BANKWEST CURTIN ECONOMICS CENTRE

WORKING PAPER SERIES

15/8: SURVEY SELF-ASSESSMENTS, REPORTING BEHAVIOUR AND THE USE OF EXTERNALLY COLLECTED VIGNETTES

Mark N. Harris, Rachel Knott, Paula Lorgelly and Nigel Rice

business.curtin.edu.au/bcec

Bankwest Curtin Economics Centre Working Paper Series
ISSN: 2202-2791
ISBN: 978-1-925083-37-8

Mark N. Harris[1] Curtin University

Rachel Knott[2] Monash University

Paula Lorgelly[3] Monash University

Nigel Rice[4] University of York

Suggested Citation

Mark N. Harris, Rachel Knott, Paula Lorgelly and Nigel Rice, December, 2015 "Survey self-assessments, reporting behaviour and the use of externally collected vignettes." Bankwest Curtin Economics Centre Working Paper 15/8, Perth: Curtin University.

Survey self-assessments, reporting behaviour and the use of externally collected vignettes*

Mark N. Harris

Curtin University, Perth, Australia

Rachel Knott

Monash University, Melbourne, Australia

Paula Lorgelly

Monash University, Melbourne, Australia

Nigel Rice

University of York, York, UK

Abstract

The anchoring vignette approach has grown in popularity as a method to adjust for reporting heterogeneity in subjective self-reports, removing bias due to systematic variation in reporting styles across study respondents. The use of anchoring vignettes, however, has been limited to surveys where both self-reports and vignette questions have been included. This diminishes their wider application. We illustrate, using an application to self-assessed health in a large household survey, how externally collected vignettes can be used to adjust for reporting heterogeneity in self-reports observed in datasets where vignettes have not been included. Given that self-reports to survey questions are an important facet of social research to understand differences across socio-economic groups and populations, we anticipate the approach described will lead to new applications of the anchoring vignette methodology.

*We are extremely grateful to the Australian Research Council and the Bankwest-Curtin Economics Centre for funding. The usual caveats apply.

Keywords: Anchoring vignettes, self-reports, reporting heterogeneity.
JEL: I1, C1, C3

1. Introduction and Background

The use of subjective scales to elicit information in the form of self-assessments or self-reports of the circumstances, preferences or beliefs of respondents are ubiquitous in social surveys. Such questions are typically inexpensive to administer and in the absence of more objective measures contain valuable information from which to infer differences across scoio-economic groups or countries. As such, the analysis of self-reported data using *likert*-type response scales forms the basis of a large amount of literature and resulting policy advice. Examples include the generic self-assessed health measure which asks respondents to rate their health using ordered response categories typically ranging from *very bad* (or *poor*) health through to *very good* (or *excellent*) health; and job and life satisfaction which use similar response scales ranging from *complete dissatisfaction* through to *complete satisfaction*.

An inherent problem with any measure using subjective categorical responses is that interpretation of the response scales are likely to vary from person to person, as will the implicit benchmarks that people use to evaluate themselves. Accordingly, responses will depend both on an objective reality and a respondent's interpretation of the subjective scale. Consequently, two individuals with identical levels of true or perceived health, for example, may rate their health differently in response to a survey question. This issue - a type of reporting heterogeneity commonly referred to as *differential item functioning* (DIF) [12] - can lead to bias when drawing inter-personal comparisons. As a result, analyses undertaking comparison using self-reported data will produce biased results and the implications and policy advice that may be forthcoming are likely to be erroneous.

A methodology for overcoming DIF is the anchoring vignette approach [10], a survey tool which has grown in popularity over the past decade in the literature on health [8], [2], work disability [9], political efficacy [10], and job and life satisfaction [1], [11]. The approach involves the use of one or more

vignettes describing situations of hypothetical individuals, which respondents evaluate in addition to their own situation. Responses to the vignettes are then used to *anchor*, or adjust for bias in self-reports introduced by DIF; such that inter-personal comparisons can be appropriately examined, resulting in more accurate policy inference.

Although this method has proved useful, its application has been limited to use in datasets where vignettes have been collected alongside self-reports of the construct of interest.[1] Here we illustrate how to use vignette responses collected externally to the main survey containing the self-reports, using generic self-assessed health in an application. The approach will be particularly valuable to researchers interested in adjusting for DIF in datasets where vignette responses have not been elicited, and hence widen the applicability of the vignette methodology. As shown below, the vignette responses may be readily available in a different dataset to the self-report of interest; or they could be newly collected in a bespoke survey. Often, when working with self-reports, researchers invariably favour large scale nationally representative, tried-and-tested datasets for their analyses. However, the absence of vignettes means they are unable to test and adjust for DIF and therefore make appropriate and robust comparisons and inference. Here we present easy-to-implement methods that combine extraneous vignette information with the self-report of interest, to make such adjustments. In addition we illustrate the robustness of the resulting parameter estimates to any observed lack of balance in covariates determining reporting behaviour across the sample of vignette responses and the sample containing self-assessed outcomes. Where imbalance occurs we weight the sample of vignette respondents to be respresentative of individuals completing the self-assessments such that

[1] Notable examples of datasets which contain both self-reports and vignettes include the Survey of Health, Ageing and Retirement in Europe (SHARE), the English Longitudinal Survey of Ageing (ELSA), the Health and Retirement Survey (HRS), and the World Health Survey (WHS).

inference can be made with respect to the principle survey of interest. Given the important role of survey self-assessments in political, economic and social science research, we anticipate the approach described will lead to new applications of the anchoring vignette methodology.

