuli over finite range, the individual's use of the scale tends to adapt such that the high and low ends of the scale are used to describe the extremes of the scale [and in between]. The agreement in the use of a scale to describe the same stimuli also increases as individuals are exposed to the same range of stimuli. However, unlike some experiments in psychophysics where stimuli are often repeatedly administered and reevaluated by individuals, the current norm for administering anchoring vignettes is one-shot each.

Following the self-evaluation question, individuals were asked to consider a series of short stories that described individuals with health problems on the same dimension, prefaced by the following instructions:

> Imagine that the people described below are the same age that you are. Using the same scale that you used on the preceding page when talking about aspects of your own health, how would you rate the health of these people?

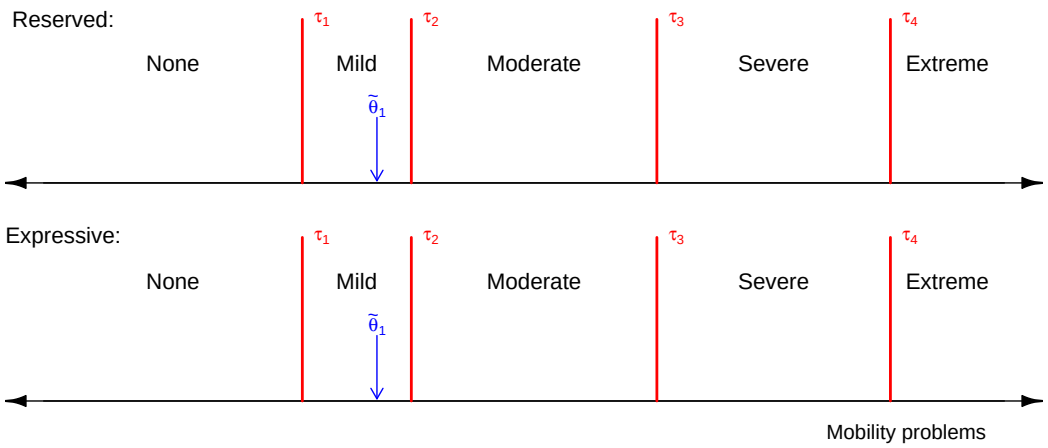
For example, the following is a vignette that describes issues with mobility:

> Richard is able to move his arms and legs but requires assistance in standing up from a chair or walking around the house. Any bending is painful and lifting is impossible. Overall, how much of a problem does Richard have with moving around?

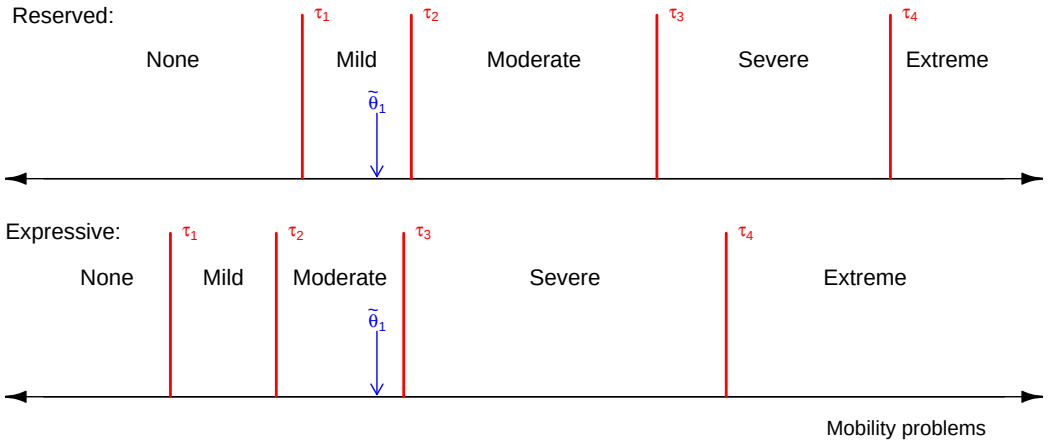
Respondent were provided with the the same five point scale to rate the vignette as they were given for their self-evaluation. The full list of self-evaluation questions and vignettes is provided in the appendix, and the order in which a subset of vignettes were asked will be discussed in the next section.

Figure 1 illustrates two scenarios for how people might differ in chopping up the continuum of "problems moving around", represented by the horizontal lines, into the five ordered response categories. I label the two types of people in each scenario as 'Reserved" and "Expressive". The scenarios vary in terms of the relative positions of the vertical "cut-points" labeled by τ symbols. Intervals between successive cut-points define response categories, and these intervals are named in the figures. I will use the psychometric phrase "Differential Item Functioning" (DIF) to generically refer to any difference in the location of cut-points across respondents.

The cut-points in the two panels of Figure 1(a) are aligned—there is no disagreement between Reserved and Expressive respondents in the mapping of the latent scale to the discrete categories. Standard models of ordered categories (e.g., ordered probit, polytomous Mokken



(a) No DIF



(c) DIF with no shared cutpoint locations

FIGURE 1: Examples of mappings between a latent dimension (horizontal axis) and observed ordered categories (vertically divided bins) describing problems with personal mobility.

Year/survey item	N	Description
1957 in school survey	10,317	1/3 random sample of all Spring 1957 high school graduates
2004 viable sample	9,018	Not confirmed dead
2004 phone survey completed	7,265	80% response rate
2004 mail survey completed	6,845	76% response rate
2004 phone and mail surveys completed	6,279	70% response rate

TABLE 1: Sample description and response rates. From Flynn, Smith, and Freese (2006)

scaling) assume that all respondents apply the same cut-points when dividing up the latent scale into categories. The panels in Figure 1(b) do not agree on the location of any cut-points, and is an example of a general form of DIF. For the same degree of problems with their disability, reserved respondents will use usually use a lesser response category than an expressive individual. For example there exist places on this continuum that an expressive person would declare a "Moderate" problem but a reserved person would declare "None". These figures are simple examples to illustrate the types of problems that are within the scope of the current analysis, though they are in fact derived from the empirical analysis that follows.

The perceived location of the vignette on the continuum of problem with mobility is included in these figures as the θ symbol and vertical arrow. The simplest assumption of the use of vignettes is that everyone perceives the same vignette at a single location, such that it is possible to detect differences in the use of scales: in 1(b) a Reserved respondent would describe the vignette as "Mild" while an Expressive respondent would describe the same vignette as "Moderate". In the next section I will elaborate on methods for combining the ratings of vignettes with a self-evaluation.

3 Design of Experiment

The survey experiments were administered to random subsets of the graduate panel of the Wisconsin Longitudinal Study (WLS). In 2004 the graduate sample was asked to complete a mail-response survey. The respondents were initially selected into the panel as part of the

	Order of Vign. Presentation				
	1st	2nd	3rd	4th	N
Form A	MO-C	AF-D	MO-A	AF-B	1240
Form B	AF-A	MO-B	AF-D	MO-C	1882
Form C	MO-D	AF-C	MO-B	AF-A	1277
Form D	AF-B	MO-A	AF-C	MO-D	1880

TABLE 2: Summary of order of presentation of vignettes by form, and sample sizes

1/3 random sample of students graduating from Wisconsin high-schools in 1957. As such respondents ranged in age from 63 to 66. The sample has been described by Flynn, Smith, and Freese (2006) and sample sizes and response rates are presented in Table 1.

