Using externally collected vignettes to account for reporting heterogeneity in survey self-assessment

MARK N HARRIS†, RACHEL J KNOTT‡*, PAULA K LORGELLY§, and NIGEL RICE††

†School of Economics, Finance and Property, Curtin University, Perth, Australia
‡Centre for Health Economics, Monash University, Melbourne, Australia
§Department of Applied Health Research, University College London, London, UK
††Centre for Health Economics & Department of Economics and Related Studies, University of York, York, UK

*Corresponding author: rachel.knott@monash.edu +613 9905 0772
Building H, Monash University, 900 Dandenong Rd, Caulfield East VIC 3145, Australia

Abstract

The anchoring vignette approach has grown in popularity as a method to adjust for reporting heterogeneity in self-reported survey questions, 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-reported health in a large household survey, how externally collected vignettes can be used to adjust for reporting heterogeneity in datasets that have not included anchoring vignettes. Given that self-reported measures are an important facet of social, health and economic research, we anticipate the approach described will lead to new applications of the anchoring vignette methodology.

Keywords: Anchoring vignettes; reporting heterogeneity; differential item functioning; self-reported measures; survey measures.

JEL: I1, C1, A1

Acknowledgements and funding sources

This research was funded by an Australian Research Council Discovery Project Grant (DP110101426), a Bankwest-Curtin Economics Centre grant, and a Monash Business School, Monash University grant.

Conflict of interest

The authors have no conflicting interests to declare.

1. Introduction and Background

The use of self-reported questions to gather information about people's circumstances, preferences or beliefs are ubiquitous in social surveys. However 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 or reference points that people use to evaluate themselves. Consequently, two individuals with identical levels of the underlying construct of interest, may rate themselves differently in response to a survey question. This issue, a type of reporting heterogeneity commonly referred to as *differential item functioning* (DIF) (Murray et al., 2002), can lead to bias when drawing inter-personal comparisons; such that comparative analyses undertaken using self-reported data may produce biased results, and the implications and policy advice that may be forthcoming are likely to be erroneous. For example, Knott et al. (2017a) discusses these implications with respect to measuring health and quality of life, noting that they may not just be problematic for analyses of self-reports but also for measures which rely on preference elicitation.

A methodology for overcoming DIF is the anchoring vignette approach (King et al., 2004), a survey tool which has grown in popularity over the past decade in the literature on health (Bago d'Uva et al., 2008, Grol-Prokopczyk et al., 2011, Knott et al., 2017b, Mu, 2014), work disability (Kapteyn et al., 2007), political efficacy (King et al., 2004), and wellbeing (Angelini et al., 2014, Bertoni, 2015). 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.

Although this method has proved useful, its application to date has been limited to datasets where vignettes have been collected alongside self-reports of the construct of interest. Here we illustrate how to use vignette responses collected externally to the main survey containing the self-reports, using a generic measure of self-reported health (SRH) as an application.

2. Methods

Our approach involves a simple extension of the hierarchical ordered probit (HOPIT) with vignettes as proposed by King et al. (2004). The likelihood function consists of two components, namely the component relating to the self-assessment, $L_{i,HOPIT}$ (the structural component), and the vignettes component $L_{q,V}$, which are linked only by the common parameters of the boundary or threshold equations (see Appendix A for further details of the $HOPIT$ model). For ease of notation, assume that there is only one vignette assessment, then the likelihood function could be written

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

where $i = 1, \ldots, N$ indexes the main sample and $q = 1, \ldots, Q$ indexes the vignettes sample. The only requirement, other than the implicit assumption that the DIF problem is the same across the two samples (note that the assumption of response consistency, RC requires

that each individual i must use the same response scale to evaluate both their own self-reports, and the experience(s) described by the vignette(s)), is that the vector of variables included in the boundary (or threshold) equation (term $\mathbf{z}$ of Appendix A), are available in both the vignette sample and the main sample of interest. Additionally, as in the case where vignettes are collected alongside self-reports, we need to assume both RC and vignette equivalence, VE, hold.[1] Imposing common support across the two samples in the covariates of the boundary equations ($\mathbf{z}$), 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.[2]