2. Methods

The ordered probit

Self-reported measures requiring responses on *likert*-type scales are invariably analysed using ordered probit/logit models [7]. Underlying the standard ordered probit (OP) model is a latent variable, y^*, which is a linear (in unknown parameters, β_y) function of observed characteristics (with no constant term) $\mathbf{x}$; a disturbance term (unrelated to any observed heterogeneity in the model), ε_y; and its relationship to certain boundary parameters, μ, such that:

$$y^* = \mathbf{x}'\boldsymbol{\beta}_y + \varepsilon_y, \tag{1}$$

translating into observed $j = 0, \ldots, J-1$ outcomes via the mapping

$$y = \left\{ \begin{array}{ll} j & \text{if } \mu_{j-1} \leq y^* < \mu_j \quad for\ j = 0, \ldots, J-1, \end{array} \right. \tag{2}$$

where $\mu_{-1} = -\infty$ and $\mu_{J-1} = +\infty$; and to ensure well-defined probabilities, $\mu_{j-1} \leq \mu_j$, $\forall j$.

Under the assumption of normality, the probabilities for each ordered outcome are $\Phi\left(\mu_0 - \mathbf{x}'\boldsymbol{\beta}_y\right)$, for $j = 0$; $\left[\Phi\left(\mu_{j-1} - \mathbf{x}'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_j - \mathbf{x}'\boldsymbol{\beta}_y\right)\right]$ for $j = 1, \ldots, J-2$; and $1 - \Phi\left(\mu_{J-2} - \mathbf{x}'\boldsymbol{\beta}_y\right)$, for $j = J-1$, respectively; where $\Phi(.)$ denotes the standard normal distribution function evaluated at its argument. The (log) density for this model for a $i = 1, \ldots, N$ random sample of individuals is simply given by

$$\ln L_{OP}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \sum_{j=0}^{J-1} d_{ij} \left[\Pr\left(y_i = j \,|\mathbf{X}\right)\right], \tag{3}$$

where d_{ij} is a function returning one if individual i chooses outcome j, and zero otherwise, and $\boldsymbol{\theta}$ denotes all model parameters [7].

Heterogeneity in boundary parameters

Results from the ordered probit model may be biased if the response scales that individuals use to evaluate themselves vary systematically across individuals**,** in which case individual variation should be allowed for in the boundary parameters, μ_{ij} (also referred to as inter-category thresholds or cut-points; see for example, [16], [14], [4], [7], [6]). An easy way to incorporate this is simply to let μ_{ij} depend on a set of observed characteristics $\mathbf{z}_i$ such that $\mu_{ij} = \mathbf{z}_i'\gamma_j$. However, to help identification and ensure well-defined probabilities many authors ([8][9][1][11][15]) adopt a hierarchical ordered probit ($HOPIT$) approach by specifying the boundaries as:[2]

$$\begin{aligned} \mu_{i0} &= \mathbf{z}_i'\boldsymbol{\gamma}_0 \\ \mu_{ij} &= \mu_{ij-1} + \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}_j\right) \\ &\vdots \end{aligned} \tag{4}$$

The model is typically estimated by maximum likelihood techniques, where the implicit $\mu_j s$ in equation (3) are replaced by those of equation (4).

Since the first threshold is specified linearly, the corresponding elements of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}$ are not separately identifiable for variables that appear in both $\mathbf{x}$ and $\mathbf{z}$, in the absence of any further information. Exclusion restrictions can overcome this issue, such that $\mathbf{x}$ and $\mathbf{z}$ are distinct vectors [14]; however empirically, exclusion restrictions are often difficult to justify. An alternative approach involves the use of (anchoring) vignettes which consist of brief statements describing situations of hypothetical individuals. Respondents

[2]Note the model is also sometimes referred to as the Compound Hierarchical Ordered Probit Model ($CHOPIT$).

are asked to evaluate these vignettes in addition to their own self-assessments; where situations portrayed in the vignettes relate to the same construct of interest as the self-assessments - for example, health as in our illustration below.

Say we have $k = 1, \ldots, K$ possible vignettes, where each k vignette is asked on the same $j = 0, \ldots, J-1$ ordinal scale as the self-report of interest. The observed response, y_{ik}, to each $k = 1, \ldots, K$ possible vignette is determined as before, such that $y_{ik} = j$ if $\mu_{ik}^{j-1} \leq y_{ik}^{*} < \mu_{ik}^{j}$, $k = 1, \ldots, K$; $j = 0, \ldots, J-1$; with $y_{ik}^{*} = \alpha_k + \varepsilon_{ik}$ and $\varepsilon \sim N(0, \sigma_k^2)$ and orthogonal to all observed covariates in the model. Usually the simplifying assumption that $\sigma_k^2 = \sigma_v^2 \; \forall k$ is made. Importantly, heterogeneity across these response scales is once more allowed for by specifying the boundaries as a function of threshold variables, $\mathbf{z}_i$ (where typically $\mathbf{z}_i \equiv \mathbf{x}_i$).

The approach relies on the identifying assumptions of *response consistency* (RC) - that the response scale used by each individual, i, is the same across self- and vignette-assessments; and *vignette equivalence* (VE) - that vignettes are interpreted in the same way and on the same unidimensional scale across respondents [10]. The RC assumption amounts to restricting all coefficients in all of the reporting parts of the model (the boundary parameters: $\boldsymbol{\gamma}_j \forall j$) to be the same; *i.e.*, $\boldsymbol{\gamma}$ in the $HOPIT$ (self-assessment) part of the model is identical to that in the $k = 1, \ldots, K$ $HOPIT$ parts of the vignette equations.[3] With all of these elements in place the (log-)likelihood function will consist of two distinct parts: one relating to the self-report of interest ($\ln L_{HOPIT}$), and the other relating to the vignette component of the model ($\ln L_{V,k}$):

$$\ln L = \ln L_{HOPIT} + \sum_{k} \ln L_{V,k}, \tag{5}$$

[3] A useful summary of the various restriction strategies available to the researcher in the presence of vignettes, is given by [13].

where the first term is a function of α_k, σ_v and $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$ and the second a function of $\boldsymbol{\beta}$ and $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$. These two components are linked through the common boundary parameters $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$, and so do not factorise into two independent models.

