Identification Strategies in Survey Response Using Vignettes

Luisa Corrado and Melvyn Weeks

May 2010

CWPE 1031

Identification Strategies in Survey Response Using Vignettes

Luisa Corrado*
Faculty of Economics, University of Cambridge and University of Rome Tor Vergata
Melvyn Weeks†
Faculty of Economics and Clare College, University of Cambridge

May 24, 2010

Abstract

In this paper we explore solutions to a particular type of heterogeneity in survey data which is manifest in the presence of individual-specific response scales. We consider this problem in the context of existing evidence on cross-country differences in subjective life satisfaction, and in particular the extent of cross-country comparability. In this instance observed responses are not directly comparable, and inference is compromised.

We utilise two broad identification strategies to account for scale heterogeneity. Keeping the data fixed, we consider a number of estimators based on alternative generalisations of the ordered response model. We also examine a number of alternative approaches based on the use of additional information in the form of responses on one or more additional questions with the same response categories as the self-assessment question. These additional questions, referred to as anchoring vignettes, can under certain conditions, be used to correct for the resultant biases in model parameters.

Keywords: Vignettes, ordered response, generalised ordered response, stochastic thresholds, attitudinal surveys.

*Faculty of Economics, Robinson Building, CB3 9DD, Cambridge. E-mail: lc242@cam.ac.uk.
†Faculty of Economics, Robinson Building, CB3 9DD, Cambridge. E-mail: mw217@cam.ac.uk.

1 Introduction

The use of survey data to study the determination of individual preferences is now firmly rooted in the social sciences. Although the theory and application of revealed preference has been a pivotal component of empirical analysis of individual choice, it is increasingly the case that firms, policy makers, and governments are interested in eliciting preferences over outcomes that are inherently difficult to measure, and in some cases over states of the world yet to be realised. An example of this is the rapidly expanding area of attitudinal research which utilises survey data to capture and represent key aspects of individuals situation. This includes surveys of consumer satisfaction over products and services (see Rossi, Gilula, and Allenby (2001)), surveys of job satisfaction (Kristensen and Johansson (2006)), health (Bago D'Uva, Lindeboom, O'Donnell, and Van Doorslaer (2009); Peracchi and Rossetti (2009); Salomon, Tandon, and Murray (2004)), political efficacy (King, Murray, Salomon, and Tandon (2004)), work disability (Kapteyn, Smith, and Van Soest (2007)), and corruption (Olken (2007)).

A fundamental barrier to inference using survey response is that respondents exhibit variation in the manner in which they utilise ratings scale. This problem has been noted in a number of areas. In marketing Rossi, Gilula, and Allenby (2001) consider the effect of respondents who differ in their use of scale, with, for example, some respondents using only the extreme points on the ratings. Baumgartner and Steenkamp (2001) provide a comprehensive analysis of response styles using data from a large cross country sample, and find a significant effect of response style on the observed responses. King and Wand (2007) emphasise the more general issue of the extent to which self-assessment responses are interpersonal comparable.

The accuracy of the survey information is critical for policy formation. The performance of public services are often compared and used as a tool to promote best practice. In some instances self-assessment surveys are conducted across countries and in this respect the issue of comparability is even more pertinent. For example, in 2001 the WHO launched the World Health Survey, an extensive cross country survey designed to elicit patients views and attitudes across a range of health-care experiences including choice of provider, quality of service, and waiting times. The importance of providing a correct statistical framework to analyse this type of data is paramount since these studies often produce country rankings leading to policy responses which must be properly informed.

In this paper we explore solutions to a particular type of heterogeneity in survey data which is manifest in the presence of individual-specific response scales. We consider this problem in the context of the existing evidence on the cross-country differences in life satisfaction. Previous studies on life satisfaction such as Kapteyn, Smith, and Van Soest

(2009) have focussed on a small number of pre-selected countries that are believed to be comparable. In this study we extend the analysis to a larger set and focus on the extent of cross-country comparability in terms of life satisfaction. Over the last twenty years the growing empirical evidence on the determinants[1] of life-satisfaction has fostered a debate on how to make comparisons across countries (Diener (2006), Kahneman, Krueger, Schkade, Schwarz, and Stone (2004)). Although across many studies Denmark and the Scandinavian Countries have persistently received high ranks (Inglehart and Klingemann (2000)), there remains some doubt as to what extent these rankings depend upon true variation in life satisfaction or are confounded by simultaneous cross-country variation in response scales. In the presence of scale heterogeneity the mapping from the underlying latent variable, say y^*, to the ordinal response may differ across respondents. Put differently, if we fix y^* across two individuals, but response scale are different, then observed ratings will differ.

We utilise two broad identification strategies to account for scale heterogeneity. Keeping the data fixed, we consider a number of methods based on alternative generalisations of the ordered response model. We also examine a number of approaches based on the use of additional information. One approach to this problem has been to collect supplementary data in the form of responses on one or more additional questions with the same response categories as the self-assessment question. These additional questions, referred to as anchoring vignettes, can under certain conditions, be used to correct for the resultant biases in model parameters (see King and Wand (2007); King, Murray, Salomon, and Tandon (2004)).

In utilising vignette information to account for scale heterogeneity we consider the identifying assumption of vignette equivalence. The two extreme cases are that (i) individuals from all countries are comparable and (ii) all individuals possess different response scales and it is not possible to engage in meaningful comparisons. Our testing strategy considers the possibility that comparability is located between these two extremes. We test the validity of this approach and in doing so construct groups of countries which are comparable.

The paper is organised as follows. In section 2 we consider the general problem of heterogeneity in survey response and position the vignette approach alongside other methodologies that have sought to address this problem. In section 3 we introduce the ordered response model in conjunction with a number of generalisations which have been proposed as a means to account for scale heterogeneity. In section 4 we consider generalisations of the ordered response model that are based on additional data. Section 5 introduces the

[1] See, for example, Clark, Frijters, and Shields (2008), Frey and Stutzer (2007).

data and in section 6 we present the results. Section 6 concludes.

2 Heterogeneity and Survey Response

In this section we position the problem of scale heterogeneity within the broader context of inference problems in survey response. One of the most difficult problems is the presence of measurement error. In the most generic sense measurement error represents the difference between the actual value of a quantity and the value obtained by a measurement. In the presence of random error repeating the measurement will improve the situation. If however the measuring instrument is systematically biased then additional measurements will not help. Pudney (2008) notes that the term measurement error does not convey the true nature of the problems with survey data given that social scientists do not actively measure but passively record responses. In this context we can think of the measuring instrument as both the scientist and the respondent. In the case of attitudinal surveys a respondents task might be to measure his own life satisfaction based on an ordinal scale and a tolerance or threshold parameter. If these thresholds are individual-specific then parameter estimates representing the impact of specific determinants of life satisfaction are likely to be biased if the model specification imposes fixed thresholds. Bound, Brown, and Mathiowetz (2001) provide an excellent survey of measurement error in survey data. The authors consider measurement error in household survey of health related variables, making a distinction between generally more continuous variables such as health care expenditure and utilisation, and the self-reporting of health related conditions.

Our point of departure is the notion of inter-individual comparability. As an example, one might postulate that individuals from different cultures and languages may perceive and respond to questions on self-assessment in different ways. Such response heterogeneity, if not accounted for, may confound inter-individual comparisons. A potential solution to this problem is the inclusion of fixed effects, as a way to account for unobserved heterogeneity at the country level. However, the problem with this approach is that these fixed effects may both determine variation in levels of the particular self-assessment, as well as differences across individuals in the manner in which a given level is reported.[2]

In considering alternative approaches to account for this form of heterogeneity, one of the problems is that the literature is fragmented across a number of rather disparate areas. In econometrics, measurement error considered within the classical errors-in-variables framework, and characterised by a fully observed continuous dependent variable with mea-

[2]A number of studies have attested that individuals may still use different scales when reporting well-being even if they live in the same country (see, for example, Van Praag (1971), and Ferrer-I-Carbonell and Frijters (2004)).

surement error affecting one or more explanatory variables, will generate biased and inconsistent parameter estimates, with a general tendency towards attenuation. Kreider (1999) discusses the problem of measurement error for self-reported health and in particular work disability in the context of models of labour force participation. However, the focus here is the impact of likely overreporting of disability on parameter estimates associated with one or more explanatory variables.

In the context of attitudinal surveys where observed responses are often discrete, the disjunction between what is observed and the underlying latent construct is nonlinear and mediated by an observational rule. In this context it is important to separate two behaviourial processes: the actual behaviour the analyst is seeking to *measure*, and the response behaviour of those individuals who provide answers to survey questions. Since a covariate of interest may affect both processes, the critical question is whether the observed response and attendant data, in combination with a particular estimator, is sufficient to separate these two processes.