Estimation of the $HOPIT$ model is undertaken by maximising the likelihood in equation (1). 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; thus, balance is maintained across the characteristics determining reporting behaviour. 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, $\mathbf{z}$ if respondents to

[1] Along with the common modeling assumptions of Normality, homoscedasticity, etc. See Appendix A and King et al. (2004) for details about the assumptions of RC and VE.

[2] It is worth noting that if desired, it is also possible (but not necessary) to include additional variables in the structural equation relating to $L_{i,HOPIT}$ (term $\mathbf{x}$ of Appendix A), such as additional variables affecting the self-reported outcome of interest that are available in the main sample but not the vignettes sample.

the vignettes are not fully representative of the sample of individuals completing the self-assessment drawn from the main survey of interest. Since reporting behaviour is identified on vignette sample respondents, and imposed on the main sample via the assumption of RC, a lack of balance in the characteristics determining reporting behaviour may lead to biases in the estimated coefficients of the structural equation (term β_y of Appendix A). In such circumstances, the respondents can be weighted such that balance in the characteristics entering the boundary equations, $\mathbf{z}$, 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 main survey samples falling into each distinct strata, with strata defined by the multivariate distribution of the set of reporting behaviour characteristics, $\mathbf{z}$, under consideration. A small covariate set, particularly those requiring coarsening, together with common support across the set of reporting behaviour 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 $Q_j, \left(j = 1, \ldots, J, \text{with } \sum_{j=1}^{J} Q_j = Q\right)$ and the corresponding number in the main sample

is N_j, then the likelihood in equation (1) can be weighted such that:

$$\ln L = \sum_{i=1}^{N} \ln L_{i,HOPIT} + \sum_{q=1}^{Q} w_{ij} \ln L_{q,V}, \tag{2}$$

where $w_{ij} = \frac{N_j}{N}\frac{Q}{Q_j}, \quad j = 1, \ldots, J$. Maximising the likelihood in equation (2) imposes reporting behaviour identified on a sample displaying greater balance across the characteristics thought, a priori, to be important drivers of reporting styles, which strengthens claims for RC.

3. Empirical example

We illustrate the approach empirically by correcting for DIF in SRH in the widely used $HILDA$ survey, using vignette responses collected in a bespoke online survey. We focus on the generic SRH question in $HILDA$ which asks respondents: *In general, would you say your health is excellent, very good, good, fair or poor?*. Three vignettes were included describing health states of differing levels of severity (see online appendix - Part B). The categories available to respondents when rating the vignettes are the same as those available for SRH in $HILDA$. Importantly, the online survey also contained a set of questions on socio-demographic characteristics of respondents which correspond to questions asked in $HILDA$. For further details about the online survey, and for descriptive statistics of each sample, refer to the online appendix - Part C.

We consider two estimation samples. The first, herein the *full sample*, simply pools the two datasets (main and external) - $HILDA$ and the vignette sample. The second, referred

to as the *weighted sample*, weights the vignettes sample in the likelihood as outlined in equation (2) (see Part D of the online appendix for details on the construction of weights for this sample). Our interest is in estimating the determinants of SRH, but adjusted for observed reporting behaviour. For simplicity, and following the predominant empirical literature, we specify that the variables in the structural component of $lnL_{i,HOPIT}$ $(\mathbf{x})$ are the same as the reporting behaviour component $(\mathbf{z})$ and adopt a standard set of demographic variables similar to those used elsewhere to model SRH (Contoyannis et al., 2004, Balia and Jones, 2008) (see Table C1 of online appendix). An ordered probit, (OP), 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 1 contain parameter estimates and corresponding standard errors for an OP model 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 ($\leq 5\%$). Space constraints do not allow for a detailed discussion, but it is important to point out that since the OP model fails to adjust for differences in reporting behaviour, the estimated effects represent composite parameters reflecting differences in true underlying health status together with differences in reporting styles.