Adjusting for reporting heterogeneity with external vignettes

To the best of our knowledge, DIF-adjustment has thus far been restricted to situations where both vignettes and self-reports are contained in a single survey dataset. Many large scale social surveys, however, do not contain vignettes, but hold a wealth of data on self-assessments.[4] We show that the lack of vignettes in such surveys should not necessarily preclude the use of the $HOPIT$ approach to adjust for systematic reporting behaviour.

The exposition above makes it clear that the two parts of the likelihood are linked only by the common boundary parameters, $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$. For ease of notation, assume that there is only one vignette assessment, then the above likelihood could equivalently be written

$$\ln L = \sum_{i=1}^{N} \ln L_{i,HOPIT} + \sum_{q=1}^{Q} \ln L_{q,V}, \tag{6}$$

where $i = 1, \ldots, N$ and $q = 1, \ldots, Q$ could potentially index two different samples. The only requirement, other than the implicit assumption that the DIF problem is the same across the two samples (for RC to be maintained), is that the vector of variables, $\mathbf{z}$, in the vignette sample are the same as those included in the boundary equation for the principal sample of interest. Additionally, as in the case where vignettes are collected alongside self-reports, we need to assume both RC and VE hold. We further need to assume that the boundary equations (5) representing reporting behaviour are correctly specified. Imposing common support across the two samples in the char-

[4]For example, the British Household Panel Survey (BHPS), Understanding Society (USoc), and the Household, Income and Labour Dynamics in Australia (HILDA).

acteristics, $\mathbf{z}$, will further strengthen claims for RC by ensuring reporting behaviour in the main sample does not involve extrapolation of reporting behaviour identified on the vignette sample. For example, since adjusting for country-specific reporting behaviour has been found to be important to improve inter-country comparability of self-reports ([10][9][1][15]), samples forming Q and N would, at the very least, be required to be taken from the countries under comparison.

It is worth noting that the approach does not exclude the specification of additional variables in the structural equation ($\mathbf{x}$); all that is required is variable equivalence across samples with respect to $\mathbf{z}$. This assumes that such variables do not determine reporting behaviour and are appropriately excluded from $\mathbf{z}$. These might consist of variables that are only available in the principle dataset of interest.

Representativeness of vignettes survey

Estimation of the $HOPIT$ model is undertaken by maximising the likelihood in equation (6). Analyses of surveys that contain both self-assessments and vignettes typically set $N = Q$ by restricting the sample to observations where respondents provide non-missing information on both types of questions. By construction, therefore, in surveys containing responses to both vignettes and self-assessments balance is maintained across the characteristics determining reporting behaviour. Consistent estimation of the parameters, β_y, then rests on the validity of the assumptions of RC and VE.

Where vignettes are drawn from a separate sample to the self-assessments, the contribution to the likelihood is likely to be dominated by observations contained within the latter (in our example, this is the Household, Income and Labour Dynamics of Australia ($HILDA$) survey) as, in general, $N \gg Q$. In addition, the two samples may display imbalance with respect to characteristics, $\mathbf{z}$. That is, respondents to the vignettes may not be fully representative of the sample of individuals completing the self-assessment drawn from the principle survey of interest. Since reporting behaviour is identified

on vignette sample respondents, and imposed on the main sample via the assumption of response consistency, a lack of balance in the characteristics determining reporting behaviour may lead to biases in estimates of β_y. This is likely to be exacerbated further where there is a lack of common support in the characteristics, $\mathbf{z}$, across the two samples.[5] This may be particularly important where the principle sample of interest contains elements of $\mathbf{z}$ beyond support observed in the vignette sample, and hence the relationship between the set of covariates, $\mathbf{z}$, and reporting behaviour requires extrapolation to regions outside common support when applied to the main survey of interest.

Weighting

In circumstances where the vignette sample is not representative of respondents in the main survey of interest, the respondents can be weighted such that balance in the characteristics, $\mathbf{z}$, is approximately achieved across the two sources of data. Assume the set of characteristics of reporting behaviour is small and the majority of variables are discrete, which typically is the case in applications. Weighting can be achieved by first coarsening any cardinal variables into appropriate intervals and counting the number of respondents in both the vignette and principle survey samples falling into each distinct strata, with strata defined by the multivariate distribution of the set of characteristics, $\mathbf{z}$, under consideration. A small covariate set, particularly those requiring coarsening, together with common support across the set of characteristics, $\mathbf{z}$, helps to ensure there are few strata populated by respondents from only one of either the vignette or main survey sample. Assume all possible combinations of the discrete and coarsened variables, $\mathbf{z}'$, observed across sample respondents produces J strata. If the number of vignette respondents falling within a given strata is

[5] That is, where the empirical density of the characteristics determining reporting behaviour in the vignette sample and for $HILDA$ do not overlap.

$Q_j, \left(j = 1, \ldots, J, \text{with } \sum_{j=1}^{J} Q_j = Q\right)$ and the corresponding number in the main sample is N_j, then the likelihood in equation (6) can be weighted such that:

$$\ln L = \sum_{i=1}^{N} \ln L_{i,HOPIT} + \sum_{q=1}^{Q} w_{ij} \ln L_{q,V}, \tag{7}$$

where $w_{ij} = \frac{N_j}{N} \frac{Q}{Q_j}, \quad j = 1, \ldots, J.$

Since weighting produces a vignette sample more representative of the principle sample of interest, maximising the likelihood in equation (7) imposes reporting behaviour identified on a sample displying greater balance across the characteristics though, a priori, to be important drivers of reporting styles. This strengthens claims for RC.