Measurement error in the dependent variable in a discrete choice setting is generally understood as arising from an error in either the recording or reporting of a response. This may occur simply because the respondent misunderstands the question. For example, models of employment tenure depend upon responses to questions on whether an individual changed jobs over a given period. Measurement error in this context is manifest as a transposition of the integer response, and is therefore synonymous with misclassification. In this instance the true discrete response is recorded with error, generating a series of false positive and false negatives. Hausman, Abrevaya, and Scott-Morton (1998) propose a modified maximum likelihood estimator where the probability of misclassification depends on the value of the true response, say $\tilde{y}_i$, namely $\tau_0 = \Pr(y_i = 1|\tilde{y}_i = 0)$ and $\tau_1 = \Pr(y_i = 0|\tilde{y}_i = 1)$. As the authors state, considering the misclassification model in this context, we observe that the model is indistinguishable from a standard binary choice model with heterogeneity in the response process over 3 types of individuals. For type 1 individuals τ_0 represent the fraction of individuals who always respond with a one independent of the observed $\mathbf{x}_i$, whilst for type 2 individuals τ_1 denotes the fraction of individuals who always respond with a zero. The remaining fraction (type 3) represent those individuals whose behaviour is consistent with the standard binary choice model. Estimates of parameters τ_0 and τ_1 are then used to account for a problem of heterogeneity rather than misclassification.

Seen in this light it is instructive to consider the convolution of scale heterogeneity and variation in y^* as a misclassification problem. However, as demonstrated below, given that our focus is on data generated by attitudinal surveys, and in particular a set of ordinal responses over J choices, the process of deconvolution is more difficult unless restrictive

identification assumptions are imposed. As an example, we consider the distribution of reported outcomes $y_i = \{j\}$ over a population of individuals (indexed by i) as determined by two components: y_i^* and $\boldsymbol{\alpha}_i = \{\alpha_{ij}\}$, the former denoting a true unobserved objective measure, with $\boldsymbol{\alpha}_i$ denoting a vector of individual-specific threshold parameters. Assuming at the outset that these are constant across individuals, namely $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$, the observational rule or response process, provides a mapping from y_i^* to y_i, namely

$$y_i = \sum_{j=1}^{J+1} \mathbf{1}(\alpha_{j-1} < y_i^* < \alpha_j) \times j, \tag{1}$$

where $\mathbf{1}(.)$ denotes the indicator function. α_0 and α_{J+1} are set at $-\infty$ and $+\infty$ respectively. Without loss of generality we assume that y_i^* is additive in a linear index $\mathbf{x}_i'\boldsymbol{\beta}$ and an error ε_i. We now introduce individual-specific response scales, $\xi_{ij} = g_j(\alpha_j, I_{ij}, \omega_{ij})$ where $g_j(.)$ denotes a transformation of the following arguments: $I_{ij} = \mathbf{v}_i'\boldsymbol{\gamma}_j$ is a linear index dependent on observables $\mathbf{v}_i$, α_j is a threshold constant, and ω_{ij} represent unobservables. The mapping is then given by

$$y_i = \sum_{j=1}^{J+1} \mathbf{1}(\xi_{ij-1} < y_i^* < \xi_{ij}) \times j. \tag{2}$$

The fundamental difference between the two observational rules (1) and (2) is that in the former the presence of threshold constants imposes a rather innocuous identification constraint in that $\mathbf{x}_i$ cannot contain a constant. However, in (2) we observe the potential identification problem that may confound inference. If thresholds ξ_{ij} vary as a function of observed covariates $\mathbf{x}_i$, then given the ordered response model contains a single index, we see that inference on parameters $\boldsymbol{\beta}$ may be compromised given the convolution of variation in thresholds $g_j(\alpha_j, I_{ij}, \omega_{ij})$ and $\mathbf{x}_i'\boldsymbol{\beta}$.

In the case of misclassification in binary response, and considering the case where the true response is zero, the identification problem is to separate $\Pr(y_i = 1|\tilde{y}_i = 0)$ from $\Pr(y_i = 0|\tilde{y}_i = 0)$, namely to allow for heterogeneity in response by estimating the fraction of false positives. Although similar to the problem of scale heterogeneity in ordered responses in that we wish to separate $\Pr(y_i = j|\xi_{ij})$ from $\Pr(y_i = j|\mathbf{x}_i'\boldsymbol{\beta})$, there are a number of fundamental differences. First, in the case of binary response the misclassification problem is naturally bounded with errors manifest as either false positives or false negatives. In the case of either ordered or multinomial response, the possible number of errors is $J(J-1)$, such that this type of correction becomes difficult to implement when J is large. Second, the proposed solution to the missclassification

problem considered by Hausman, Abrevaya, and Scott-Morton (1998) and extended to the multinomial case by Ramalho (2001), relies on the identification assumption of conditional independence

$$\Pr(Y = y|\tilde{Y} = \tilde{y}, \mathbf{x}) = \Pr(Y = y|\tilde{Y} = \tilde{y}) \tag{3}$$

namely that conditional on the true unobserved response, the reported outcome is independent of the individual characteristics $\mathbf{x}_i$. In our case we wish to allow the possibility that misclassification depends upon both observed and unobserved heterogeneity at the level of the individual[3]

In section 3 we introduce the ordered response model and focus upon a number of generalisations that have been proposed. These generalisations take the data as given and are based upon alternative estimators. In section 4 we examine a number of additional generalisations where the focus is in supplementing the information set.

3 Generalised Ordered Response

To consider a class of generalised ordered response models we present the following canonical model

$$\mathbf{y}_i^* \sim \Psi(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*), \tag{4}$$

where $\mathbf{y}_i^* = \{y_{im}^*\}$ denotes a $M \times 1$ vector of latent variables, and $\mathbf{y}_i = \{y_{im}\}$ denotes a $M \times 1$ vector of responses with each element $y_{im} \in (1, J)$. $\Psi(.)$ denotes a multivariate distribution with mean $\boldsymbol{\mu}_i^*$ and covariance matrix $\boldsymbol{\Sigma}_i^*$.[4] The observational rule for the m^{th} element of $\mathbf{y}_i$ is given by

$$y_{im} \quad = \quad 1(\xi_{j-1,im} < y_{im}^* < \xi_{j-1,im}) \times j, \quad j = 1, ..., J.$$

For $M > 1$ represents cases where respondents provide information on multiple assessments, either in the form of panel or a multiple question survey.

Our point of departure is the benchmark ordered response (OR) model, based on a single self-assessment ($M = 1$). This model is a single index model, characterised by an unobserved latent variable y_i^* which is generally assumed additive in a linear index $\mathbf{x}_i'\boldsymbol{\beta}$ and an error term ε_i. The mapping from y_i^* to an observed response is given by the observational rule (1). Identification of mean equation parameters $\boldsymbol{\beta}$ is achieved by

[3] Krieder and Pepper (2008) consider the problem of making inference on disability using potentially corrupt self-assessment data. In introducing a classification error model, the authors utilise a nonparametric bounding methodology and demonstrate the price of various identification assumptions.

[4] A common distributional assumption used in these models is normality. For $M > 1$ this gives rise to the multivariate ordered probit model.

assuming scale homogeneity in the form of constant threshold parameters. Standard location conditions can be achieved by excluding a constant from $\mathbf{x}_i$.

There exist a large number of generalisations of the benchmark ordered response model,[5] including Pudney and Shields (2000), finite mixtures models (see Eluru, Bhat, and Hensher (2008), Greene and Hensher (2010)), and generalised thresholds which depend upon both observables and unobservables (Cunha, Heckman, and Navarro (2007)). In the case of the latter generalisation the authors discuss the economic foundations of ordered *choice* models, and present a useful discussion of stochastic threshold models. Below we consider a number of these extensions, with particular emphasis on models that account for scale heterogeneity.

3.1 Generalised Threshold Models

A common extension of the benchmark ordered response model (1) accommodates scale heterogeneity by allowing thresholds to vary across individuals due to observables. Deterministic threshold models were first proposed by Maddala (1983) and Terza (1985). An immediate obstacle to this generalisation is the single linear index characteristic of the ordered response model. For example, if we let individual thresholds be a linear deterministic function of $\mathbf{x}_i$, say $\xi_{ij} = \alpha_j + \mathbf{x}_i'\boldsymbol{\gamma}_j$, where $\boldsymbol{\gamma}_j$ is a $K \times 1$ vector of parameters, then the identification problem stems from the convolution of two linear indexes. To see this we rewrite $\mathcal{P}_{ij} = \Pr(y_i = j)$ as

$$\mathcal{P}_{ij} = F(\alpha_j - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\gamma}_j)) - F(\alpha_{j-1} - \mathbf{x}'(\boldsymbol{\beta} - \boldsymbol{\gamma}_{j-1})), \tag{5}$$

where $F(.)$ denotes the distribution function. From (5) we observe that this generalised (linear) threshold model is not separately identified from a heterogenous parameter model, namely $\boldsymbol{\theta}_j = \boldsymbol{\beta} - \boldsymbol{\gamma}_j$.

To see the ramifications of this problem in the context of cross-country comparability of self-assessed life satisfaction, we consider the identification of country effects. Letting C denote the number of countries, such effects might be estimated by a $(C-1)\times 1$ parameter vector $\boldsymbol{\beta}_C \subset \boldsymbol{\beta}$. Respondents living in different countries may have systematically different true levels of satisfaction and, in addition, the scale on which the level of life satisfaction are reported may differ by country. These effects can be estimated by a parameter vector $\boldsymbol{\gamma}_{jC} \subset \boldsymbol{\gamma}_j$. However, in the context of the linear index model (as in (5)), the information from self-reported assessments is not sufficient to separately identify variation in true levels which may be attributed to country effects ($\boldsymbol{\beta}_C$) from scale heterogeneity ($\boldsymbol{\gamma}_{jC}$).