Individuals in the sample were randomly assigned to receive one of four forms as part of the mail survey. The vignettes that were included on each form, and the sequence of presentation, is summarized in Table 2. The wording of the vignettes associated with each label in this table is presented in the appendix. The naming convention of the vignette labels is constructed by combining an abbreviation of the health domain as the prefix, (AF)ffect or (MO)bility, and a suffix, A, B, C or D. The order of the suffix letters also reflects the judgment of the survey writers that the vignettes can nominally be ordered in terms of increasing severity of the problems described (D being the most severe). The table also includes the number of respondents who returned the questionnaire.[3]

In the following analysis I focus on contrasting the behavior of individuals who received a particular vignette first in the sequence vignettes, and the behavior of individuals who received this same vignette last in the sequence. Thus the individuals in one treatment arm were asked to evaluate a "target" vignette prior to the presentation of any other vignette, and individuals

[3]The uneven sample sizes is due to an administrative feature of giving the survey to half the graduate sample combined with distributing the surveys in blocks. "Because these forms were administered by 1/10th random replicate to save money, we could not have a 25% administration of each form. Instead, Forms B and D were each administered to 3 replicates (30% of the sample each) and Forms A and C were each administered to 2 replicates (20%)." (Freese and Hauser, 2006)

in the other treatment arm were asked to evaluate three other vignettes and then the "target" vignette. I will refer to these two alternative treatments as receiving the target vignette "first" or "last", respectively. For the "last" treatment, I will refer to the three vignettes generically as the "prefatory" vignettes.

For clarity, I present in Table 3 the four experiments that will be analyzed, along with the ordering of vignettes that are associated with the alternative treatments. I will refer to the four experiments as AF-A, AF-B, MO-C and MO-D, reflecting the label of the target vignette. Each the forms and hence individuals appear in two treatments, one in each health domain. If an individual is assigned to the "first" treatment for an affect vignette, then she is also assigned to the "last" treatment for a mobility vignette experiment (and vice versa).

The affect vignettes that are evaluated by these experiments are relatively mild in nature, while the the mobility vignettes that are evaluated by these experiments are relatively dire. The prefatory vignettes in the affect experiments include two mobility vignettes and one affect vignettes. Conversely, the prefatory vignettes in the mobility experiments include one mobility vignette and two affect vignettes. In the mobility experiments, the prefatory mobility vignettes describe situations that are relatively nicer than the target vignette, while the prefatory affect vignette in the affect experiment is relatively worse off than the target.

A note on the mode of administering the survey is appropriate. Prior research would suggest that the mode of mail survey will potentially attenuate an ordering effect. Individuals can scan ahead, and indeed the design of the survey puts all the vignettes on one page. As such one can consider the findings of the effect of order effects to be underestimated. However, there were no instructions about how to read the page (i.e., no indication that the respondent should read all vignettes before beginning to rate any of them). Moreover, when the target vignette is given "first", the other vignette in the same domain is not adjacent, so a respondent would need to skim two questions ahead to get comparison in the same domain.

Treatment		Order of Vign. Presentation 1st	2nd	3rd	4th	Form
AF-A	First	AF-A	MO-B	AF-D	MO-C	B
	Last	MO-D	AF-C	MO-B	AF-A	C
AF-B	First	AF-B	MO-A	AF-C	MO-D	D
	Last	MO-C	AF-D	MO-A	AF-B	A
MO-C	First	MO-C	AF-D	MO-A	AF-B	A
	Last	AF-A	MO-B	AF-D	MO-C	B
MO-D	First	MO-D	AF-C	MO-B	AF-A	C
	Last	AF-B	MO-A	AF-C	MO-D	D

TABLE 3: Summary of order of presentation of vignettes by experiment, and sample sizes

4 Models of DIF and DIF correction

4.1 Definitions

Consider a set of individuals, $i \in \{1, \ldots, N\}$ each with a scalar value $\tilde{y}_i$. The premise of this analysis is that a researcher is unable to observe $\tilde{y}_i$, but individuals could in principle be compared on this implicit unidimensional latent scale. When posed with the task of self-evaluation on an ordinal scale, it is assumed that the respondent simply divides the latent scale into mutually exclusive and exhaustive intervals representing the survey response categories and picks the category in which their $\tilde{y}_i$ lies. With τ_{ik} representing the cut-point separating category k and $k+1$, the mapping between the latent location $\tilde{y}_i$ and observed self-evaluation category choice y_i is thus defined as,

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

Implicit in an analysis of ordered categorical responses is that there is agreement on the ordering of the cut-points, though people may differ in where cut-points are located. To define $K+1$

categories there needs to be K cut-points.[4]

Differential Item Functioning (DIF) is used here to describe any difference between two respondents in the location of cut-points. If $\tau_{ik} \neq \tau_{i'k}$ for any k, then the response of i and i' are subject to *DIF*, and thus comparisons on the basis of their category choice may infer an incorrect interpersonal ordering.

Anchoring vignettes are designed to describe individuals at different locations along the same unidimensional scale as the self-evaluations. Each individual may have their own idiosyncratic view on the location of the vignette on the latent scale, $\tilde{z}_{ij}$. Different methods of analysis of vignettes make different assumptions about how much people may disagree over the location of the vignette. It is useful to postulate that there is a true or average location of the vignette, $\tilde{\theta}_j$; for example if differences in the perceived locations of vignettes are due to an additive error, we would have $\tilde{z}_{ij} = \tilde{\theta}_j + \epsilon_{ij}$. In the same way that $\tilde{y}_i$ is mapped to y_i, the perceived location of anchoring item j is mapped into a response category z_{ij},

$$z_{ij} = k \quad \Leftrightarrow \quad \tau^j_{i,k-1} \leq \tilde{z}_{ij} < \tau^j_{ik}$$

The possibility that the cut-points could vary with the evaluation of each anchor is indicated by the j superscript on each τ_{ik}.

4.2 Non-parametric estimators

Non-parametric estimators seek to correct for possible DIF in the self-evaluations by ranking an individual's self-evaluation relative to her own ratings of the vignettes. The logic is that if she uses the same reasoning and standards to evaluate herself and the vignettes, we can extract interpersonally comparable information by using this transformed data that is relative to the vignettes that people all evaluate in common.

[4]It is useful to define an extended set of cut-points to include cut-points τ_0 and τ_{K+1} at the lower and bounds of $\tilde{Y}$. If $\tilde{Y} \in \mathbb{R}$ then $\tau_0 = -\infty$ and $\tau_{K+1} = \infty$.

Example	Observed Responses	Definitions in terms of shared and idiosyncratic locations			B	C
1.	$y_i < z_{i1}$	$\tilde{y}_i$	$<\tau_{i,z_{i1}-1}<$	$\tilde{\theta}_1$	1	1
2.	$y_i = z_{i1}$	$\tau_{i,z_{i1}-1}$	$< \tilde{y}_i, \tilde{\theta}_1 <$	$\tau_{i,z_{i1}}$	{ 1,2 }	2
3.	$y_i < z_{i2}$	$\tilde{\theta}_1$	$< \tau_{i,z_{i1}} <$	$\tilde{y}_i$	2	3

TABLE 4: Calculations for the nonparametric scales B, and C for relative orderings of a self-assessment, y_i and one anchor ratings z_1

In this section, I briefly review the two standard non-parametric estimators that transforms self-evaluation ratings into ranks; see Wand (2007a) for an axiomatic derivation and discussion of these methods.