Although our empirical example considers self-assessed health; clearly the approach is applicable to any self-reports of interest, provided that appropriate vignettes (i.e., vignettes relating to the same construct as the self-report and using the same response scales) have been collected in other data sources. Researchers may therefore choose to administer their own ancillary survey and collect vignette responses on a (potentially smaller) sample to that for which self-assessments are derived; this is the approach utilised in the empirical example below. Alternatively, certain waves of existing household surveys already contain vignette components which might be used to externally adjust for DIF.[6]

It is important to reiterate that the approach relies on the implicit assumption that reporting styles are similar across samples; so the approach is unlikely to work if, for example, vignette data from a survey in one country is merged with self-reported data from a survey in another country or from

[6]For instance, SHARE (wave 1 and 2) and ELSA (wave 3) include vignettes on health and health limitations; ELSA (wave 3) and the HRS (2007 wave) contain vignettes on work disability; while SHARE (wave 2) also contains vignettes on life and job satisfaction, political influence and health care responsiveness.

clearly distinct time periods. In such situations the assumption of response consistency is unlikely to be tenable.

We consider two estimation samples. The first simply pools the two surveys - $HILDA$ and the vignette sample - to estimate the $HOPIT$ model. We refer to the joint sample of pooled $HILDA$ and vignette samples as the *full sample*. Second, to improve balance and enhance the representativeness of vignette sample respondents to respondents in the principle survey of interest we weight the former in the likelihood as outlined in equation (7). Henceforth we refer to this as the *weighted sample*.

3. Empirical example

We illustrate the approach empirically by correcting for *DIF* in self-assessed health in the widely used Household, Income and Labour Dynamics of Australia ($HILDA$) survey, using vignette responses collected in an online survey involving 5,034 Australian respondents. We focus on the generic self-assessed health question in $HILDA$ which asks respondents "*In general, would you say your health is excellent, very good, good, fair or poor?*". The online survey was conducted in April 2014 and in August 2015 and targeted a representative sample of Australians aged 18-65. Three vignettes were included describing health states of differing levels of severity which are presented in the appendix. The categories available to respondents when rating the vignettes are the same as those available for self-assessed health in $HILDA$. Importantly, the online survey also contained a set of questions on socio-demographic characteristics of respondents which correspond to questions asked in $HILDA$.

The top panel of Table 1 displays descriptive statistics for both samples, where analysis of $HILDA$ is restricted to the latest wave (*i.e.*, wave 13 or 2013) and those aged between 18 and 65 at the time of survey ($n = 12009$), so as to be as comparable as possible with the auxiliary survey. The two samples are similar in terms of age (mean age of the $HILDA$ sample is

40.8 years; mean age of the vignette sample is 41.4), gender (53% of the $HILDA$ sample and 52% of the vignette sample are female), marital status (approximately 63% of the $HILDA$ sample and 59% of the vignette sample are married) and migrant status (21% of the $HILDA$ sample and 25% of the vignettes sample are born in countries other than Australia). The vignette sample is, however, slightly more highly educated (69% versus 62%) and yet more likely to be unemployed (10% versus 4%). The final column of the table presents p-values for tests of difference in means (age) or difference in proportions across the two samples, under the null of equality. As can be seen, with the exception of gender we reject the null at conventional levels of significance, again indicating differences across the two samples which may be important for drawing inference relevant to the population of respondents of the self-assessments contained within $HILDA$.

Insert Table 1 about here

The two samples differ most notably for education and labour market status. This is a likely consequence of using an online survey which recruited respondents via a panel company - internet users are more likely to be better educated than the general public, but possibly more likely to be unemployed as they are paid to undertake such surveys. It is worth noting, however, that for the example presented here, while there is imbalance across the two samples, there is common support over the set of characteristics.

As outlined above, we also consider a weighted sample of vignette respondents with weights designed to improve the representativeness of the sample with respect to the principle sample of interest. This is achieved by firstly coarsening age - the only cardinal variable - into 5-year age groups and secondly, considering the distinct strata formed from the set of coarsened and binary variables. In total this leads to 721 strata observed in the data. For each strata the number of individuals within $HILDA$ and the number within the vignette sample are computed. These can then be used

to compute the weights required to produce a distribution of respondents in the vignette sample representative of the distribution in $HILDA$, but scaled to the original sample size of 5034. Of the 721 possible strata, 504 were populated by both vignette and $HILDA$ sample members. These are the vignette respondents to which weighting is applied. A further 49 strata contained only vignette, and 94 only $HILDA$ respondents. To maintain the sample size the latter individuals are included in the weighting with a weight of unity. Their inclusion is at the expense of compromising the ability of weighting to produce a sample fully representative of $HILDA$ across the full set of characteristics, $\mathbf{z}$, as there remain combinations of $\mathbf{z}'$ only observed in $HILDA$ or the vignette sample. Weighting in this way, however, produces greater balance in covariates across the two samples. This can be seen in the bottom panel of Table 1. Eyeballing the summary statistics across $HILDA$ and the weighted vignette sample, shows improved balance across all covariates. This is supported by formal statistical tests of the difference in means and proportions. Were this not the case, then strata containing only vignette or $HILDA$ respondent could be removed from analysis (by applying a weight of zero).

Our interest is in estimating the determinants of self-reported health observed through the specification of an outcome model (equation (1)) but adjusted for observed reporting behaviour. For simplicity, and following the predominant empirical literature, we specify $\mathbf{x} \equiv \mathbf{z}$ and adopt a standard set of demographic variables similar to those used elsewhere to model self-assessed health [5], [3]. These characteristics are summarised in Table 1. An ordered probit is applied to the *full sample* and the $HOPIT$ model is estimated on the *full sample*, and the *weighted sample*.

Columns 2 and 3 of Table 2 contains parameter estimates and corresponding standard errors for an ordered probit model, which does not correct for DIF - see equation (3). This is estimated on the sample of $HILDA$ respondents alone. The dependent variable is increasing in health ($y = 0$ denotes

poor health; $y = 4$ is *excellent* health). All parameter estimates are significant at conventional levels ($\leq 5\%$). As expected age is decreasing in health; women report better health than men and education is positively associated with health (where individuals who did not finish year 12 form the reference category). Employment is associated with better health than unemployment (the baseline category), and not being in the labour market is associated with worse health than being unemployed, presumably reflecting selection out of the labour market on the basis of ill-health. Being married is positively associated with health. Being born in a country other than Australia is associated with reporting better health than respondents born in Australia. Since the ordered probit model fails to adjust for differences in reporting behaviour, the estimated effects are associations, representing composite parameters reflecting differences in true underlying health status together with differences in reporting styles.