[5] Excellent coverage of a number of generalisations are to be found in Greene and Hensher (2010) and Boes and Winkelmann (2005).

In order to separately identify these two effects additional information is required. To consider these approaches, using the notation introduced in section 2, we first write the probability for choosing j as

$$\mathcal{P}_{ij} = F(g(\alpha_j, I_{ij}, \omega_{ij}) - \mathbf{x}_i'\boldsymbol{\beta}) - F(g(\alpha_{j-1}, I_{ij-1}, \omega_{ij-1}) - \mathbf{x}_i'\boldsymbol{\beta}). \tag{6}$$

Using this canonical representation below we consider a number of approaches to identification.

3.1.1 Identification Strategies

It is important to position the identification strategy in the context of the objectives of the study. In the strategies outlined below our focus is upon instances where the analyst observes both responses on one or more ordinal responses, alongside a set of individual characteristics ($\mathbf{x}_i$). The objective is then to control for scale heterogeneity, utilising all or some of the observed heterogeneity, in order to make valid inference on the impact of observed characteristics on the observed response.

Pudney and Shields (2000) maintain a linear index specification but achieve identification of mean parameters $\boldsymbol{\beta}$ by partitioning $\mathbf{x}_i$ into possibly overlapping subsets. Letting $\mathbf{x}_i^M \subset \mathbf{x}_i$ and $\mathbf{v}_i^T \subset \mathbf{x}_i$ denote the covariates used in the mean (M) and the threshold (T), the identification condition is that each partition contain at least one unique variable. Although the resulting set of zero restrictions generates identification without the need to introduce a nonlinear transformation, there are a number of problems with this approach. First, it is not possible to estimate the impact of any given covariate on both y^* and the response process.[6] Second, model probabilities $\mathcal{P}_{ij}$ for ordered response models, as evident from (6), are calculated as the difference between two distribution functions, evaluated at their respective arguments. As a consequence response models based on a linear index threshold model cannot guarantee that model probabilities will be positive.

Letting $\mathbf{v}_i = \mathbf{x}_i$ and $\xi_{ij} = \texttt{exp}(\alpha_j + \mathbf{x}_i'\boldsymbol{\gamma}_j)$ then we have the nonlinear threshold specification. The presence of the same set of covariates in the threshold component as in the mean equation dictates that identification is achieved through functional form. Although this route to identification has met disapproval in the empirical economics literature, Greene and Hensher (2010) make a useful argument, noting that there is no underlying theory which dictates the linearity of the index for either the conditional mean or the thresholds.

A mixed generalised ordered response model represents a further generalisation of

[6] See Pudney and Shields (2000) for an insightful discussion of circumstances when this does not represent a constraint.

the ordered response model. The most general form of this model allows parameters in both the mean and threshold component to vary across individuals (see Greene and Hensher (2010)). Other recent examples of this generalisation is the stochastic threshold model considered by Cunha, Heckman, and Navarro (2007), and Eluru, Bhat, and Hensher (2008), where thresholds are now allowed to depend on both observables and unobservable. In this instance we write

$$\xi_{ij} = \texttt{exp}(\alpha_j + \mathbf{x}_i'\boldsymbol{\gamma}_j + \sigma_{\xi_j}\omega_{ij}) \tag{7}$$

Threshold parameters are now given by threshold constants α_j, threshold standard deviations σ_{ξ_j}, alongside the effects of observables $\boldsymbol{\gamma}_j$. A standard specification of the random individual effects is based on $\omega_{ij} \sim N(0,1)$.

In some cases the analyst may not observe covariate information in the form of $\mathbf{x}_i$. For example, Rossi, Gilula, and Allenby (2001), control for scale heterogeneity utilising a Bayesian Hierarchical approach. Relative to the approaches considered above, a notable feature is that identification is dependent upon observing, for each respondent, a $1 \times M$ vector of ordinal responses $\mathbf{y}_i$. A typical format of this type of data might be that the first question is designed to elicit a general preference a given product, with subsequent questions focussing on product attributes.

The specification of the vector of latent variables is given by

$$\mathbf{y}_i^* = \boldsymbol{\mu} + \tau_i\boldsymbol{\iota} + \sigma_i\mathbf{z}_i, \tag{8}$$

where $\boldsymbol{\iota}$ is $M \times 1$ unit vector. $\boldsymbol{\mu}_i^* = \boldsymbol{\mu} + \tau_i\boldsymbol{\iota}$ is a $M \times 1$ vector of means and $\boldsymbol{\Sigma}_i^* = \sigma_i\boldsymbol{\Sigma}$ is the $M \times M$ covariance matrix. (8) allows for a respondent-specific location (τ_i) and scale (σ_i) parameters, fixed across M. It is instructive to note that although heterogeneity is introduced into the specification of the vector of latent variable $\mathbf{y}_i^*$, with the threshold parameters fixed, an alternative (observationally equivalent) approach would be to fix $\mathbf{y}_i^*$ and introduce heterogeneity through the threshold parameters. One such specification would be to recast (8), with $\mathbf{z}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, as a model with respondent specific thresholds $\boldsymbol{\alpha}_i = \tau_i + \sigma_i\boldsymbol{\alpha}$. The authors point out that the chosen route to identification is dependent on observing multiple responses per respondent, and given that this is generally not large, a parsimonious specification is required. The two primary objectives of the study are also important considerations here. The authors focus upon the measurement of the relationship between product satisfaction and preferences for product attributes, and the identification of individual-specific preference parameters. In this instance the set of parameters that represent scale heterogeneity are of interest in themselves, and this may be done either through location and scale shifts introduced at the level of the latent variables or the threshold parameters. An interesting extension of this approach would be to utilise

this method to facilitate valid inference on the impact of covariates on the observed set of responses.

4 Incorporating Additional Information

The identification strategies considered above have taken as given the available data, and sort to overcome the identification problem through a combination of partitioning $\mathbf{x}_i$, functional form, or alternate estimators based upon stochastic threshold specification. In the spirit of the standard instrumental variables estimators, an alternative route to identification is to use additional data. An approach which has found prominence in a number of disciplines, accounts for measurement error by utilising additional survey response information.[7] Kotlarski (1967) exploits the classical properties of the measurement errors, and demonstrates that in all cases it is better to use two or more noisy measures rather than a single, more precisely defined measure. Browning and Crossley (2009) extend this result, relaxing the classical assumptions.

In the context of attitudinal surveys, two types of additional information have been used to circumvent the problem of scale heterogeneity. One approach has been to locate one or more indicators of the latent construct y^* that is not subject to reporting heterogeneity (see, for example, Bago D'Uva, Lindeboom, O'Donnell, and Van Doorslaer (2009) and Van Soest, Delaney, Harmon, Kapteyn, and Smith (2007)). Bound, Brown, and Mathiowetz (2001) review a number of studies of self-report, emphasising the role of validation sources. An alternative strategy which has been predominantly used in the political and health sciences, utilises additional survey responses on a hypothetical situation, a so-called vignette, that is fixed for all respondents. The '*repeated measures*' in this context comprise a set of responses on the self-assessment survey, and responses across a set of vignettes. If all individuals perceive the description in the vignette in the same way then any systematic variation in answers to vignettes can be attributed to scale heterogeneity.

The vignette approach is based on the following sub-model

$$z_{il}^* = \theta_l + \sigma_\omega \omega_{il}, \quad l = 1, ..., L \tag{9}$$

$$z_{il} = \sum_{j=1}^{J+1} \mathbf{1}(\xi_{ij-1} \leq z_{il}^* \leq \xi_{ij}) \times j, \tag{10}$$

where z_{il}^* is the unobserved latent variable corresponding to vignette l. The fundamental premiss of the vignette methodology is that there exists a true (objective) unobserved

[7] Examples of this approach include Li and Vuong (1998), Schennach (2004), and Delaigle, Hall, and Meister (2008).

level of the latent variable, which apart from iid sampling error (ω_{il}) is constant across individuals (θ_l). The observational rule in (10) is assumed to be the same as for the self-assessment model (2), with the thresholds determined by the same set of explanatory variables. The self-assessment component is the standard latent variable regression with linear index $\mathbf{x}_i'\boldsymbol{\beta}$ with the following threshold specification

$$\xi_{ij} \quad = \quad \xi_{ij-1} + \exp(\alpha_j + \mathbf{x}_i'\boldsymbol{\gamma}_j), \qquad j = 2, ..., J-1, \tag{11}$$

where $\xi_{i1} = \alpha_1 + \mathbf{x}_i'\boldsymbol{\gamma}_1$. This model, which we refer to as Hopit[8] with vignettes is comprised of a self-assessment component given by (2) and normal errors, a threshold specification given by (11), and a vignette component given by equations (9) and (10). This specification facilitates the identification of scale heterogeneity and the requisite ordering of thresholds.