Table 1 also presents results of the $HOPIT$ model. The fourth and fifth columns show coefficient estimates and standard errors for the *full sample*. Note the high significance levels for parameters in the threshold equations, indicating a significant degree of DIF across the reporting behaviour 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. These findings 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, we 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). Similar results can be seen across coefficients for age and employment status, while the effects for education and migrant status become more prominent under the $HOPIT$ model.

Columns 6 and 7 of Table 1 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 for a number of covariates. 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*. Similar effects can be seen for other variables, particularly across

gender, education and labour market status.

4. Concluding remarks

While anchoring vignettes have been widely used to identify and correct for DIF, their use in the literature thus far has been limited to analyses of datasets containing both self-assessments and vignette questions (e.g. Bago d'Uva et al. (2008), Bertoni (2015), Knott et al. (2017b)). 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 how information on vignettes can be incorporated without losing the ability to generate inference on the target survey of interest, (which, as in the example provided, may be a population representative household survey), through weighting to create better balance in covariates determining reporting behaviour.

Although our empirical example considers self-reported health, 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 this paper. Alternatively, certain waves of existing household surveys already contain vignette components which might be used to externally adjust for DIF.[3] Given that self-reports to survey ques-

[3]For instance, SHARE (wave 1 and 2) and ELSA (wave 3) include vignettes on health and health limita-

tions are an important facet of health, economic and social science research, we anticipate this approach will lead to new applications of the anchoring vignette methodology.

tions; 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.

References

V. Angelini, D. Cavapozzi, L. Corazzini, and O. Paccagnella. 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, 2014.

N. Au and P. K. Lorgelly. Anchoring vignettes for health comparisons: an analysis of response consistency. *Quality of Life Research*, 23(6):1721–1731, 2014.

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

S. Balia and A. M. Jones. Mortality, lifestyle and socio-economic status. *Journal of Health Economics*, 27(1):1–26, 2008.

M. Bertoni. Hungry today, unhappy tomorrow? Childhood hunger and subjective wellbeing later in life. *Journal of Health Economics*, 40:40–53, 2015.

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

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

A. Kapteyn, J. P. Smith, and A. Van Soest. Vignettes and self-reports of work disability in the United States and the Netherlands. *The American Economic Review*, pages 461–473, 2007.

G. King, C. Murray, J. Salomon, and A. Tandon. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98 (1):191–207, 2004.

R. J. Knott, N. Black, B. Hollingsworth, and P. K. Lorgelly. Response-scale heterogeneity in the EQ-5D. *Health Economics*, 26(3):387–394, 2017a.

R. J. Knott, P. K. Lorgelly, N. Black, and B. Hollingsworth. Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. *Social Science & Medicine*, 190:247–255, 2017b.

R. Mu. Regional disparities in self-reported health: Evidence from Chinese older adults. *Health Economics*, 23(5):529–549, 2014.

C. J. Murray, A. Tandon, J. A. Salomon, C. D. Mathers, and R. Sadana. Cross-population comparability of evidence for health policy. *Health Systems Performance Assessment: Debates, Methods and Empiricism*, pages 705–713, 2002.

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

S. Pudney and M. Shields. Gender, race, pay and promotion in the British nursing profession: Estimation of a generalized ordered probit model. *Journal of Applied Econometrics*, 15:367399, 2000.