INSERT TABLE 2 ABOUT HERE

Table 2 also presents results of the $HOPIT$ model. The fourth and fifth columns show coefficient estimates and standard errors for the *full sample* of $HILDA$ combined with the vignette sample. Note the high significance levels for parameters in the threshold equations, indicating a significant degree of reporting heterogeneity across the characteristics contained in $\mathbf{z}$. For example, the coefficient for female is positive and significant (at 5%) in the first threshold equation. This indicates that, on average, women use a higher threshold between the categories representing *poor* and *fair* health compared to men, indicating they are more likely to make use of the *poor* health category. However, this effect is offset by a larger and negative coefficient in the second threshold ($j = 1$) indicating that women also tend to apply a lower threshold between *fair* and *good* health and are more likely to make use of the *good* health response category than men. Similarly, individuals with higher levels of education (Tertiary or Year 12) are more likely to down

report underlying true health by making greater use of the response category *poor*, and less use of the *excellent* category (the corresponding parameters in the thresholds for both $j = 0$ and $j = 3$ are positive and significant at conventional levels, with the exception of Year 12 in the $j = 0$ threshold). However, they are also more likely to report *very good* and less likely to report *fair* or *good* health compared to individuals with less than year 12 education (although apart from teriary edication in the $j = 2$ threshold, these effects are not significant at conventional levels). In contrast, married individuals are less likely to report *poor* health and are more likely to report *excellent* and *very good* health than those who are not married, indicating that married people, in general, tend to over report their health status. These findings strongly imply that misleading conclusions are likely to result when considering models that do not account for such reporting heterogeneity.

Many of the parameters in the outcome equation of interest retain statistical significance in the $HOPIT$ model. However, many of the covariates decline in magnitude and/or significance - some even changing sign. For example, although the ordered probit model suggests that females are in better health than males ($\beta = 0.086, t - stat = 4.355$), the $HOPIT$ model for the full sample reveals a positive but lower and less significant relationship ($\beta = 0.067, t - stat = 2.207$). This suggests that the health of females is worse than indicated by the ordered probit model; this effect follows from the discussion above of the respective parameters for female in the threshold equations.

We also see that the coefficient on marital status moves from large, positive and significant (at the 1% level) in the ordered probit model (columns 2 and 3) to positive but closer to zero and non-significant in the $HOPIT$ model (columns 4 and 5). Again this concurs with the notion that married individuals over-report their health status by making less use of the *poor* health category than unmarried counterparts.

Similar results can be seen across coefficients for age, which decrease in

magnitude when comparing the ordered probit and *HOPIT* models; education, where the gradient is more prominent in the *HOPIT* estimates; and employment status, where the coefficients for employed and not in the labour force are lower in the *HOPIT* model than the ordered probit model. Conversely, the coefficient for migrants increases in value when moving from the ordered probit to the *HOPIT* model. This effect is due to reporting behaviour reflected in the positive and significant respective coefficients in the first and second boundary equations, indicating a greater use of the response categories for *poor* and *fair* health for migrants than Australian natives. This conforms with findings elsewhere that reporting norms vary substantively across cultures and countries ([10], [9], [1], [15]).

Insert Table 3 about here

Columns 6 and 7 of Table 2 present estimates based on the *weighted sample*. Focusing on estimated parameters in the outcome equation and compared to results from the *full sample*, weighting makes a difference with respect to both coefficients and standard errors, particularly for the effects of age, gender, education and labour market status. For example, while the substantive effect of age remains similar to the results of the *full sample*, the magnitudes in absolute terms increase substantially for the *weighted sample*. Similarly, differences in health by gender increase considerably, with women on the whole experiencing better health than men. The magnitudes of the education variables also increases in the weighted sample. The relative penalty in terms of health status of being out of the labour force compared to the employed remains approximately the same in the *weighted sample* compared to the *full sample*. The magnitudes of the effects, however, are higher in the former compared to the latter. For other covariates (maried and migrant status), the observed changes are relatively small and the substantive effects remain the same as those observed for the unweighted *full sample*.

An alternative comparison of the coefficients across samples is provided in Table 3 which presents partial effects for reporting *poor* and *excellent*

health.[7] The first set of results are partial effects and standard errors for the ordered probit model estimated on the $HILDA$ sample. The remaining columns present partial effects for the $HOPIT$ model estimated on the *full sample* and seperately on the *weighted sample*.

Partial effects for reporting *poor* health vary considerably across estimators and samples. With the exception of married and migrant status, there are large discrepancies between the ordered probit and $HOPIT$ estimates (based on the *full sample*) with many of the partial effects estimated on the latter sample being around half the corresponding value of the ordered probit. A comparison between the *weighted* and unweighted (*full sample*) partial effects from the $HOPIT$ model on the whole show greater similarity than comparison with the ordered probit model. Age is perhaps the exception, where a notable difference exists between the two sets of $HOPIT$ results. We further observe large discrepancies in the partial effects of reporting *excellent* health (bottom panel of Table 3) when comparing ordered probit to $HOPIT$ results. This is particularly the case for the effects of not being in the labour market and marital status. Estimates for the former are negative and highly significant in the ordered probit model, but positive and not significant in the $HOPIT$ model estimated on the *weighted sample*. Martial status, loses statistical significance at conventional levels and is approximately half the estimated effect in the $HOPIT$ model compared to the ordered probit.