The two critical identifying assumptions are (i) *vignette equivalence* and (ii) *response consistency*. The first assumption requires that the description of the vignette is perceived to correspond to the same state by all respondents. Response consistency requires that individuals use the response category in the self-assessment question in the same way when they evaluate hypothetical scenarios in the vignettes. The identifying assumption of vignette equivalence implies that any systematic differences in observed responses z_i can be attributed to scale heterogeneity.

Although in utilising this approach scale heterogeneity is identified with the use of a single vignette, there are gains to observing multiple vignette responses for each individual. To see this assume that the design of a single vignette depicts a situation at the *lower* end of the distribution for life satisfaction, with the self-assessed responses clustered in the *upper* end of the distribution. Since this vignette depicts a situation at the lower end of the distribution, this provides information to correct for scale heterogeneity over self-assessment responses located at this point. If we now include responses on an additional vignette, and assume that these responses span other parts of the distribution, then we have additional information to capture the full extent of scale heterogeneity. Note that the additional information used by Rossi, Gilula, and Allenby (2001) is in the form of multiple responses, which facilitates the identification of respondent-specific location and scale parameters. In this context a large location parameter, τ_i, would indicate overuse of a particular value while a large scale parameter, σ_i, would indicate the presence of extreme response styles. One limitation of identifying scale heterogeneity through a set of individual specific scale and location parameters, is that this approach cannot, for example, accommodate a situation where individual response heterogeneity is polarised across two extremes of the distribution.

[8]The use of the acronym Hopit, hierarchical ordered probit, follows that of Greene and Hensher (2010).

Combining the self-assessment and vignette components, the likelihood function for the Hopit model can be written as[9]

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{z}) = L_y(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}) \times L_z(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathbf{z}), \tag{12}$$

where $L_y(.)$ and $L_z(.)$ denote, respectively, the likelihood functions for the two sub-models. $\boldsymbol{\beta}$ is a $K \times 1$ vector of mean parameters, $\boldsymbol{\theta}$ is a $L \times 1$ vector of vignettes constants and $\boldsymbol{\gamma}$ is a $(J \times K) \times 1$ vector of threshold parameters.[10] In most cases the parameters of interest are $\boldsymbol{\beta}$.

The likelihood for the self-assessment component $L_y(.)$ is given by

$$L_y(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}) \propto \mathbf{\Pi}_{i=1}^{N} \mathbf{\Pi}_{j=1}^{J} [F(\xi_{ij} - \mathbf{x}_i' \boldsymbol{\beta}) - F(\xi_{ij} - \mathbf{x}_i' \boldsymbol{\beta})]^{\mathbf{1}(y_i = j)},$$

and the likelihood for the vignette component $L_z(.)$ is given by

$$L_z(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathbf{z}) \propto \mathbf{\Pi}_{l=1}^{L} \mathbf{\Pi}_{i=1}^{N} \mathbf{\Pi}_{j=1}^{J} [F(\xi_{ij} - \boldsymbol{\theta}_l) - F(\xi_{ij-1} - \boldsymbol{\theta}_l)]^{\mathbf{1}(z_{il} = j)}.$$

Note that i indexes individuals that provide both self-assessment responses and also responses to the vignettes. Since the variance of the self-assessment sub-model is normalised to one, we are able to identify the variance of the latent variable in the vignette model, σ_ω.

Examining $L_y(\boldsymbol{\beta}, \boldsymbol{\gamma})$ we observe that the likelihood depends upon both mean and scale parameters, and there lies the potential identification problem. Since the identification assumption of the vignette approach depends upon vignette equivalence, thereby precluding the inclusion of $\mathbf{x}_i$, we see that the accompanying assumption of response consistency, facilities the identification of $\boldsymbol{\gamma}$ given that these parameters are present in the vignette likelihood $L_z(\boldsymbol{\theta}, \boldsymbol{\gamma})$.

4.1 Testing

A number of tests have been designed to test the validity of the identifying assumptions of vignette equivalence and response consistency. Bago D'Uva, Lindeboom, O'Donnell, and Van Doorslaer (2009) test both assumptions in a study of cognitive functioning and mobility related health problems in the UK. An important caveat here is that without additional information it is not possible to test for response consistency. This occurs given that with the vignette sub-model we have a just-identified model. Such a test is feasible

[9] King, Murray, Salomon, and Tandon (2004) refer to this as a compound hierarchical model.

[10] Note that the dimension of $\boldsymbol{\gamma}$ corresponds to the most general model. More parsimonious versions, for example, restricting all threshold parameters to be equal, are possible.

when the analyst has additional objective indicators of the latent construct which are believed to be free of reporting error. With this information the identification assumption is that any systematic variation in assessments that exists after conditioning on objective indicators can be considered scale heterogeneity. As a result, we can then think of the model as overidentified, and response consistency can then be tested.

As Kapteyn, Smith, and Van Soest (2009) note, it is extremely difficult to locate objective indicators for life satisfaction. Therefore, in this study we focus on testing vignette equivalence based upon whether respondents in all countries in our sample perceive the same "true" value of the vignette. We identify countries for which vignette equivalence holds and use this to locate country groups with the same underlying interpretation of the vignette. Put another way our testing strategy is designed to identify the extent to which respondents are comparable across countries. To operationalise our test we first rewrite (9)

$$z^*_{i1} = \theta_1 + \sigma_\omega \omega_{i1} \tag{13}$$

$$z^*_{il} = \theta_l + \lambda_r(\theta_l D_{ic}) + \sigma_\omega \omega_{il} \quad c = 1, ..., C-1 \quad l = 2, ..., L \tag{14}$$

where $D_{ic} = 1$ for individual i resident in country c. The two extreme cases are that all/none of the observations in the sample are comparable. However, the most likely situation is that individual responses across certain groups of countries are comparable.

In summary we estimate a joint model with a likelihood given by (12), and replace the vignette equations (9) and (10) with equations (13) and (14).[11] Non zero elements in $\boldsymbol{\lambda} = \{\lambda_r\}$ indicate systematic variation in the perception of a vignette relative to a reference country thereby invalidating the vignette equivalence assumption, implying that for one or more countries we cannot identify reporting heterogeneity using the information from the vignettes. Those countries for which we cannot reject the null form a group, such that within this partition comparability is based on individuals sharing the same interpretation of the vignette.

5 Data

We use data from the second wave (2006) of the Survey of Health, Ageing and Retirement in Europe (SHARE)[12] which provides information on health, psycho-cognitive ability, socio-economic status and social support for individuals aged 50+ living in private Eu-

[11] The first vignette acts as reference and cannot depend on the same country dummies as the second vignette as the model would be unidentified (see Bago D'Uva, Lindeboom, O'Donnell, and Van Doorslaer (2009)).

[12] see SHARE (2009).

ropean households. A separate project COMPARE utilises a random sample from the SHARE sample, and includes vignettes on health, well-being, job satisfaction and work disability for eleven countries covering Scandinavia (Denmark and Sweden) Continental Europe (France, Belgium, Germany, Poland, Netherlands, Czech Republic) and Southern Europe (Italy, Greece and Spain). In our study we restrict attention to respondents for whom the vignette information is available.[13]

Table 1 describes the variables used in the analysis. We consider traditional economic factors such as household income, household size plus a set of socio-demographic controls represented by gender, age and numbers of year in education. A number of existing studies (Helliwell and Putnam (2005), Lyubomirsky, Sheldon, and Schkade (2005)) have noted that economic factors account for only about 10% of the variation in life satisfaction across individuals, emphasising the influence of non-economic factors, such as being married or being in a stable relationship, being active in the community and helping others. Personal health and education are other determinants.

As we can see from Table 2 our vignette sample equally represents women and men, with respondents, on average aged 65. In the vignette sample 67% of respondents live with their partner, and in 2% of the cases the partner never worked. Respondents seem to prefer socialising activity (22%) and voluntary work (15%) to other forms of community involvement such as educational (8%) or political activity (5%). Finally 11% are part of a religious organisation. Since the use of self-reported health is prone to endogeneity with respect to self-reported life satisfaction, we include in the analysis more objective measures of respondents' health represented by the number of chronic diseases, a measure of depression and whether the respondent has any limitation in daily activities. Respondents rate their depression on average at 2.25 on a scale between 0-11, 44% declare to have limitation in daily activity and have at least one chronic disease.

5.1 Cross-Country Comparisons using Vignettes

Vignettes have been employed in a growing number of surveys including the English Longitudinal Study of Ageing (ELSA), the World Health Surveys (WHS) and the Survey of Health, Ageing and Retirement in Europe (SHARE). Table 3 presents a number of key studies in the vignette literature. Kapteyn, Smith, and Van Soest (2007) utilise a vignette approach to disentangle true differences in work disability from potential differences in response scales. Given that this is done for two countries, the United States and the Netherlands, the imposition of vignette equivalence at the country level is less restric-

[13]Two types of vignettes were randomly assigned to the respondents: type A for respondents younger than 65 and type B for respondents 65 years and older. They differ with regard to question order and gender of the people described in the statements.

tive.[14] Kapteyn, Smith, and Van Soest (2009) have recently extended this approach to analyse the determinants of life satisfaction for the same pair of countries; they find that after correcting for differences in response scales, the conclusion that the Dutch are more satisfied with their lives remains valid.