TABLE 1
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)
$(Age/100)^2$	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)
$(Age/100)^2$			-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)
$(Age/100)^2$			-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)
$(Age/100)^2$			-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)
$(Age/100)^2$			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)

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

Online Appendix

Part A: The traditional HOPIT model

The HOPIT model consists of two components; namely a self-assessment component (the structural equation), and a component relating to reporting behaviour, which is identified by the vignettes. Formally, assuming our measure of interest is self-reported health, suppose the true latent health of individual i, y^*, is a linear function (in unknown parameters, β_y) of observed characteristics $\mathbf{x}$ and a disturbance term (unrelated to any observed heterogeneity in the model), ε_y;

$$y^* = \mathbf{x}'\boldsymbol{\beta}_y + \varepsilon_y, \tag{3}$$

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

$$y = \left\{ \begin{array}{ll} j & \text{if } \mu_i^{j-1} \leq y^* < \mu_i^j \quad for\ j = 0, \ldots, J-1. \end{array} \right. \tag{4}$$

(where to avoid clutter in notation, subscripts denoting and individual have been omitted where not strictly necessary). The boundary (or threshold) parameters, μ_i^j, depend on a set of observed characteristics, $\mathbf{z_i}$, such that

$$\begin{array}{rcl} \mu_i^0 & = & \mathbf{z}_i'\boldsymbol{\gamma}_0 \\ \mu_i^j & = & \mu_i^{j-1} + \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}_j\right) \\ & \vdots & \end{array} \tag{5}$$

Anchoring vignettes aid in identification of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_\mathbf{y}$, which otherwise would not be separately identifiable for variables that appear in both $\mathbf{x}$ and $\mathbf{z}$.[4]

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 $\varepsilon \sim N\left(0, \sigma_k^2\right)$ and orthogonal to all observed covariates in the model. Usually the simplifying assumption that

[4]unless, for example, exclusion restrictions are used, such that $\mathbf{x}$ and $\mathbf{z}$ are distinct vectors (Pudney and Shields, 2000). However exclusion restricitions are often difficult to justify in practice.

$\sigma_k^2 = \sigma_v^2 \ \forall k$ 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 $\mathbf{z}_i \equiv \mathbf{x}_i$).

Identification of the parameters relies on the 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 (King et al., 2004). 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.[5] 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 ($\ln L_{V,k}$):

$$\ln L = \ln L_{HOPIT} + \sum_k \ln L_{V,k}, \tag{6}$$

where the first term is a function of $\boldsymbol{\beta}$ and $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$ and the second a function of α_k, σ_v and $\mu_i^j\left(\boldsymbol{\gamma}_j\right)$. Thus 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.

[5] A useful summary of the various restriction strategies available to the researcher in the presence of vignettes, is given by Peracchi and Rossetti Peracchi and Rossetti (2013).

Part B: Anchoring vignettes for self-reported health

For details of the construction of the vignettes, refer to (Au and Lorgelly, 2014). Note that vignettes were sex-matched, according to the gender of respondents.

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.

Part C: Sample information and characteristics

The online survey was conducted in April 2014 and in August 2015 and targeted a representative sample of Australians aged 18-65. Descriptive statistics for the online (external vignette) sample and the (main) HILDA sample, corresponding to wave 13 (*i.e.*, 2013) for HILDA and those aged between 18 and 65, so as to be comparable with the external sample are presented below in Table B1. 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 in this paper, while there is imbalance across the two samples, there is common support over the set of characteristics.

TABLE C1

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 (proportions) based on t-statistic with 17041 degrees of freedom.

Part D: Weighting of data in empirical example

Weighting was achieved in the empirical example by firstly coarsening age into 5-year age groups and secondly, considering the distinct strata formed from the set of coarsened and binary variables. 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 vignette sample size of 5034. Of the 720 possible strata,[6] 504 were populated by both vignette and $HILDA$ sample members. These are the vignette respondents to which the weighting procedure outlined in equation (2) applied. A further 49 strata contained only vignette respondents, and 94 only $HILDA$ respondents. To maintain the sample size these two sets of 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 C1; there is improved balance across all covariates. This is supported by formal statistical tests of the difference in means and proportions (final columns of Table C1).