The set of results for the $HOPIT$ model display substantial evidence of reporting behaviour observed through the significance of covariates $\mathbf{z}$ in the threshold equations (see Table 2). This strongly suggests that inference based on the ordered probit model is erroneous and that the $HOPIT$ model is preferred due to its ability to adjust for reporting heterogeneity. To investigate the determinants of self-assessed health in the $HILDA$ survey

[7]Partial effects for reporting *fair*, *good* or *very good* health are available on request.

this requires merging information on vignettes from an auxiliary source of data, in the case presented here, from an online survey. Whilst $HILDA$ and the online survey have common support across the set of characteristics $\mathbf{z}$ and $\mathbf{x}$, the vignette survey over-sampled educated and unemployed individuals. Weighting this sample to be more representative of $HILDA$ resulted in some notable changes to parameter estimates in the main outcome equation for self-assessed health. This is particularly the case for those characteristics where the two orginal samples lacked balance.

Weighting is undertaken to increase balance (and hence representativeness) across the set of characteristics determining reporting behaviour in the vignette and $HILDA$ surveys. This renders the specification of reporting behaviour less vulnerable to being determined by specific characteristics of the vignette sample that are not reflected in the main survey of interest, likely to have been designed to be representative of the population. Accordingly, inference that follows is appropriate to the outcomes in the principle survey, and not overly influenced by the idiosyncrasies of the sub-population of respondents who volunteer to complete internet surveys.

4. Conclusions

Our empirical example based on a large household survey containing rich socio-economic variables together with self-reported health status, but lacking information on vignettes, illustrates how the $HOPIT$ approach, using externally collected vignettes, can be applied to correct for reporting heterogeneity inherent in subjective measures of interest. In our example, the socio-economic deteminants of self-reported health is the focus of interest with health reported on a 5-point ordered categorical scale. The methodology of anchoring vignettes, and the application of information from externally collected vignettes is, however, applicable to other spheres and outcome in social science research where comparison across individuals, socio-economic groups, or countries is the primary focus and where there is good reason to

believe that the self-reports of interest contain reporting heterogeneity.

While anchoring vignettes have been widely used to correct for reporting heterogeneity, their use in the literature thus far has been limited to analyses of datasets containing both self-assessments and vignette questions. In this paper we demonstrate how vignette responses collected externally to the main dataset of interest can be used to correct for reporting heterogeneity (provided that the relevant assumptions of RC and VE hold). We also show through weighting to create better balance in covariates determining reporting behaviour how information on vignettes can be incorported without losing the ability to generate inference on the target survey of principle interest, which, as in the example provided, may be a population representative household survey. Given that self-reports to survey questions are an important facet of political, economic and social science research, we anticipate this approach will lead to new applications of the anchoring vignette methodology.

[1] Angelini, V., D. Cavapozzi, L. Corazzini, and O. Paccagnella (2014): "Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases," *Oxford bulletin of Economics and Statistics*, 76(5), 643–666.

[2] Bago d'Uva, T., E. Van Doorslaer, M. Lindeboom, and O. O'Donnell (2008): "Does reporting heterogeneity bias the measurement of health disparities?," *Health economics*, 17(3), 351–375.

[3] Balia, S., and A. M. Jones (2008): "Mortality, lifestyle and socio-economic status," *Journal of health economics*, 27(1), 1–26.

[4] Boes, S., and R. Winkelmann (2006): "Ordered response models," *AStA Advances in Statistical Analysis*, 90(1), 167–181.

[5] Contoyannis, P., A. Jones, and N. Rice (2004): "The dynamics of health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19, 473–503.

[6] Greene, W., M. Harris, B. Hollingsworth, and P. Maitra (2014): "A Latent Class Model for Obesity," *Economics Letters*, 123, 1–5.

[7] Greene, W., and D. Hensher (2010): *Modeling Ordered Choices*. Cambridge University Press.

[8] Grol-Prokopczyk, H., J. Freese, and R. M. Hauser (2011): "Using anchoring vignettes to assess group differences in general self-rated health," *Journal of health and social behavior*, 52(2), 246–261.

[9] Kapteyn, A., J. P. Smith, and A. Van Soest (2007): "Vignettes and self-reports of work disability in the United States and the Netherlands," *The American Economic Review*, pp. 461–473.

[10] King, G., C. Murray, J. 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.

[11] Kristensen, N., and E. Johansson (2008): "New evidence on cross-country differences in job satisfaction using anchoring vignettes," *Labour Economics*, 15(1), 96–117.

[12] Murray, C. J., A. Tandon, J. A. Salomon, C. D. Mathers, and R. Sadana (2002): "Cross-population comparability of evidence for health policy," *Health systems performance assessment: debates, methods and empiricism*, pp. 705–713.

[13] Peracchi, F., and C. Rossetti (2013): "The heterogeneous thresholds ordered response model: identification and inference," *Journal of the Royal Statistical Society Series A*, 176(3), 703–722.

[14] Pudney, S., and M. Shields (2000): "Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model," *Journal of Applied Econometrics*, 15, 367399.

[15] Rice, N., S. Robone, and P. C. Smith (2012): "Vignettes and health systems responsiveness in cross-country comparative analyses," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2), 337–369.

[16] Terza, J. (1985): "Ordered Probit: A Generlization," *Communications in Statistics - A. Theory and Methods*, 14, 1–11.

5. Appendix A

Anchoring vignettes for self-assessed health

(Note that vignettes were gender specific)

Vignette 1:

Rob (Rebecca) is able to walk distances of up to 500 metres without any problems but feels puffed and tired after walking one kilometre or walking up more than one flight of stairs. He (she) is able to wash, dress and groom himself/herself, but it requires some effort due to an injury from an accident one year ago. His (her) injury causes him (her) to stay home from work or social activities about once a month. Rob (Rebecca) feels some stiffness and pain in his (her) right shoulder most days however his (her) symptoms are usually relieved with low doses of medication, stretching and massage. He (she) feels happy and enjoys things like hobbies or social activities around half of the time. The rest of the time he (she) worries about the future and feels depressed a couple of days a month.