King, Murray, Salomon, and Tandon (2004) have applied the vignette methodology to political efficacy in another two country study. They find that without accounting for response scale heterogeneity Chinese seem to have more political influence than the Mexicans. However, after controlling for scale heterogeneity the conclusion reverses. In a within country study Delaney, Harmon, Smith, and Van Soest (2007) utilise self-assessment and vignettes in a survey on drinking behavior among students at a major university in Ireland. The self-assessment and vignette responses are then combined in a joint estimation to identify the varying thresholds.

Kristensen and Johansson (2006) utilise a similar approach to reconsider the empirical regularity that in cross-country studies certain countries are persistently ranked high with respect to a number of measures, including job satisfaction (see, for example, Blanchflower and Oswald (1999)). In particular the authors examine the extent to which cross-country differences in reported job satisfaction may be attributed to scale heterogeneity. The identifying assumption is that vignette equivalence holds across seven EU countries.

In this study we utilise vignettes as identification instruments in comparing life satisfaction across the eleven European countries covered by the SHARE data. For life satisfaction respondents are first asked to rate the following question: "How satisfied are you with your life in general?". The self-assessment question is rated according to the scale: "Very Dissatisfied", Dissatisfied", "Neither Satisfied not Dissatisfied", "Satisfied", "Very Satisfied". Respondents are then faced with the following two anchoring vignettes:

- John is 63 years old. His wife died 2 years ago and he still spends a lot of time thinking about her. He has 4 children and 10 grandchildren who visit him regularly. John can make ends meet but has no money for extras such as expensive gifts to his grandchildren. He has had to stop working recently due to heart problems. He gets tired easily. Otherwise, he has no serious health conditions.

- Carry is 72 years old and a widow. Her total after tax income is about € 1,100 per month. She owns the house she lives in and has a large circle of friends. She plays bridge twice a week and goes on vacation regularly with some friends. Lately she has been suffering from arthritis, which makes working in the house and garden

[14]Given that Kapteyn, Smith, and Van Soest (2007) compare work disability across two countries, the authors also examine the extent to which variation in how respondents translate a true disability level into a reported indicator affects inference on disability within a country.

painful.

Respondents are asked to rate the level of life satisfaction of the two hypothetical individuals described in the above vignettes using the same scale as the self-assessment question. Figure 1 and Table 4 report the distribution across countries of the respondent's life satisfaction for the self-assessment and the two vignettes scored on an ordinal scale between one and five. As shown in Table 2 the average for self-reported life satisfaction across countries is around 4, for the first vignette is 2.6 and for the second vignette 3.52. Looking at the distribution of self-reported life satisfaction in Table 4 we note that in Italy and the Czech Republic only 7% of the respondents are "Very Satisfied" with their life while in Denmark this percentage is 41% and in Sweden is 31%. It is also interesting to note that in Italy and Greece around 13% of respondents class themselves as "Very Dissatisfied" or "Dissatisfied", whereas in the Scandinavian Countries this percentage drops to 2% for Denmark and Sweden and 1% for Netherlands.

This data suggests there may exist scale heterogeneity in our sample based upon observed differences in the way respondents rate the level of life-satisfaction for any given vignette. For example, when rating the level of satisfaction of John (vignette 1), only 1% of Danish rate him as "Very Dissatisfied" while in Italy the percentage is 13%. Assuming vignette equivalence this variation can be used to control for scale heterogeneity when undertaking self-assessment.

6 Results

In this section we present the results of the benchmark ordered probit model and compare it with two versions of the generalised ordered probit model given by deterministic and stochastic thresholds. We also present the results of the Hopit model which combines a self-assessment and vignette model component. In line with the main focus of the paper in Table 8 we provide cross-country rankings of life satisfaction for each specification and assess their sensitivity to model specification. In testing for vignette equivalence, we identify groups of countries which are comparable in terms of the interpretation of the vignette.

6.1 Ordered Probit and Generalised Ordered Probit

Table 5 presents the results for the ordered and generalised ordered probit models with deterministic and stochastic thresholds.[15] Column two presents the results for the ordered

[15]The estimation of the ordered probit and generalised ordered probit uses LIMDEP (Greene (2007). The deterministic version of generalised ordered probit model is estimated by maximum likelihood while

probit model. We observe that life satisfaction is higher for women and is weakly affected by the numbers of years in education. Life satisfaction increases with household income (given household size) and falls with any health limitation and with the number of chronic disease and depression. Respondents who are currently living with their partner or spouse report higher levels of life satisfaction relative to those who are alone. Finally, any form of community involvement exerts a positive effect on life satisfaction, particularly if the respondent is active in politics, religion, social activity or volunteering.

Column three of Table 5 presents the results for the generalised ordered probit with deterministic thresholds.[16] Although our specification allows thresholds to vary as a function of the same set of covariates as in the mean equation, for the sake of expediency we choose to present the parameter estimates for a particular threshold, namely $(\hat{\gamma}_4)$, which represents the category "Very Satisfied". The observed differences in the mean equation parameters relative to ordered probit, indicate that the threshold specification is accounting for some degree of heterogeneity. However we have few priors as to how accounting for scale heterogeneity in this way might affect the mean parameters with respect to the ordered probit results. For example, for the deterministic threshold model we find that both gender and income are now not significant whereas we observe a much larger effect on life satisfaction of community involvement, particularly volunteering and social activity. Column four reports the parameter estimates for the upper threshold. Respondents who receive help, do any voluntary work or are involved in any social activity utilise a higher threshold for the category "Very Satisfied". Note also that few controls are significant especially in the threshold equation suggesting that a more parsimonious model for the threshold equation might be considered.

In the last two columns of Table 5 we present the results of a generalised ordered probit model with stochastic thresholds. In this model the combination of nonlinearities via the functional form and the random thresholds are used to identify response scale heterogeneity. The particular variant of the stochastic threshold model we consider is:

$$\xi_{ij} = \xi_{ij-1} + \texttt{exp}(\alpha_j + \sum_{c=1}^{C-1} D_{ic}\gamma_c + \sigma_{\xi_j}\omega_{ij}) \quad j = 2, ..., J-1, \tag{15}$$

where $\omega_{ij} \sim N(0,1)$. $D_{ic} = 1$ is a dummy variable equal to 1 if individual i is resident in country c; γ_c is the associated country effect. Although the stochastic threshold model represents a useful extension of the ordered response model in accounting for individual

the stochastic version by maximum simulated likelihood.

[16]Note that for all models estimated in LIMDEP the parameters for the first threshold are restricted to be zero. A constant is included in the estimation of the mean equation but is not reported in the table of results.

reporting heterogeneity, we encountered a number of numerical problems when estimating the parameters of this model. Eluru, Bhat, and Hensher (2008) and Greene and Hensher (2010)) report similar estimation problems. In our model the threshold specification includes a threshold constant and standard deviation, together with a full set of country dummies. We impose zero restrictions on the elements of $\boldsymbol{\gamma}$ which are associated with covariates which vary across individuals.[17]

In Table 5 we present estimates for the mean ($\hat{\boldsymbol{\beta}}$), together with the mean and standard deviation of the threshold random effects ($\hat{\alpha}_j$, $\hat{\sigma}_{\xi j}$). Estimates of threshold parameters for the country effects ($\hat{\boldsymbol{\gamma}}_C$) are presented in Table 8. The mean equation parameters of this more parsimonious model are reported in column five of Table 5. As a result of the parameter restrictions we note that in a number of instances parameter estimates are similar to the ordered probit. Given that the stochastic generalised ordered probit model nests the ordered probit we perform a likelihood ratio test and find that this generalised model represents an improvement. There are also differences in the estimation of country dummies in the two models and this is discussed in Table 8.

6.2 Hierarchical Ordered Probit

Table 6 presents the results of the mean equation for the Hopit model.[18] Column two presents the results for the mean equation for all countries, while columns 4-7 present the parameter estimates for groups of countries which are located on the basis of a test of vignette equivalence which is discussed below. We again report the parameter estimates for the threshold category $\hat{\gamma}_4$, "Very Satisfied". The bottom part of Table 6 reports the results for the vignette equation which includes estimates of the mean parameters for the two vignettes, $\hat{\theta}_1$ and $\hat{\theta}_2$, and the standard deviation parameter $\hat{\sigma}_\omega$.

In the results for all countries we observe that relative to males, females are still more likely to report higher levels of life satisfaction with a magnitude similar to the ordered and generalised stochastic ordered probit estimates. However, we again emphasise that we have few priors on the expected direction of the change in the mean equation parameters. We note, however, that age is now significant and household income exhibits a larger effect relative to the ordered and generalised ordered probit models. The level of community involvement, especially volunteering and social activity, exhibit relatively large effects.

To assess whether the Hopit model facilitates comparability across all countries we perform a test for vignette equivalence. To perform the test we maximise the log of the

[17] In extending our work we plan to explore a number of richer specifications of the stochastic generalised ordered probit model incorporating in the threshold equation a subset of the controls.