[6] A result of there being 10 age groups; 2 groups each for gender, marital status and migrant status; and 3 groups each for education and labour force status.

Online Appendix

Part A: The traditional HOPIT model

The HOPIT model consists of two components; namely a self-assessment component (the structural equation), and a component relating to reporting behaviour, which is identified by the vignettes. Formally, suppose the true latent health of individual i, y^*, is a linear function (in unknown parameters, β_y) of observed characteristics $\mathbf{x}$ and a disturbance term (unrelated to any observed heterogeneity in the model), ε_y;

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

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

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

(where to avoid clutter in notation, subscripts denoting and individual have been omitted where not strictly necessary). The boundary (or threshold) parameters, μ_i^j, depend on a set of observed characteristics, $\mathbf{z_i}$, such that

$$\begin{array}{rcl} \mu_i^0 & = & \mathbf{z}_i'\gamma_0 \\ \mu_i^j & = & \mu_i^{j-1} + \exp\left(\mathbf{z}_i'\gamma_j\right) \\ & \vdots & \end{array} \tag{3}$$

Anchoring vignettes aid in identification of γ_0 and $\beta_{\mathbf{y}}$, which otherwise would not be separately identifiable for variables that appear in both $\mathbf{x}$ and $\mathbf{z}$.[1]

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 $\varepsilon \sim N\left(0, \sigma_k^2\right)$ 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 $\mathbf{z}_i \equiv \mathbf{x}_i$).

Identification of the parameters relies on the 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 (**?**). The RC assumption amounts to restricting all coefficients in all of the reporting parts of the model (the boundary parameters: $\gamma_j \forall j$) to be the same; *i.e.*, γ 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.[2] 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 ($\ln L_{V,k}$):

[1] unless, for example, exclusion restrictions are used, such that $\mathbf{x}$ and $\mathbf{z}$ are distinct vectors (**?**). However exclusion restricitions are often difficult to justify in practice.

[2] A useful summary of the various restriction strategies available to the researcher in the presence of vignettes, is given by Peracchi and Rossetti **?**.

$$\ln L = \ln L_{HOPIT} + \sum_k \ln L_{V,k}, \tag{4}$$

where the first term is a function of β and $\mu_i^j(\gamma_j)$ and the second a function of α_k, σ_v and $\mu_i^j(\gamma_j)$. Thus these two components are linked through the common boundary parameters $\mu_i^j(\gamma_j)$, and so do not factorise into two independent models.

Part B: Anchoring vignettes for self-reported health

For details of the construction of the vignettes, refer to (**?**). Note that vignettes were sex-matched, according to the gender of respondents.

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.

Part C: Sample information and characteristics

The online survey was conducted in April 2014 and in August 2015 and targeted a representative sample of Australians aged 18-65. Descriptive statistics for the online (external vignette) sample and the (main) HILDA sample, corresponding to wave 13 (*i.e.*, 2013) for HILDA and those aged between 18 and 65, so as to be comparable with the external sample are presented below in Table B1. 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 in this paper, while there is imbalance across the two samples, there is common support over the set of characteristics.

TABLE C1

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 (proportions) based on t-statistic with 17041 degrees of freedom.

Part D: Weighting of data in empirical example

Weighting was achieved in the empirical example by firstly coarsening age into 5-year age groups and secondly, considering the distinct strata formed from the set of coarsened and binary variables. 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 vignette sample size of 5034. Of the 720 possible strata,[3] 504 were populated by both vignette and $HILDA$ sample members. These are the vignette respondents to which the weighting procedure outlined in equation (**??**) applied. A further 49 strata contained only vignette respondents, and 94 only $HILDA$ respondents. To maintain the sample size these two sets of 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 C1; there is improved balance across all covariates. This is supported by formal statistical tests of the difference in means and proportions (final columns of Table C1).

[3] A result of there being 10 age groups; 2 groups each for gender, marital status and migrant status; and 3 groups each for education and labour force status.