Vignette 2:

Chris (Christine) is suffering from an injury which causes him (her) a considerable amount of pain. He (she) can walk up to a distance of 50 metres without any assistance, but struggles to walk up and down stairs. He (she) can wash his (her) face and comb his (her) hair, but has difficulty washing his (her) whole body without help. He (she) needs assistance with putting clothes on the lower half of his (her) body. Since having the injury Chris (Christine) can no longer cook or clean the house himself (herself), and needs someone to do the grocery shopping for him (her). The injury has caused him (her) to experience back pain every day and he (she) is unable to stand or sit for more than half an hour at a time. He (she) is depressed nearly every day and feels hopeless. He (she) also has a low self-esteem and feels that he (she) has become a burden.

Vignette 3:

Kevin (Heather) walks for one to two kilometres and climbs three flights of stairs every day without tiring. He (she) keeps himself neat and tidy and showers and dresses himself each morning in under 15 minutes. He (she) works in an office and misses work one or two days per year due to illness. Kevin (Heather) has a headache once every two months that is relieved by taking over-the-counter pain medication. He (she) remains happy and cheerful most of the time, but once a week feels worried about things at work. He (she) feels very sad once a year but is able to come out of this mood within a few hours.

Table 1: Descriptive statistics

	HILDA				VIGNETTES SAMPLE				Difference	
Variable	Mean	Std. Dev	Min.	Max.	Mean	Std. Dev	Min.	Max.	Z	P-value
					Full sample					
		(N = 12009)				(N = 5034)				
Self-assessed health	2.469	0.944	0	4						
Explanatory variables										
Age	40.78	13.78	18	65	41.39	13.28	18	65	-2.66*	0.008
Female	0.533	0.499	0	1	0.517	0.500	0	1	1.96	0.056
Tertiary education	0.617	0.486	0	1	0.694	0.461	0	1	-9.55	0.000
Year 12	0.178	0.382	0	1	0.149	0.356	0	1	4.60	0.000
Less than year 12	0.205	0.404	0	1	0.157	0.364	0	1	7.28	0.000
Employed	0.744	0.436	0	1	0.672	0.469	0	1	9.58	0.000
Not in labour force	0.212	0.409	0	1	0.224	0.417	0	1	-1.74	0.082
Unemployed	0.044	0.206	0	1	0.103	0.305	0	1	-14.63	0.000
Married	0.634	0.482	0	1	0.589	0.492	0	1	5.52	0.000
Migrant	0.210	0.407	0	1	0.246	0.430	0	1	-5.17	0.000
Vignettes										
V1					3.132	0.872	0	4		
V2					1.442	0.815	0	4		
V3					0.361	0.784	0	4		
					Vignette weighted sample					
		(N = 12009)				(N = 5034)				
Self-assessed health	2.469	0.944	0	4						
Explanatory variables										
Age	40.78	13.78	18	65	40.95	13.71	18	65	-0.736*	0.462
Female	0.533	0.499	0	1	0.536	0.499	0	1	-0.36	0.720
Tertiary education	0.617	0.486	0	1	0.630	0.483	0	1	-1.60	0.111
Year 12	0.178	0.382	0	1	0.173	0.379	0	1	0.78	0.435
Less than year 12	0.205	0.404	0	1	0.197	0.398	0	1	1.19	0.236
Employed	0.744	0.436	0	1	0.755	0.430	0	1	-1.51	0.132
Not in labour force	0.212	0.409	0	1	0.205	0.404	0	1	1.24	0.306
Unemployed	0.044	0.206	0	1	0.040	0.196	0	1	1.18	0.239
Married	0.634	0.482	0	1	0.633	0.482	0	1	0.12	0.902
Migrant	0.210	0.407	0	1	0.197	0.398	0	1	1.91	0.056
Vignettes										
V1					3.149	0.853	0	4		
V2					1.476	0.837	0	4		
V3					0.385	0.828	0	4		

* Comparison of means based on t-statistic with 17041 degrees of freedom.

Table 2: Ordered Probit and HOPIT Results

	Ordered Probit		HOPIT			
	HILDA sample		Full sample		Weighted sample	
	Coefficient	s.e.	Coefficient	s.e.	Coefficient	s.e.
Outcome equation						
Constant			2.688***	(0.187)	2.818***	(0.195)
Age/100	-4.542***	(0.514)	-1.287	(0.790)	-1.922***	(0.800)
$(\text{Age}/100)^2$	3.633***	(0.614)	1.236	(0.939)	2.069***	(0.953)
Female	0.086***	(0.020)	0.067**	(0.030)	0.109***	(0.031)
Tertiary education	0.277***	(0.026)	0.307***	(0.041)	0.357***	(0.040)
Year 12	0.249***	(0.033)	0.310***	(0.052)	0.365***	(0.051)
Employed	0.374***	(0.048)	0.278***	(0.062)	0.224***	(0.076)
Not in labour force	-0.117**	(0.051)	-0.290***	(0.067)	-0.334***	(0.081)
Married	0.129***	(0.021)	0.034	(0.032)	0.028	(0.033)
Migrant	0.050**	(0.024)	0.210***	(0.036)	0.210***	(0.038)
Vignettes constants						
V1			3.770***	(0.162)	3.819***	(0.171)
V2			1.714***	(0.159)	1.756***	(0.168)
V3			-0.032	(0.158)	-0.018	(0.166)
Threshold equations						
$\mu^{j=0}$						
Constant	-2.672***	(0.108)				
Age/100			1.991**	(0.775)	1.984***	(0.794)
$(\text{Age}/100)^2$			-1.651*	(0.912)	-1.485	(0.937)
Female			0.070**	(0.029)	0.078***	(0.030)
Tertiary education			0.107***	(0.040)	0.138***	(0.039)
Year 12			0.099*	(0.052)	0.118**	(0.051)
Employed			0.033	(0.050)	-0.021	(0.076)
Not in labour force			0.141**	(0.057)	0.151*	(0.080)
Married			-0.155***	(0.031)	-0.158***	(0.032)
Migrant			0.142***	(0.034)	0.153***	(0.037)
$\mu^{j=1}$						
Constant	-1.707***	(0.106)	-0.020	(0.139)	0.123	(0.143)
Age/100			1.244*	(0.680)	0.391	(0.680)
$(\text{Age}/100)^2$			-0.773	(0.789)	0.104	(0.791)
Female			-0.094***	(0.025)	-0.069***	(0.026)
Tertiary education			-0.014	(0.032)	0.008	(0.031)
Year 12			-0.019	(0.043)	0.030	(0.042)
Employed			-0.117***	(0.042)	-0.075	(0.060)
Not in labour force			-0.200***	(0.047)	-0.075	(0.064)
Married			-0.002	(0.026)	-0.001	(0.027)
Migrant			0.025	(0.028)	0.003	(0.030)
$\mu^{j=2}$						
Constant	-.583***	(0.105)	0.203*	(0.110)	0.227**	(0.111)
Age/100			-0.103	(0.551)	0.026	(0.553)
$(\text{Age}/100)^2$			-0.072	(0.653)	-0.164	(0.655)
Female			0.012	(0.021)	0.023**	(0.021)
Tertiary education			-0.105***	(0.026)	-0.113***	(0.025)
Year 12			-0.058*	(0.035)	-0.084**	(0.034)
Employed			-0.021	(0.040)	-0.076	(0.047)
Not in labour force			-0.093**	(0.044)	-0.161***	(0.051)
Married			0.057**	(0.022)	0.053**	(0.022)
Migrant			0.004	(0.025)	0.015	(0.026)
$\mu^{j=3}$						
Constant	0.620	(0.105)	0.026	(0.109)	-0.035	(0.112)
Age/100			0.419	(0.543)	-0.125	(0.547)
$(\text{Age}/100)^2$			0.849	(0.651)	0.525**	(0.656)
Female			0.025	(0.020)	0.025	(0.020)
Tertiary education			0.074**	(0.030)	0.082***	(0.029)
Year 12			0.104***	(0.037)	0.119***	(0.036)
Employed			0.075*	(0.044)	0.052	(0.055)
Not in labour force			-0.095*	(0.050)	-0.109*	(0.060)
Married			0.059***	(0.022)	0.071***	(0.023)
Migrant			-0.042*	(0.025)	-0.025	(0.026)
1/s			0.893***	(0.011)	0.868***	(0.011)