[18] This model is estimated using conditional maximum likelihood and implemented by the `STATA` module `GLLAMM` - see Rabe-Hesketh and Skrondal (2008).

likelihood given by (12), utilising a variant of the vignette component given by (14). The test results are presented in Table 7. In column two we present the mean parameter estimates for the vignette equation, $\hat{\theta}_1$ and $\hat{\theta}_2$, and the parameter estimates for the country interaction effects, $\hat{\theta}_2 * D_{ic}$ with $c = 1, ..., C-1$. Parameter estimates for $\hat{\theta}_2 * D_{ic}$ which are not significantly different from zero, indicate a group of homogenous countries in the sense that the interpretation of the vignettes are comparable. This group is formed by Germany, the Netherlands, France and Greece. We now restrict the sample to the remaining countries, and employ the same test of vignette equivalence to determine whether there exist additional groupings.[19] As reported in column four, we are able to locate a second group of countries formed by Spain, Italy, Denmark, Belgium and the Czech Republic, with Poland forming a singleton. A likelihood ratio test strongly rejects the model that assumes vignette equivalence for all countries.

Parameter estimates for each subgroup are presented in column four to seven of Table 6. In Group 1 all countries are EU member states while the second includes one accession country (Czech Republic). Previous studies on the determinants of life satisfaction across Europe have found that while economic factors are not as important for the EU-15, especially for the Scandinavian and Continental European countries, they are still important drivers of life satisfaction for the accession countries and the Mediterranean countries (see Aslam and Corrado (2007)). In comparing parameter estimates for the two subgroups we find similar results. For example, marital status exerts a significantly larger effect on life satisfaction in the first group of core European countries. In the second group income and poor health, particularly limitation in daily activities, play a greater role.

6.3 Country Rankings

In Table 8 we report parameter estimates and indicators of significance for country dummies for model specifications presented in the previous sections. Using the parameter estimates for the country dummies we also construct a ranking of countries in terms of reported life satisfaction, and comment on the variation in these rankings across different specifications.

The rankings generated by the ordered probit model (column 2) indicate that life satisfaction is highest for Denmark and lowest for Greece. These results, along with high rankings for Sweden and Netherlands are in line with the findings of a number of previous studies ((Inglehart and Klingemann (2000)). In columns 4 and 6 we report rankings for two generalisations based on, respectively, deterministic and stochastic thresholds. In both cases we note substantial differences in the rankings relative to the benchmark model.

[19] For both tests we utilise a significance level of 10%.

Although these differences demonstrate an impact of the different model specifications, and in particular how rankings on life satisfaction change dependent upon alternative strategies to identify scale heterogeneity, there are a number of factors which limit the inference we can make. First, across all models the ranks are based on parameter estimates, which, in a number of instances are not statistically significant. Related, it is also the case that the testing strategy is not capable of generating a full set of rank order statistics, together with interval estimates for these ranks. For example, if we compare the results for the deterministic and stochastic threshold models, alongside the Hopit results using all countries, we observe that the Netherlands is ranked 2nd and 1st. In addition for the same three models three countries are consistently in the top 5 countries (Spain, Netherlands and Poland), although the estimated ranks differ. Although these results demonstrate a degree of robustness in terms of locating countries which are ranked either high or low, the full ranking exhibits variation dependent upon the identification strategy. Put another way, given that these ranks are statistics with sampling distributions, the data may not be sufficient to generate a full set of order statistics.

A limitation of the generalised ordered probit models considered here is that there exists a maintained assumption that conditional on the use of the respective identification strategies, scale heterogeneity can be accounted for, and reliable inference conducted using the mean equation parameters. When using the Hopit model, the additional vignette information has provided an alternative identification strategy, with an advantage that we can test for cross-country comparability. In testing for vignette equivalence we rejected comparability across the full set of countries, and located two groups of countries which are directly comparable. The rankings for the groups are given in columns 10 and 12. We also find the one country, Poland, appears to interpret the vignette in a different way from the other countries.

7 Conclusion

In this paper we have considered the problem of inference in cross-country surveys of life satisfaction. In particular we have examined the extent to which the impact of country of residence on life satisfaction is confounded by scale heterogeneity. Although our findings suggest that existing models are able to differentiate between high and low ranked countries, our results suggest that the complete rankings for life satisfaction depend on the identification strategy. For the model specification based upon vignettes, the rejection of comparability across all countries provides a question over the design of the vignette. Increasingly social scientist have become actively involved in the design of survey questionnaires, including, for example, consideration of possible instruments in anticipation of

endogeneity problems. The use of vignettes in conjunction with self-assessment responses represents a similar development, and in this regard there may be scope for pilot studies to explore the issue of vignette equivalence. Existing work by one of the authors (Weeks (2010)) has considered a related set of problems encountered when ranking stochastically ordered distributions. We are currently developing these methods within a Bayesian framework and extending to the present analysis. In addition given that the primary focus of much of the extant literature is on point identification, we will consider the extent to which methods based upon partial identification may be of use.

References

Aslam, A., and L. Corrado (2007): "No Man is An Island: The Inter-personal Determinants of Regional Well-Being in Europe," Cambridge Working Papers in Economics 0717, Faculty of Economics, University of Cambridge.

Bago D'Uva, T., M. Lindeboom, O. O'Donnell, and E. Van Doorslaer (2009): "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity," HEDG Working Paper 09/30. The University of York.

Baumgartner, H., and J. E. M. Steenkamp (2001): "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38(2), 143–156.

Blanchflower, D. G., and A. Oswald (1999): "Well-Being, Insecurity, and the Decline of American Job Satisfaction," Mimeo, University of York.

Boes, S., and R. Winkelmann (2005): "Ordered Response Models," Working Paper No. 0507, Socioeconomic Institute, University of Zurich.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Learner. Edward E. Leamer, North Holland, Volume 5.

Browning, M., and T. F. Crossley (2009): "Are Two Cheap, Noisy Measures Better Than One Expensive, Accurate One?," IFS Working Paper W09/01.

Clark, A. E., P. Frijters, and M. A. Shields (2008): "Relative Income, Happiness and Utility: An Explanation for the Easterlin Paradox and Other Puzzles," *Journal of Economic Literature*, 46(1), 95–114.

Cunha, F., J. J. Heckman, and S. Navarro (2007): "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds," NBER Technical Working Paper Series 2007.

Delaigle, A., P. Hall, and A. Meister (2008): "On Deconvolution with Repeated Measurements," *The Annals of Statistics*, 36(2), 665–685.

Delaney, C., A. Harmon, P. Kapteyn, J. P. S. Smith, and A. L. Van Soest (2007): "Validating the Use of Vignettes for Subjective Threshold Scales," RAND Labour and Population Working Paper WP-501.

Diener, E. (2006): "Guidelines for National Indicators of Subjective Well-Being and Ill-Being," *Journal of Happiness Studies*, 7(4), 397–404.

Eluru, N., C. R. Bhat, and D. A. Hensher (2008): "A Mixed Generalised Ordered Response Model for Examining Pedestrian and Bicylist Injury Severity Level in Traffic Crashes," *Accident Analysis an Prevention*, 40, 1033–1054.

Ferrer-I-Carbonell, A., and P. Frijters (2004): "How Important is Methodology for the Estimates of the Determinants of Happiness," *Economic Journal*, 114, 641–659.

Frey, B., and A. Stutzer (2007): "Should National Happiness Be Maximised?," Zurich IEER Working Paper No. 306.

Greene, W. H. (2007): *LIMDEP 9.0 Econometric Modeling Guide*. Econometric Software Inc.

Greene, W. H., and D. A. Hensher (2010): *Modeling Ordered Choices: A Primer*. Cambridge University Press, Cambridge.

Hausman, J. A., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the Dependent Variable in a Discrete Response Setting," *Journal of Econometrics*, 87, 239–269.

Helliwell, J., and R. D. Putnam (2005): "The Social Context of Well-Being," in *The Science of Well-Being*, ed. by A. Huppert, B. Keverne, and N. Baylis. Oxford University Press, Oxford.

Inglehart, R. F., and H.-D. Klingemann (2000): *Subjective Well-Being Across Cultures*chap. Genes, Culture, Democracy and Happiness, pp. 165–184. MIT Press.

Kahneman, D., A. B. Krueger, D. Schkade, N. Schwarz, and A. Stone (2004): "Toward National Well-Being Accounts," *American Economic Review*, 94(2), 429–434.

Kapteyn, A., J. P. Smith, and A. Van Soest (2007): "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands," *American Economic Review*, 97(1), 461–472.

——— (2009): "Life Satisfaction," IZA Discussion Paper 4015.

King, G. A., 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(1), 191–207.

King, G. A., and J. Wand (2007): "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes," *Political Analysis*, 15, 46–66.

Kotlarski, I. (1967): "On Characting the Gamma and the Normal Distribution," *Pacific Journal of Mathematics*, 20, 69–76.

Kreider, B. (1999): "Latent Work Disability and Reporting Bias," *Journal of Human Resources*, 34(4), 734–769.

Krieder, B., and J. Pepper (2008): "Inferring Disability Status from Corrupt Data," *Journal of Applied Econometrics*, 23, 329–349.

Kristensen, N., and E. Johansson (2006): "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes," Working Paper No. 06-1, Department of Economics, Aarhus School of Business.

Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Lyubomirsky, S., K. Sheldon, and D. Schkade (2005): "Pursuing Happiness: The Architecture of Sustainable Change," *Review of General Psychology*, 9, 111–131.