The non-parametric ranking, B, converts ordinal ratings into relative ranks. The simplest case occurs when a self-evaluation is not tied with any anchoring response, and the anchor responses at least weakly adhere to the same ordering across respondents. In this case, B can is defined as,

$$B_i = j \;\Leftrightarrow\; z_{i,j-1} < y_i < z_{i,j} \tag{2}$$

Comparing this function with Equation (1) which defines the self-evaluation response y_i, Equation (2) has the same logic. However, instead of postulating relationships among unobservable quantities, B makes the same relative comparisons using only observed responses. Just as y_i can take on $K+1$ categories given K cutpoints, there are at most $J+1$ possible scalar values of B given J anchoring items. A more general presentation is provided in the Appendix, and full details are in Wand (2007a).

In the case considered here with a single vignette being used in each experiment, we have two possible true states of the world: $\tilde{y}_i < \tilde{\theta}$ or $\tilde{\theta} < \tilde{y}_i$ and hence we could observe $B = 1$ or $B = 2$. However, we could also observe $z_i = y_i$ such that it is not be possible to know whether the individual is better or worse off than the vignette. In the case of a tie it is nonetheless

possible to put bounds on quantities of interest using different scenarios for how these ties would be broken if we were omniscient. Table 4 summarizes possible observed outcomes in terms of unobserved quantities and values of B.

A related method of ordering individuals, called C (King et al., 2004), is also based on the relative ordering of self (y_i) and the vector of anchor ratings (z_i). If the anchoring items had strictly ordered responses, $z_{ij} < z_{ij'}$ for all $j < j'$, this method remaps observed responses as,

$$C_i' = \begin{cases} 1 & \text{if } y_i < z_{i1} \\ 2 & \text{if } y_i = z_{i1} \\ 3 & \text{if } z_{i1} < y_i < z_{i2} \\ \vdots & \vdots \\ 2J+1 & \text{if } y_i > z_{iJ} \end{cases} \tag{3}$$

This mapping identifies where among the $2J + 1$ possible positions relative to the anchors is an individual's self-evaluation. Examples of C for one vignette are included in Table 4 and a more general presentation of C is presented in the Appendix.

The essential difference between B and C lies in the information that is claimed to exist when a self-response is tied with an anchoring object. C assumes that there is more information than does B, however, this has been shown to come at the cost of needing to believe that the location of every individual's cut-points are the same (no DIF) or that cutpoint locations across are related across respondents in a very special way; see Wand (2007a) Proposition 1 for details. Implications of the assumptions underlying these non-parametric methods that are amenable to testing by the survey experiment will be presented at the end of this section.

4.3 Parametric models

It is also possible to use a parametric approach to jointly model self-evaluation, vignette responses, and cut-points to account for DIF. The location at which each respondent perceives the location of a vignette is assumed to be drawn from a common parametric stochastic distribution. The location of the cut-points are assumed to be some parametrized function of observed characteristics associated with each respondents.[5]

As noted above, if we designate the "true" location of vignette j as θ_j and the perceptual deviances ϵ are additive and distributed as F (some cumulative distribution F), then the probability of perceiving an anchoring item in a particular interval of the latent scale, and hence the probability of observing a particular rating, can be characterized generically as,

$$P(z_{ij} = k) = F((\tau_{ik} - \theta_j)) - F((\tau_{ik-1} - \theta_j)). \tag{4}$$

It is often assumed F is a standard normal distribution, denoted Φ. As such and assuming also a consensus among all people on the location of cut-points,

$$P(z_{ij} = k) = \Phi((\tau_k - \theta_j)/\sigma_j) - \Phi((\tau_{k-1} - \theta_j)/\sigma_j). \tag{5}$$

If we set $\theta_j = 0$ and $\sigma_j = 1$ we would have the standard ordered probit. Between the extremes of Equations 4 and 5, there is a range of alternative restrictions on the cutpoints, reflecting varying ways of characterizing the amount of common understanding about the ordinal scale across individuals and within groups.

Current common practice in the estimation of cutpoints has allowed for individuals to differ in the location of cutpoints as function of observable covariates, but conditional on those

[5]Some researchers make further a priori assumptions about the distribution of latent self-evaluation $\tilde{y}_i$ (King et al., 2004), but this is not necessary (Wand, 2007b) and is not done here.

covariates the cut-points of individuals are assumed to be identical. The ability of this type of model to account for DIF depends on the adequacy of the covariates and the appropriateness of the functional form in their parametrization. In the linear additive formulation of the model,[6]

$$\tau_{ik} \quad = \quad \tau_{ik-1} + \gamma_k X_{ik}, \qquad 0 < k < K, \quad \tau_{ik} > \tau_{ik-1} \tag{6}$$

where $\tau_{0i} = 0$ for notational convenience.

As with standard latent variable models of discrete responses, the location and scale are not identified. Given that I will be analyzing each vignette separately, in the most restricted models I will set the location-scale for each vignette in the same way for ease of interpretation of coefficients, $\theta_j = 0$ and $\sigma_j = 1$. The location of the the cut-points τ_{ki} will be estimated relative to this scale. For the other anchoring items, their locations and amount of perceptual error are also relative to this scale. The log-likelihood is thus,

$$L_i = \sum_{k=1}^{K} I(z_{ij} = k) \log(P(z_{ij} = k \mid X_i))$$

4.4 Testable implications

The assumptions needed to make reliable comparisons using the non-parametric methods have been derived axiomatically in Wand (2007a), and the assumptions of the parametric model have been enumerated in Wand (2007b). I provide here a unified summary of the assumptions that are subject to being tested within the context of the survey experiment.

The common assumption of all methods of analysis based on vignettes is that a respondent uses the same cut-points for evaluating themselves and all of the vignettes. Formally,

Assumption A. *For all* i, j, k $\tau_{ik}^{j} = \tau_{ik}$.

[6]Implicit is the restriction on the γ parameters sufficient to produce $\tau_{ik} - \tau_{ik-1} > 0$ for all k. Alternatively, this can be done without constraints by specifying $\tau_{ik} = \tau_{ik-1} + \exp(\gamma_k X_{ik})$. However, this exponentiated formulation also implies non-linear interactions between the effects of the covariates.

This property has been termed "response consistency" by King et al. (2004). See Wand (2007a) Lemma 1 for further details.

The methods differ however in the assumptions about how the perceived location of the vignette may vary across individuals. The non-parametric method C requires that every individual perceive the location of the vignette at the same location.

Assumption B.1. *For all* i, i', j, k $\tilde{\theta}_j = \tilde{z}_{ij}$

This property has been termed "vignette equivalence" by King et al. (2004). See Wand (2007a) Proposition 1 for further details.

The parametric model assumes that the perceptions of individuals are drawn from a common continuous distribution, denoted as F in the previous section. Indeed, the parametric model requires that there be "sufficient" disagreement in the perceived location of each vignette—if there is complete agreement as in Hypothesis 1, then the parametric model is not identified and cannot be estimated. Given that I will follow convention and also specify F as the normal distribution, one need only further specify that the perceptual deviations from θ_j have the same first two moments for all respondents,

Assumption B.2. $E(\tilde{z}_{ij}) = \theta_j$, $V(\tilde{z}_{ij}) = \sigma_j^2$.