Table 3: Ordered Probit and HOPIT Results - partial effects

	Ordered Probit		HOPIT			
	HILDA sample		Full sample		Weighted sample	
	Coefficient	s.e.	Coefficient	s.e.	Coefficient	s.e.
	Partial effects of reporting *poor* health					
Outcome equation						
Constant			-0.082***	(0.007)	-0.082***	(0.007)
Age/100	0.200***	(0.025)	0.099***	(0.028)	0.113***	(0.027)
$(Age/100)^2$	-0.160***	(0.028)	-0.088***	(0.032)	-0.103***	(0.032)
Female	-0.004***	(0.001)	-0.00009	(0.001)	-0.001	(0.001)
Tertiary education	-0.012***	(0.001)	-0.006***	(0.001)	-0.006***	(0.001)
Year 12	-0.011***	(0.002)	-0.006***	(0.002)	-0.007***	(0.002)
Employed	-0.016***	(0.002)	-0.007***	(0.002)	-0.007***	(0.003)
Not in labour force	0.005***	(0.002)	0.013***	(0.002)	0.014***	(0.003)
Married	-0.006***	(0.001)	-0.006***	(0.001)	-0.005***	(0.001)
Migrant	-0.002**	(0.001)	-0.002*	(0.001)	-0.002*	(0.001)
	Partial effects of reporting *excellent* health					
Outcome equation						
Constant			0.506***	(0.031)	0.506***	(0.031)
Age/100	-0.890***	(0.101)	-0.850***	(0.143)	-0.873***	(0.143)
$(Age/100)^2$	0.712***	(0.121)	0.592***	(0.172)	0.615***	(0.172)
Female	0.017***	(0.004)	0.013**	(0.005)	0.012***	(0.005)
Tertiary education	0.054***	(0.005)	0.050***	(0.008)	0.048***	(0.008)
Year 12	0.049***	(0.007)	0.036***	(0.009)	0.034***	(0.009)
Employed	0.073***	(0.009)	0.065***	(0.013)	0.073***	(0.014)
Not in labour force	-0.023***	(0.010)	0.003	(0.014)	0.012	(0.016)
Married	0.025**	(0.004)	0.012*	(0.006)	0.010*	(0.006)
Migrant	0.010**	(0.005)	0.017***	(0.007)	0.013**	(0.007)

* Significant at 10%; ** significant at 5%; *** signifcant at 1%.

The Bankwest Curtin Economics Centre is an independent economic and social research organisation located within the Curtin Business School at Curtin University. The Centre was established in 2012 through the generous support from Bankwest (a division of the Commonwealth Bank of Australia), with a core mission to undertake high quality, objective research on the key economic and social issues of relevance to Western Australia.

The Centre's research and engagement activities are designed to influence economic and social policy debates in state and Federal Parliament, regional and national media, and the wider Australian community. Through high quality, evidence-based research and analysis, our research outcomes inform policy makers and commentators of the economic challenges to achieving sustainable and equitable growth and prosperity both in Western Australia and nationally.

The Centre capitalises on Curtin University's reputation for excellence in economic modelling, forecasting, public policy research, trade and industrial economics and spatial sciences. Centre researchers have specific expertise in economic forecasting, quantitative modelling, microdata analysis and economic and social policy evaluation.

A suite of tailored and national economic models and methods are maintained within the Centre to facilitate advanced economic policy analysis: these include macroeconomic and time series models, micro(simulation) models, computable general equilibrium (CGE) models, spatial modelling methods, economic index analysis, and behavioural modelling methods.

CONTACT

Bankwest Curtin Economics Centre
Curtin University
Kent Street Bentley WA 6102
GPO Box U1987 Perth WA 6845

Tel: +61 8 9266 2873

bcec@curtin.edu.au
business.curtin.edu.au/bcec