Maddala, J. (1983): *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Olken, B. A. (2007): "Corruption Perception Vs. Corruption Reality," Working Paper, Harvard University.

Peracchi, F., and C. Rossetti (2009): "Gender and Regional Differences in Self-Rated Health in Europe," CEIS Research Paper No. 142.

Pudney, S. (2008): "Survey Response Error and Econometric Analysis," Lecture Notes, University of Essex.

Pudney, S., and M. Shields (2000): "Gender, Race, Pay and Promotion in the British Nursing Profession: Estimation of a Generalised Ordered Probit Model," *Journal of Applied Econometrics*, 15, 367–399.

Rabe-Hesketh, S., and A. Skrondal (2008): *Multilevel and Longitudinal Modeling Using Stata*. Chapman and Hall, CRC Press, Boca Raton, FL, 2 edn.

Ramalho, E. A. (2001): "Regression Models for Choice-Based Samples with Misclassification in the Response Variable," *Journal of Econometrics*, 106, 171–201.

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), 20–31.

Salomon, J. A., A. Tandon, and C. J. L. Murray (2004): "Comparability of Self-Rated Health: Cross Sectional Multi-Country Survey Using Anchoring Vignettes," *British Medical Journal*, 328, 258–260.

Schennach, S. M. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72(1), 33–75.

SHARE (2009): "Survey of Health, Ageing and Retirement in Europe, Guide to Release 2.3.0 Waves 1 and 2.," Mannheim Research Institute for the Economics of Ageing.

Terza, J. (1985): "Ordered Probit: A Generalization," in *Commucations in Statistics - A. Theory and Methods*, vol. 14, pp. 1–11.

Van Praag, B. M. S. (1971): "The Welfare Function of Income in Belgium: An Empirical Investigation," *European Economic Review*, 2, 337–369.

Van Soest, A., L. Delaney, C. Harmon, A. Kapteyn, and J. P. Smith (2007): "Validating the Use of Vignettes for Subjective Threshold Scales," IZA Discussion paper No. 2860, CentER Discussion Paper Series No. 2007-43, RAND Working Paper No. WR-501.

Weeks, M. (2010): "Subset Selection for Stochastically Ordered Distributions: An Application to Efficiency Measurement," mimeo, Faculty of Economics, University of Cambridge.

Table 1: List of Variables

Variables	Description
Life Satisfaction	How satisfied are you with your life in general? 1=Very dissatisfied; 2= Dissatisfied 3= Neither satisfied nor dissatisfied; 4= Satisfied, 5= Very satisfied
Female	Dummy=1 if respondents is female
Age	Respondent's age at the time of interview
Age Squared	Square of respondent's age
Years Education	Numbers of years in education
Household Size	Number of people living in the household
Household Income	Log of total income received by all household members in last month (In Euros and PPP adjusted)
Living with Partner	Dummy=1 if respondent lives with a spouse or partner
Partner never Worked	Dummy=1 if partner ever done paid work
Limitation	Dummy=1 if respondent has limitations with activities
Depression	Depression scale: ordinal 0 - 11 [high]
Chronic Diseases	Number of chronic diseases
Received Help	Dummy=1 if respondent received help from outside the household
Voluntary Work	Dummy=1 if respondent has done any voluntary/charity work
Educational Activity	Dummy=1 if respondent has attended educational or training course
Sport/Social Activity	Dummy=1 if respondent has gone to sport, social or other kind of club
Political Activity	Dummy=1 if respondent has taken part in political or community organization
Religious Activity	Dummy=1 if respondent has taken part in religious organization
Country	Dummy=1 if respondent lives in Germany, Sweden, Netherlands, Spain, Italy, France Denmark, Greece, Belgium, Czech Republic, Poland.

Table 2: Descriptive Statistics

Variable	Mean	S.D.	Min	Max
Life Satisfaction	3.90	0.77	1	5
Vignette 1	2.66	0.81	1	5
Vignette 2	3.52	0.84	1	5
Female	0.53	0.50	0	1
Age	64.90	9.81	50	97
Years Education	11.21	3.99	1	25
Household Size	2.09	1.05	1	10
Log Household Income	7.56	1.20	0	13.82
Living with Partner	0.67	0.47	0	1
Partner never Worked	0.02	0.15	0	1
Limitation	0.44	0.50	0	1
Depression	2.25	2.19	0	11
Chronic Diseases	1.68	1.54	0	10
Received Help	0.24	0.42	0	1
Voluntary Work	0.15	0.35	0	1
Educational Activity	0.08	0.26	0	1
Social Activity	0.22	0.42	0	1
Religious Activity	0.11	0.31	0	1
Political Activity	0.05	0.21	0	1
N	3927			

Source: Share Data Wave-2 2006.

Table 3: Studies that have used Vignettes

Application	Countries	Testing	Authors
Political Efficacy	Mexico and China	No	King et al. 2004
Job Satisfaction	European countries	No	Kristensen and Johanson 2008
Work Disability/Life Satisfaction	US and Netherlands	No	Kapteyn et al. 2007, Kapteyn et al. 2009
Drinking Behaviour	Ireland	Response Consistency	Delaney et al. 2007
Functioning Mobility and Related Health Problems	UK	Vignette Equivalence and Response Consistency	Bago D'Uva et al. 2009

Table 4: Distribution by Country

	Country											
	Germany	Sweden	Netherlands	Spain	Italy	France	Denmark	Greece	Belgium	Czechia	Poland	
						%						
Life Satisfaction												
Very dissatisfied	0	0	0	0	1	1	0	1	1	0	2	
Dissatisfied	4	2	1	7	12	4	2	11	5	5	6	
Neither satisfied nor dissatisfied	20	13	6	13	23	24	7	37	17	23	27	
Satisfied	62	54	70	66	58	58	50	37	57	65	57	
Very satisfied	14	31	23	14	7	12	41	14	19	7	9	
Vignette 1												
Very dissatisfied	2	8	4	4	13	7	1	9	8	1	5	
Dissatisfied	30	51	37	52	43	50	28	38	50	29	44	
Neither satisfied nor dissatisfied	50	33	51	25	35	38	45	37	33	53	35	
Satisfied	18	8	8	19	9	6	25	13	9	16	15	
Very satisfied	0	1	0	1	0	0	0	3	1	0	1	
Vignette 2												
Very dissatisfied	0	1	8	0	3	2	1	1	1	1	2	
Dissatisfied	12	11	11	14	13	16	5	23	13	4	15	
Neither satisfied nor dissatisfied	32	32	35	29	34	43	19	33	27	29	25	
Satisfied	51	50	43	52	49	38	61	33	48	58	48	
Very satisfied	5	6	3	5	2	2	15	10	13	8	10	
						Number						
Vignette Sample	590	305	256	163	363	223	573	183	459	464	348	3,927
Non-Response Vignette 1	10	8	10	6	8	12	28	0	38	18	17	155
Non-Response Vignette 2	10	7	12	7	8	11	30	0	38	19	15	157
Full Sample	2568	2745	2661	2228	2983	2968	2616	3243	3169	2830	2467	30478

Table 5: Ordered and Generalised Ordered Probit

Life Satisfaction	Ordered Probit	Generalised Ordered Probit			
		Deterministic		Stochastic	
	$\widehat{\beta}$	$\widehat{\beta}$	$\hat{\gamma}_4$	$\widehat{\beta}$	$\hat{\gamma}$
Female	0.15**	0.22	0.02	0.16**	-
Age	0.04	0.03	-0.01	0.04*	-
Age Squared	-0.01	-0.01	0.01	-0.01	-
Years Education	0.01*	-0.01	-0.01	0.01**	-
Household Size	-0.05*	-0.01	0.01	-0.05*	-
Log Household Income	0.09**	-0.01	-0.03	0.10**	-
Living with Partner	0.29**	0.09	-0.03	0.32**	-
Partner never Worked	-0.03	1.06	0.22	-0.04	-
Limitation	-0.32**	-0.24	0.02	-0.35**	-
Depression	-0.18**	-0.15**	-0.01	-0.19**	-
Chronic Diseases	-0.05**	-0.07	0.01	-0.05**	-
Received Help	0.05	0.56**	0.12**	0.06	-
Voluntary Work	0.15**	0.61**	0.13*	0.17**	-
Training Activity	0.06	0.03	0.01	0.07	-
Social Activity	0.14**	0.67**	0.15**	0.17**	-
Religious Activity	0.14*	-0.16	-0.08	0.15**	-
Political Activity	0.22*	-0.33	-0.17	0.24**	-
Country Dummies	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
α_1	-	-		-	
α_2	1.18**	1.18*		0.18*	
α_3	2.27**	1.19**		0.12*	
α_4	4.27**	1.85**		0.82**	
σ_{α_1}	-	-			
σ_{α_2}	-	-		0.14	
σ_{α_3}	-	-		0.11	
σ_{α_4}	-	-		0.35**	
Sample Size	3927	3927		3927	
Log-likelihood.	-3700.75	-3617.30		-3665.3	
LR (Prob>=χ^2)	-	166 (0.00)		70 (0.00)	

$+\ p < 0.10$, $*\ p < 0.05$, $**\ p < 0.01$.