The non-parametric method B is able to provide credible comparisons with or without agreement in the perceived location of a vignette. The restriction needed for B to produce credible comparisons is that

Assumption B.3. *For all* i, i', j, k $sign(\tilde{\theta}_j - \tilde{y}_i) = sign(\tilde{z}_{ij} - \tilde{y}_i)$

This assumption allows individuals to have any perception of the location up to the constraint that they are not so much in disagreement such that the perceived location is on the opposite side, relative to each respondent's own location $\tilde{y}_i$, as the true location $\tilde{\theta}_j$. Specifically, $\tilde{\theta}_j < \tilde{y}_i < \tilde{z}_{ij}$ and $\tilde{z}_{ij} < \tilde{y}_i < \tilde{\theta}_j$ are precluded by this assumption. See Lemma 1(b) in Wand (2007a).

This implication is weaker than the "vignette equivalence" property which requires that every individual perceives the anchoring objects in exactly the same latent location: $\tilde{\theta}_j \equiv \tilde{z}_{ij}$ for all i, j. The B method also does not require there necessarily be disagreement about the vignette location as is required by the parametric model.

The randomized survey experiment enables a joint test of these assumptions. Under the scenario that the ordering of the vignette presentation has no effect on the perception of the vignettes or the standards for using the scale, we have equality of the potential outcomes under the two treatment, $t \in \{First, Last\}$,

$$\tau_i(t = First) = \tau_i(t = Last) \quad \text{and} \quad \tilde{z}_i(t = First) = \tilde{z}_i(t = Last) \tag{7}$$

The randomization of individuals to each treatment implies that one treatment group should not systematically differ from another in any respect. Most importantly, the groups should not systematically differ in the distribution of cut-points for evaluating themselves or in the perception of where the vignette is located (if it differs at all across individuals).

Thus, if Assumption A (response consistency) and Assumption B.1 (vignette equivalence) hold then we should see no significant difference in the distribution of ratings of a vignette across treatment groups within an experiment. Let r_{tk} be the proportion of individuals who rate a vignette as category k in treatment arm $t \in$ {First,Last}. The assumption of $z_i(t = First) = z_i(t = Last) = \theta$ maintained by C and the random assignment to treatment implies that there should be no association between ratings and treatment:

Hypothesis 1. $r_{fk} = r_{lk}$ *for all* k

Finding that the distribution of responses differ across the treatment arms, and hence rejecting Hypothesis 1, means that for at least one treatment group respondents can either not be using the same standards for their self-evaluation or the perceived vignette location varies by treatment. Failing to reject this hypothesis, however, is not evidence that respondents are

indeed using the same standards to evaluate themselves and the vignette, but at least the data is consistent with the theory that standards are are not affected by the range and order of the stimuli presented.

The test of the assumptions of the parametric model, namely that Assumption A (response consistency) and Assumption B.2 (common perceptual error) jointly hold, is identical to the test of Hypothesis 1. Due to randomization, the two treatment groups should have the same a priori distribution of perceptual errors and the same distribution of cut-points. As such we would expect the proportions of responses to also be the same under the parametric model.

The parametric model has the additional feature, however, that it is possible to decompose differences between treatment groups into effects that are due to differences in cutpoints and differences in perception of the vignette. This leads to three alternatives of interest,

$$\begin{array}{lll} \text{(i)} & \tau_i(t = First) = \tau_i(t = Last) \quad \text{and} \quad \tilde{z}_i(t = First) = \tilde{z}_i(t = Last) \\ \text{(ii)} & \tau_i(t = First) = \tau_i(t = Last) \quad \text{and} \quad \tilde{z}_i(t = First) \neq \tilde{z}_i(t = Last) \\ \text{(iii)} & \tau_i(t = First) \neq \tau_i(t = Last) \quad \text{and} \quad \tilde{z}_i(t = First) \neq \tilde{z}_i(t = Last) \end{array} \tag{8}$$

If 8(i) cannot be rejected when compared to (ii) or (iii), then the data is consistent with the use of the parametric model. Otherwise, if (ii) cannot be rejected relative to (iii) then the data is consistent with using anchoring vignettes in a fixed order, but not in the randomized order; this finding would at least suggest a strategy of holding fixed the order of vignettes in a future survey rather than being a critique of the use of vignettes overall. If we reject (i) and (ii) when tested against (iii), then this is an indication of a failure of the assumptions needed by the parametric model.

As previously mentioned, the location and scale need to be set in order for the ordinal model to be identified. There are many equivalent ways of doing this. I set the scale by fixing the variance of the vignettes ($V(z_i(t = First))V(z_i(t = Last)) = \sigma = 1$). I also set the location of the mean of the vignette for the First treatment group ($E(z_i(t = First)) = \theta_f = 0$). The

location for the Last treatment group needs to also be fixed (e.g., also setting $E(z_i(t = Last)) = \theta_f = 0$) or by constraining a cutpoint to be the same across treatment groups. I adopt the latter formulation though I emphasize that they are equivalent. Thus the mean of the Last treatment group vignette (θ_f) is thus estimable as are differences in the widths of cutpoints between treatment groups. I will refer to this least restricted model as "M3".

Since the parametric model will include additional individual specific covariates in X, it will be useful to decompose the vector of parameters into the subset of cut-point parameters that are not a function of the assignment to a treatment (γ_k') and those that are (γ_k''), such that $\gamma_k = (\gamma_k', \gamma_k'')$. In the empirical analysis that follows, γ_k' will be parameters that allow cutpoints to vary as a function of the sex and psychological characteristics of each respondent, while γ_k'' will allow cutpoints to differ if the respondent received the target vignette last (i.e., a dummy variable for treatment assignment). If 8(i) holds, then

Hypothesis 2. $\gamma_k'' = 0$ *for all* k*, and* $E(z_i(t = Last)) = 0$.

I will refer to the model that imposes these restrictions as "M1". This is essentially the same as Hypothesis 1 re-framed in terms of parameters of the ordered probit. Alternatively, if 8(ii) holds, then the perception of vignette location may differ across treatment groups. I test whether a mean-shift alone could account for differences in the distribution of vignette ratings,

Hypothesis 3. $\gamma_k'' = 0$ *for all* k*, and* $E(z_i(t = Last)) \neq 0$.

The model that imposes restrictions on the spacing of cutpoints but not the equality of means will be referred to as M2.

The non-parametric method B can be tested on the basis of the distribution of transformed responses. Similar to the parametric model, it is not sufficient that vignette distribution is different to infer that the assumptions about the cutpoints do not hold. Even if there are different biases in the perceived location of the vignettes in different treatment arms, this need not impede correct ordering of individuals unless the Assumption B.3 is violated. The

maintained hypothesis of Assumptions A and B.3 holding can be stated as,

Hypothesis 5.

$$[P_f(B=1), P_f(B=1)+P_f(B=\{1,2\})] \cap [P_l(B=1), P_l(B=1)+P_l(B=\{1,2\})] \neq \emptyset.$$

In other words, there exists overlap in the bounds of $P(B=1)$ between the treatment groups, indicated by the subscripts f and l.

5 Empirical Analysis

The following analysis focuses on data from four survey experiments where respondents were asked a rate themselves and a vignette on two domains of health, depression and mobility. The questions that were used to elicit evaluations on each of these domains were, respectively,

- How much of a problem did [you/he/she] have with feeling sad, low or depressed?
- How much of a problem did [you/he/she] have with moving around?

The respondents were asked to answer these questions using a five category that was both numbered and labeled: (1) None; (2) Mild; (3) Moderate; (4) Severe; (5) Extreme. The analysis will focus on contrasting and testing the behavior of respondents within each of four experiments which varied the order in which the target vignette was presented.

I begin with an analysis of the univariate distributions of pre and post-treatment responses. Consistent with the intended random assignment to treatments, I find that responses to questions asked prior to posing the randomly assigned vignettes in the survey are the same across treatment groups. In contrast, the distribution of ratings of a vignette is not the same across treatments. This difference in vignette rating implies that the assumptions needed to make credible interpersonal comparisons using C do not hold, and that one should not pool responses from different ordering of vignettes in a parametric model.

I also consider the transformation of a self-evaluation by ranking it relative to an individual's rating of the vignette. I find that one cannot reject that the non-parametric method B is invariant to the ordering of vignettes.

Finally, I compare the ranking of individuals implied by their self-evaluations and by the non-parametric ranking of C using the Medical Outcomes Study Short Form 12 (SF-12) Physical Component Summary (PCS) and Mental Component Summary (MCS) (Ware, Kosinski, and Keller, 1996). I find that individuals are on average ordered by their self-evaluations in the same manner as the SF-12 scale indicates, and that there is additional information in B to distinguish the ranking of individuals.

5.1 Distributions of responses

The distribution of self-evaluations in each treatment are shown in Table 5. Most individuals declare no problem with either depression or mobility, and about a quarter declare mild problems and less than fifteen percent of the population declare a problem as moderate or worse.

There is also no sizable difference between the distribution of responses in the two arms of each experiments. Moreover, one cannot reject at a conventional .05 level the hypothesis that the distribution of self-evaluations do no systematically differ across ordering of vignettes as evaluated by a chi-square test of association including non-response as a category: χ^2 statistics of 23.8 and 9.8 for affect and mobility respectively, on 15 degrees of freedom. The large statistic for affect is of concern, and a test between treatment arms within experiment AF-B would reject at a .05 level a lack of association ($\chi^2 = 11.4$ on 5 degrees of freedom), as shown in the last two columns of Table 5. This significant empirical correlation between assignment and self-reported well-being should nonetheless make us cautious of any attempt at interpreting the effects of the experiment AF-B due to the observed imbalance on a key pre-treatment variable.[7] As such,

[7]Since the association is not due to non-response to these questions, it is unlikely that there is a selection bias

Distribution of Depression Self-ratings										
Treatment		Blank	None	Mild	Moderate	Severe	Extreme	Mean	χ^2	P_{χ^2}
AF-A	First	0.020	0.63	0.27	0.07	0.008	0.003	1.43	3.53	0.62
	Last	0.019	0.63	0.27	0.07	0.003	0.005	1.42		
AF-B:	First	0.019	0.66	0.25	0.06	0.011	0.000	1.38	11.40	0.04
	Last	0.015	0.64	0.26	0.08	0.008	0.002	1.44		
Distribution of Mobility Self-ratings										
Treatment		Blank	None	Mild	Moderate	Severe	Extreme	Mean	χ^2	P_{χ^2}
MO-C:	First	0.016	0.64	0.22	0.09	0.025	0.005	1.48	4.413	0.494
	Last	0.015	0.62	0.23	0.10	0.018	0.005	1.51		
MO-D:	First	0.020	0.63	0.22	0.10	0.024	0.009	1.50	2.916	0.712
	Last	0.019	0.63	0.23	0.09	0.022	0.005	1.48		

TABLE 5: Distribution of responses to self-evaluation questions

the other three experiments are the main objects of interest in the remaining discussion.

As another test of the random assignment of individuals across treatments, Table 6 summarizes hypotheses tests of whether the distribution of SF-12 Physical Component Summary (PCS) and Mental Component Summary (PCS) scale values differ across treatments within each experiment. The distributions of SF-12 scale values are evaluated using a bootstrapped Kolmogorov-Smirnov test (Sekhon, 2008).[8] There is no significant differences across treatment groups.

Table 7 summarizes the distribution of responses for the target vignettes across treatment. Moving down the table, the target vignettes are judged by respondents as increasingly bad on the five point scale. This can be seen either by looking at the distribution of responses in the different categories, or as a quick summary simply by looking at the simple mean, which is obtained by giving integer values 1-5 to the categories and taking the average among

between the samples. As such, using a matching method to select a subset of observations in AF-B experiment that are indeed balanced on pre-treatment variables is a potentially useful pursuit for future research.

[8]The SF-12 scales are intended to be follow a normal distribution in the population as a whole (Ware, Kosinski, and Keller, 1996), but in practice the scale is generated from discrete data and there are many ties across people in scale values. The trap KS-test provides correct size even under ties.

	SF-12 Mental Component		SF-12 Physical Component	
Experiment	D	P_D	D	P_D
AF-A	0.035	0.29	0.023	0.79
AF-B	0.026	0.65	0.031	0.44
MO-C	0.032	0.41	0.034	0.35
MO-D	0.029	0.50	0.041	0.15

TABLE 6: Equality of distribution of pre-treatment variables. Bootstrapped Kolmogorov-Smirnov (B=10,000). D is the largest observed deviance.

respondents. The median respondent places the affect vignettes in the "moderate" category while placing the mobility vignettes in the "severe" category.

Comparing the distribution of responses across treatment arms in Table 7 we see difference as great as .23 in the proportion choosing a category (MO-C extreme category: .11 vs .34). Moreover, the distribution of responses are significantly different between treatments within each experiment; the last two columns of Table 7 summarizes the hypothesis tests for each experiment. In all four experiments we can reject Hypothesis 1 that the data are consistent with the use of nonparametric method C for ordering individuals.

The comparison of the fits for alternative parametric specifications are presented in Table 8. The least restricted model (M3), described in the previous section, allows the spacing of the cutpoints to differ across treatment groups, as well as allowing different mean and variance for the same vignette across treatment groups. This model, along with the others that will be estimated, allow cutpoints to differ by the sex of the respondent as well as by whether the respondent thinks of herself as reserved.[9] The parameter estimates for M3 are included in the appendix Table 14, and the estimates of the other models are available from the author as a separate appendix.

[9] The reserved variable is a dummy variable for those agreed (slightly, moderately, or strongly) with the question "I see myself as someone who is reserved".