Table 6: Hopit with Vignettes

Life Satisfaction

	All		Group 1		Group 2	
	$\widehat{\beta}$	$\hat{\gamma}_4$	$\widehat{\beta}$	$\hat{\gamma}_4$	$\widehat{\beta}$	$\hat{\gamma}_4$
Female	0.13*	0.01	0.13	-0.06	0.11	0.02
Age	0.09**	-0.02	0.14**	-0.01	0.08+	-0.02
Age Squared	-0.01*	0.01	-0.01*	0.01	-0.01	0.01
Years Education	0.02*	-0.01	0.03*	-0.01	0.01	-0.01
Household Size	-0.03	-0.01	-0.05	0.02	-0.02	-0.01
Log Household Income	0.12**	0.01	0.10*	-0.01	0.13**	0.01
Living with Partner	0.32**	0.07*	0.48**	0.10	0.24**	0.05
Partner never Worked	0.08	0.04	0.22	0.03	0.23	0.03
Limitation	-0.39**	-0.04	-0.37**	-0.01	-0.43**	-0.08*
Depression	-0.17**	-0.02*	-0.15**	0.01	-0.18**	-0.02**
Chronic Diseases	-0.03	0.01	-0.04	-0.01	-0.03	0.01
Received Help	0.07	0.02	0.10	0.09	0.09	-0.02
Voluntary Work	0.32**	0.08*	0.23+	0.08	0.35**	0.08
Training Activity	-0.08	0.01	-0.23	-0.03	-0.01	0.01
Social Activity	0.23**	0.04	0.24*	0.03	0.20*	0.05
Religious Activity	0.11	-0.01	0.24+	0.04	0.07	0.02
Political Activity	0.22	-0.04	0.27	-0.06	0.28+	0.01
Constant	-	1.47**	-	0.92	-	1.30+
Country Dummies	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Vignette equation						
$\widehat{\theta}_1$	2.68*		4.77*		1.83	
$\widehat{\theta}_2$	3.97**		5.74**		3.30*	
$\widehat{\sigma}_\omega$	0.11**		0.04		0.13**	
Sample Size	3927		1252		2327	
Log-likelihood.	-12815		-4050		-7479	

$+\ p < 0.10$, $*\ p < 0.05$, $**\ p < 0.01$

Table 7: Hopit: Test for Vignette Equivalence

Life Satisfaction

	All Countries	Group1	Group2
$\widehat{\theta}_1$	2.83*	4.82*	0.85
$\widehat{\theta}_2$	3.80**	5.74**	2.42+
$\widehat{\theta}_2$*Netherlands	0.15	0.14	
$\widehat{\theta}_2$*France	0.21	0.21	
$\widehat{\theta}_2$*Greece	-0.12	-0.11	
$\widehat{\theta}_2$*Sweden	0.56**		-
$\widehat{\theta}_2$*Spain	0.28+		-0.29
$\widehat{\theta}_2$*Italy	0.38**		-0.19
$\widehat{\theta}_2$*Denmark	0.39**		-0.17
$\widehat{\theta}_2$*Belgium	0.70**		0.14
$\widehat{\theta}_2$*Czechia	0.47**		-0.10
$\widehat{\theta}_2$*Poland	0.25*		-0.33*
$\widehat{\sigma}_\omega$	0.11**	0.04	0.15**
Reference Country	Germany	Germany	Sweden
Sample Size	3927	1192	2675
Log-likelihood	-12783.6	-4047.4	-8693.3

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 8: Country Rankings

	Ordered Probit		Generalised Ordered Probit				Hopit					
			Deterministic		Stochastic		All		Group 1		Group 2	
Ranking	Country	$\widehat{\beta}_C$	Country	$\widehat{\beta}_C$	Country	$\widehat{\beta}_C$	Country	$\widehat{\beta}_C$	Country	$\widehat{\beta}_C$	Country	$\widehat{\beta}_C$
1)	Denmark	0.71**	Spain	1.16*	Netherlands	1.23**	Netherlands	0.93**	Netherlands	0.91**	Sweden	0.42**
2)	Sweden	0.40**	Netherlands	0.79	Poland	0.43**	Sweden	0.49**	France	0.39**	Spain	0.26+
3)	Poland	0.38**	Belgium	0.42	Czech R.	0.38*	Poland	0.43**	Greece	-0.18	Denmark	0.13
4)	Netherlands	0.37**	Denmark	0.33	Spain	0.15	France	0.39**			Italy	-0.08
5)	Spain	0.25*	Poland	0.27	Denmark	0.06	Spain	0.35*			Czech R.	-0.17
6)	Belgium	0.16*	Czech R.	0.12	Belgium	-0.07	Denmark	0.19+				
7)	Czech R.	-0.01	Sweden	0.10	Sweden	-0.14	Belgium	0.07				
8)	France	-0.10	Italy	0.01	France	-0.20	Italy	0.01				
9)	Italy	-0.33**	Greece	-0.34	Italy	-0.23	Czech R.	-0.09				
10)	Greece	-0.37**	France	-0.35	Greece	-0.43**	Greece	-0.22+				
Reference	Germany		Germany		Germany		Germany		Germany		Belgium	

$+ \ p < 0.10$, $*$ $p < 0.05$, $**$ $p < 0.01$.

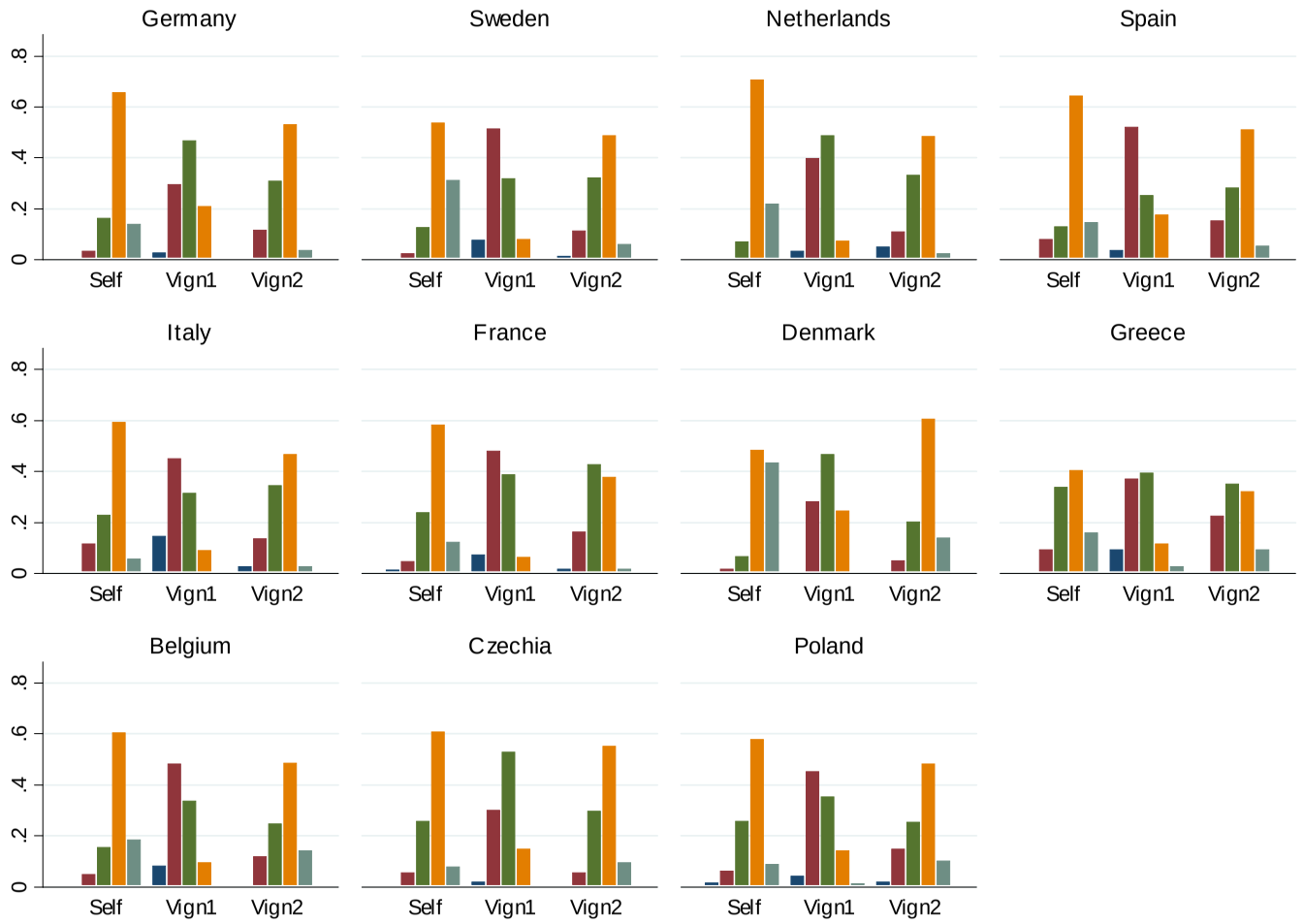

Figure 1: Distribution of Self-Reports and Vignettes.