Distribution of Depression Vignette ratings

Treatment		Blank	None	Mild	Moderate	Severe	Extreme	Mean	χ^2	P_{χ^2}
AF-A	First	0.03	0.04	0.38	0.42	0.12	0.01	2.58	13.97	0.016
	Last	0.02	0.03	0.38	0.45	0.10	0.01	2.62		
AF-B:	First	0.04	0.03	0.13	0.40	0.36	0.04	3.15	23.73	<0.001
	Last	0.04	0.02	0.16	0.45	0.29	0.05	3.07		

Distribution of Mobility Vignette ratings

Treatment		Blank	None	Mild	Moderate	Severe	Extreme	Mean	χ^2	P_{χ^2}
MO-C:	First	0.04	0.02	0.03	0.23	0.58	0.11	3.61	304.29	<0.001
	Last	0.03	0.01	0.01	0.09	0.53	0.34	4.08		
MO-D:	First	0.02	0.02	0.02	0.11	0.64	0.18	3.86	191.16	<0.001
	Last	0.04	0.02	0.01	0.06	0.47	0.39	4.10		

TABLE 7: Distribution of vignette evaluations

The most restricted model considered (M1) constrains the location of the cutpoints and the distribution of the vignette to be the same across the treatment groups. The likelihood ratio test of M3 versus M1 is also shown in Table 8, and for all experiments the constraints of M1 are rejected at any plausible level. We can thus reject Hypothesis 2 in all four experiments, and pooling of responses across different vignette orderings is not defensible in these parametric models. The label M2 refers to the the parametric model that allows differences in the mean of distribution of vignette locations (z_{ij}) across treatment groups, while constraining the cutpoint locations to be the same. The likelihood ratio tests of M2 versus M3 also reject Hypothesis 3 for all experiments.

5.2 Combining self-evaluations and vignette ratings

The distribution of ranks of each respondent's self-evaluation relative to their rating of the target vignette is presented in Table 9. In this transformed data, the differences between treatments are attenuated, and the formal tests that these distributions are the same are presented in first column of Table 10. The differences between treatments is still significant in the transformed

	Experiment			
	AF-A	AF-B	MO-C	MO-D
−Log-likelihoods of Models:				
M3, Varying τ spacing, $E_l(\theta)$	3449.6	3684.4	3067.6	3080.8
M2, Varying $E_l(\theta)$	3455.5	3692.6	3075.1	3101.5
M1, Homogeneous parameters	3455.8	3696.3	3225.8	3175.6
Likelihood Ratio Tests of Hypotheses:				
Hypothesis 2, Probability M1 same as M3	0.008	< 0.001	0.002	< 0.001
Hypothesis 3, Probability M2 same as M3	0.016	< 0.001	0.002	< 0.001

TABLE 8: Fit and hypothesis tests from ordered probit models of combining vignette ratings across treatments

data for the two mobility vignettes, but not the affect vignettes. Note that this lack of difference between treatment groups for the affect vignettes might lead some to conclude that one could defend the use C, but again it is not justified by the rejection of Hypothesis 1 in the previous subsection. I thus focus on testing Hypothesis 5 to establish whether the data is consistent with the assumptions necessary to justify the use of the nonparametric method B.

The main difference between treatment arms lies in the proportion who rate themselves as better off than the vignette and those that rate themselves the same—the proportion who rate themselves as worse of is essentially the same. Given the concentration of individuals at the low end of the self-evaluation scale, it is not surprising that it is relatively rare for any respondent to see herself as worse off than even the mildest vignette, but the proportion who do so are not significantly different between treatments; chi-square tests of this contrast are presented in the second column of Table 10. The key difference is that in three of the four experiments there are more people who put themselves at the same location as the target vignette if the target is given first.

The test of Hypothesis 5 requires calculating the intersection of the bounds of $P(B = 1)$ across treatment groups. If the intersection is the empty set, then the data is inconsistent with

Treatment		Better rating (Self < Vign)	Same rating (Self = Vign)	Worse rating (Self > Vign)	Bounds on P(B=1)	Intersection Hypothesis 5
AF-A	First	0.736	0.211	0.053	(0.736 , 0.947)	(0.769,0.947)
	Last	0.769	0.183	0.048	(0.769 , 0.952)	
AF-B	First	0.865	0.105	0.029	(0.865 , 0.970)	(0.865,0.963)
	Last	0.846	0.117	0.036	(0.846 , 0.963)	
MO-C	First	0.911	0.071	0.018	(0.911 , 0.982)	(0.954,0.982)
	Last	0.954	0.032	0.014	(0.954 , 0.986)	
MO-D	First	0.924	0.062	0.014	(0.924 , 0.986)	(0.948,0.986)
	Last	0.948	0.041	0.011	(0.948 , 0.989)	

TABLE 9: Distribution of ranking of self-evaluations relative to anchoring vignette.

	Relative Rating Vignette vs Self		Self Rated as Worse than Vign	
Experiment	χ^2	P_{χ^2}	χ^2	P_{χ^2}
AF-A	4.452	0.108	0.386	0.534
AF-B	2.244	0.326	0.798	0.372
MO-C	26.061	< 0.001	0.308	0.579
MO-D	7.340	0.025	0.224	0.636

TABLE 10: Distribution of Relative Ranks: Tests of no association between treatment and transformed responses

the use of the non-parametric ordering C. The intersections of the proportions are presented in the final column of Table 9, and for all four experiments the interval has positive length.

5.3 Evaluating comparisons using benchmark measures

In this section I use values from the SF-12 Mental and Physical Component Summary scales to compare the ranking implied by the self-evaluations and the ranking implied by the non-parametric B method. Since one cannot reject that the orderings produced by B are the same across treatment groups, I pool the response within each experiment. The following section proposes an approach to comparing scales to a benchmark.

		Mean SF-12 MCS	Mean SF-12 MCS by Relative Rank			No Diff in Means	
	Self		Better	Same	Worse	P_{t-test}	N
AF-A	None	58	58	57			1843
	Mild	54	54	53	52	(0.04 , 0.13)	812
	Mod.	45	49	45	45	(0.04 , 0.14)	201
	Sev.	33			33		17
	Extr	30		24	30		8
AF-B	None	58	58	56			1897
	Mild	53	53	52	52	(0.09 , 0.20)	752
	Mod.	44	45	44	44	(0.01 , 0.04)	190
	Sev.	35	50	34	35		29
	Extr	40			40		3

TABLE 11: SF-12 Mental Component Summary (MCS) by depression self-evaluation rating and rank relative to vignette and vignette rating.

Ideally one would like a benchmark scale that measures the exact same quantity as the self-evaluations and vignettes. The stem of the self-evaluations and vignette evaluations questions ("How much of a problem did [you/he/she] have with ...") asks for a summary of the extent and impact of a disability. The SF-12 scales were not designed to validate these specific self-evaluation question, but are built from a series of questions about the respondents (in)ability to perform certain normal tasks or actions related to mobility and depression. The SF-12 scales are formulated such that higher values indicate better health in that domain. Substantively, it is an open question how the self-evaluation responses and the SF-12 scales are related.

In Tables 11 and 12 the mean SF-12 value in the Mental and Physical domains, respectively, are summarized by both self-evaluation rating and ranking relative to the vignette (i.e., the components of C). The first order differences is across self-evaluations ratings (rows) rather than across relative ranks (columns within a row). Rarely do any rows have overlaps in means. Within a row, however, the relative information revealed by C does provide additional information—those who claim to be better off than the vignette have better ratings on the SF-12 scale than those who rate themselves as worse of than the vignette.

		Mean SF-12 PCS	Mean SF-12 PCS Rating by Relative Rank			No Diff in Means	
	Self		Better	Same	Worse	P_{t-test}	N
MO-C	None	53	53	52			1827
	Mild	45	45	42	41	(0.15 , 0.27)	671
	Mod.	35	35	31	34	(0.08 , 0.18)	284
	Sev.	27	30	28	23	(0.11 , 0.21)	55
	Extr	25		24	26		15
MO-D	None	53	53	53			1867
	Mild	45	45	45	46	< 0.001	673
	Mod.	34	34	32	42	(0.03 , 0.12)	274
	Sev.	27	29	26	24	(0.01 , 0.04)	61
	Extr	22		21	22		18

TABLE 12: SF-12 Physical Component Summary (PCS) by mobility self-evaluation rating and rank relative to vignette

Where there is sufficient data across columns within a row, I provide the formal test of whether respondents that rate themselves lower than the vignette have significantly lower SF-12 ratings than those respondents who rate themselves as better off than the vignette. Given the definition of B, the respondents who rate themselves as the vignette can be considered as either better or worse off than the vignette. I integrate over the possible permutations of the allocations of the "same as" cases via Monte Carlo simulation and present the 95 percentile interval of the t-test probabilities that the better off and worse off respondents have the same mean SF-12 values.[10]

[10]Each Monte Carlo simulation randomly assigned each respondent having a tie between the self-evaluation and vignette rating to being better or worse off. In each simulation these respondents had an equal chance of each assignment in these Bernoulli trials. Means and variances were then calculated for the better and worse off groups, as well as the t-test and probability of equality of the means. This was repeated 1,000 times and the 95 percent intervals of the probabilities are presented in Tables 11 and 12.

6 Conclusions

This is the first study to systematically consider whether the use of anchoring objects help or hinder the credible comparisons of individuals. For the data that was analyzed in this paper, the answer is that the use of anchoring vignettes are defensible and helpful when combined by one method (the nonparametric B ranks), and not defensible when combined with other methods (the nonparametric C ranks or the ordered probit models).

One motivation for randomly assigning respondents to a subset of data is to reduce the cost of using vignettes while still allowing for a parametric adjustment for differences in the use of a scale. The difficulty with this approach that has been illuminated by this study is that the use of the scale may change depending on which vignettes are asked and the order in which they are presented. The results of this study are consistent with the results from studies in psychophysics.

It is unlikely that survey research will be able to adopt methods of psychophysics of repeatedly asking people to reevaluate items until their use of the scale stabilizes. Two current research projects are underway provide alternatives to the approach considered in this paper. One is developing a computer-based survey instrument allowing individuals to adjust or amend their ratings retroactively in light of new vignettes being posed. Enabling a respondent to change how they would use the scale in light of a series of vignettes will provide further insight into how individuals use response scales. Another is to directly ask respondents to make pairwise comparisons between themselves and the vignettes. Asking respondents to undertake ranking exercises has been argued to be preferable in other contexts (Alwin and Krosnick, 1985), and the greater cognitive engagement may also prove useful in the context of self-evaluations. However, the results of this paper temper the expected benefits of relying solely on a ranking approach of direct comparisons by providing examples where there is important ordering information in the self-rating scale that is not contained in the relative ranks.

Appendix

The following is the wording of the survey questions for self-evaluation, and the list of anchoring vignettes. The self-evaluation questions were prefaced by "Overall in the last 30 days....", and then individuals were asked to rate their health on each of the following domains.

- How much of a problem did you have with moving around?
- How much difficulty did you have in vigorous activities, such as running 2 miles or cycling?
- How much of a problem did you have with feeling sad, low or depressed?
- How much of a problem did you have with worry or anxiety?

For each question, respondents could choose one of five response categories: (1) None; (2) Mild; (3) Moderate; (4) Severe; (5) Extreme.

Following the self-evaluation questions, individuals were asked to evaluate a series of short vignettes that described individuals with health problems. The vignettes were prefaced by the following instructions:

> Imagine that the people described below are the same age that you are. Using the same scale that you used on the preceding page when talking about aspects of your own health, how would you rate the health of these people?

The list of vignettes from which individuals were randomly assigned are as follows.

AF-A [Name] enjoys his [her] work and social activities and is generally satisfied with his [her] life. He [she] gets depressed every 3 weeks for a day or two and loses interest in what he [she] usually enjoys but is able to carry on with his [her] day to day activities.

AF-B [Name] worries often about his [her] health. He [She] gets depressed once a week for a day or two, thinking about what could go wrong and all the illnesses he [she] could get, but is able to come out of this mood if he [she] concentrates on something else.

AF-C [Name] feels nervous and anxious. He [She] worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests him [her]. When he [she] is alone he [she] tends to feel useless and empty.

AF-D [Name] feels depressed most of the time. He [she] weeps frequently and feels hopeless about the future. He [She] feels that he [she] has become a burden on others and that he [she] would be better dead.

For each of these vignettes, respondents were asked

- How much of a problem did [he/she] have with feeling sad, low or depressed?
- How much of a problem did [he/she] have with worry or anxiety?

MO-A [Name] is able to walk distances of up to 1/8 mile without any problems but feels tired after walking 1/2 mile or climbing up more than one flight of stairs. He [She] has no problems with day-to-day physical activities, such as carrying food from the market.

MO-B [Name] does not exercise. He [She] cannot climb stairs or do other physical activities because he [she] is obese. He [She] is able to carry the groceries and do some light household work.

MO-C [Name] has a lot of swelling in his [her] legs due to his health condition. He [She] has to make an effort to walk around his home as his [her] legs feel heavy.

MO-D [Name] is able to move his [her] arms and legs, but requires assistance in standing up from a chair or walking around the house. Any bending is painful and lifting is impossible.

For each of these vignettes, respondents were asked

- How much of a problem did [he/she] have with moving around?
- How much difficulty did [he/she] have in vigorous activities, such

Non-parametric estimators

A brief, general statement of the estimators is given here; more details are found in Wand (2007a). Define

$$B_i = \{j+1, j+2, \ldots, j'\} \;\Leftrightarrow\; z_{ij} < y_i < z_{i,j'} \tag{9}$$

To determine j and j', first identify the vignettes that have responses that are strictly less than the self-evaluation response, $\mathcal{J}_{i1} = \{j : z_{ij} < y_i\}$. Similarly let $\mathcal{J}_{i2} = \{j : z_{ij} > y_i\}$. Then $j = \min(\max(\mathcal{J}_{i1}), \min(\mathcal{J}_{i2}) - 1)$ and $j' = \max(\max(\mathcal{J}_{i1}) + 1, \min(\mathcal{J}_{i2}))$. If the anchor responses are in the correct (weak) order, these functions simplify to $j = \max(\mathcal{J}_{i1})$ and $j' = \min(\mathcal{J}_{i2})$.

C_i is constructed as the set of sequential integers

$$C_i = \{\min C_i^*, \ldots, \max C_i^*\} \tag{10}$$

using

$$C_i^* = \left\{ \begin{array}{l} 1 \times I(y_i < z_{i1}), \\ \ldots \\ (2J+1) \times I(y_i > z_{iJ}) \end{array} \right\} \setminus 0 \tag{11}$$

where $I(x) = 1$ if x is true and equals zero otherwise, and $\{s\}\backslash 0$ indicates that zero values are removed from the set $\{s\}$. Table 13 gives examples of the non-parametric mapping of the vignettes and self-evaluations into the transformed rankings.

Parameter estimates

Table 14 provides parameter estimates for the unrestricted model M1.

Example	Observed Responses	Definitions in terms of shared and idiosyncratic locations	B	C
1.	$y_i < z_{i1} \leq z_{i2}$	$-\infty < \tilde{y}_i < \tau_{i,z_{i1}-1} < \tilde{\theta}_1$	1	1
2.	$y_i = z_{i1} < z_{i2}$	$-\infty < \tau_{i,z_{i1}-1} < \tilde{y}_i, \tilde{\theta}_1 < \tau_{i,z_{i1}} < \tilde{\theta}_2$	{ 1,2 }	2
3.	$z_{i1} < y_i < z_{i2}$	$\tilde{\theta}_1 < \tau_{i,z_{i1}} < \tilde{y}_i < \tau_{i,z_{i2}-1} < \tilde{\theta}_2$	2	3
4.	$z_{i1} < y_i = z_{i2}$	$\tilde{\theta}_1 < \tau_{i,z_{i2}-1} < \tilde{y}_i, \tilde{\theta}_2 < \tau_{i,z_{i2}} < \tilde{\theta}_2$	{ 2,3 }	4
5.	$z_{i1} \leq z_{i2} < y_i$	$\tilde{\theta}_2 < \tau_{i,z_{i2}} < \tilde{y}_i < \infty$	3	5
6.	$y_i = z_{i1} = z_{i2}$	$-\infty < \tau_{i,z_{i2}-1} < \tilde{y}_i, \tilde{\theta}_1, \tilde{\theta}_2 < \tau_{i,z_{i1}} < \infty$	{ 1,2,3 }	{ 2,3,4 }

TABLE 13: Calculations for the nonparametric scales B, and C for all possible relative ordering of a self-assessment, y_i and two anchor ratings z_1 and z_2

Parameters		AF-A Coef	AF-A S.E.	AF-B Coef	AF-B S.E.	MO-C Coef	MO-C S.E.	MO-D Coef	MO-D S.E.
γ_1	Intercept	−1.73	0.07	−1.71	0.08	−1.88	0.11	−1.82	0.09
	Female	0.00	0.09	−0.28	0.10	−0.13	0.12	−0.24	0.10
	Reserved	−0.16	0.11	−0.33	0.13	−0.64	0.24	−0.15	0.13
γ_2	Intercept	1.45	0.07	0.70	0.07	0.36	0.08	0.26	0.06
	Female	0.23	0.09	0.37	0.09	0.04	0.09	0.02	0.07
	Reserved	0.09	0.11	0.19	0.13	0.22	0.21	0.03	0.09
	Vign Last	0.09	0.09	0.20	0.10	−0.17	0.09	−0.05	0.07
γ_3	Intercept	1.21	0.05	1.13	0.05	1.05	0.08	0.62	0.07
	Female	0.13	0.06	0.10	0.06	0.01	0.09	0.18	0.07
	Reserved	0.04	0.07	0.09	0.07	0.24	0.14	−0.01	0.09
	Vign Last	0.11	0.06	0.08	0.06	−0.20	0.09	−0.19	0.07
γ_4	Intercept	1.55	0.13	1.59	0.07	1.79	0.07	1.84	0.07
	Female	−0.15	0.14	−0.11	0.08	−0.04	0.07	0.07	0.07
	Reserved	−0.24	0.15	0.03	0.10	0.07	0.08	0.10	0.08
	Vign Last	−0.40	0.14	−0.27	0.08	−0.15	0.07	−0.37	0.07
θ_L		−0.12	0.09	−0.10	0.10	−0.32	0.12	−0.07	0.10

TABLE 14: Parameter estimates from ordered probit model M1

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: 111–30.

Alwin, D.F., and J.A. Krosnick. 1985. "The Measurement of Values in Surveys: A Comparison of Ratings and Rankings." *The Public Opinion Quarterly* 49 (4): 535–552.

Brady, H.E. 1989. "Factor and ideal point analysis for interpersonally incomparable data." *Psychometrika* 54 (2): 181–202.

Cronbach, L.J. 1946. "Response sets and test validity." *Educational and Psychological Measurement* 6: 475–494.

Cronbach, L.J. 1950. "Further Evidence on Response Sets and Test Design." *Educational and Psychological Measurement* 10 (1): 3.

Cunningham, W.H., I.C.M. Cunningham, and R.T. Green. 1977. "The Ipsative Process to Reduce Response Set Bias." *Public Opinion Quarterly* 41 (3): 379–384.

Flynn, K.E., M.A. Smith, and J. Freese. 2006. "When Do Older Adults Turn to the Internet for Health Information? Findings from the Wisconsin Longitudinal Study." *J Gen Intern Med* 21 (12): 1295–301.

Freese, Jeremy, and Robert M. Hauser. 2006. "WLS MEMO 145: Implementation Of World Health Survey Health Vignettes." Wisconsin Longitudinal Study, January 23, 2006.

Hamilton, DL. 1968. "Personality attributes associated with extreme response style." *Psychol Bull* 69 (3): 192–203.

Javaras, K.N., and B.D. Ripley. 2007. "An 'Unfolding' Latent Variable Model for Likert Attitude Data: Drawing Inferences Adjusted for Response Style." *Journal of the American Statistical Association* 102 (478): 454–463.

Jones, LV, and LL Thurstone. 1955. "The psychophysics of semantics: An experimental investigation." *Journal of Applied Psychology* 39: 31–36.

King, Gary, C.J.L. Murray, J.A. Salomon, and A. Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98 (01): 191–207.

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

Martin, E.A., P.C. Campanelli, and R.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 (3): 265–276.

McGarvey, H.R. 1943. "Anchoring Effects in the Absolute Judgment of Verbal Materials." Ph.D. diss. Columbia University.

Mosier, CI. 1941. "A Psychometric Study of Meaning." *Journal of Social Psychology* 13: 123–140.

Rossi, P.E., Z. Gilula, and G.M. Allenby. 2001. "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach." *Journal of the American Statistical Association* 96 (453).

Sekhon, Jasjeet S. 2008. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software*.

Smith, T.W. 2003. "Developing comparable questions in cross-national surveys." In *Cross-Cultural Survey Methods*, ed. JA Harkness, FJR van der Vijver, and P. Mohler. Wiley.

Torgerson, W.S. 1958. *Theory and methods of scaling*. Wiley New York.

Tourangeau, R., M.P. Couper, and F. Conrad. 2004. "Spacing, Position, and Order Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68 (3): 368–393.

Wand, Jonathan. 2007a. "Credible Comparisons Using Interpersonally Incomparable Data: Ranking self-evaluations relative to anchoring vignettes or other common survey questions." Stanford University.

URL: *http://wand.stanford.edu/anchors/wand_anchors.pdf*

Wand, Jonathan. 2007b. "Stochastic and Latent Class Models of Anchoring Vignettes." Stanford University.

Wand, Jonathan, Gary King, and Olivia Lau. Forthcoming. "Anchors: Software for Anchoring Vignette Data." *Journal of Statistical Software*.

URL: *http://wand.stanford.edu/anchors/*

Ware, J., M. Kosinski, and SD Keller. 1996. "A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity." *Med Care* 34 (3): 220–33.