d'Uva, Teresa Bago; Lindeboom, Maarten; O'Donnell, Owen; van Doorslaer, Eddy

Working Paper

Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity

Tinbergen Institute Discussion Paper, No. 09-091/3

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: d'Uva, Teresa Bago; Lindeboom, Maarten; O'Donnell, Owen; van Doorslaer, Eddy (2009) : Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity, Tinbergen Institute Discussion Paper, No. 09-091/3, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/86789

TI 2009-091/3
Tinbergen Institute Discussion Paper

Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity

Teresa Bago d'Uva[a, b,c]
Maarten Lindeboom[b,c,d]
Owen O'Donnell[c,e,f]
Eddy van Doorslaer[a,b,c]

[a] *Erasmus University Rotterdam;*
[b] *Tinbergen Institute;*
[c] *Netspar;*
[d] *VU University Amsterdam;*
[e] *University of Macedonia;*
[f] *University of Lausanne.*

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity

Teresa Bago d'Uva[abc*], Maarten Lindeboom[bcd], Owen O'Donnell[cef], Eddy van Doorslaer[abc]

a Erasmus University Rotterdam, b Tinbergen Institute, c Netspar, d Free University Amsterdam, e University of Macedonia, f University of Lausanne

3 November 2009

Abstract

Anchoring vignettes are increasingly used to identify and correct heterogeneity in the reporting of health, work disability, life satisfaction, political efficacy, etc. with the aim of improving interpersonal comparability of subjective indicators of these constructs. The method relies on two assumptions: *vignette equivalence* – the vignette description is perceived by all to correspond to the same state; and, *response consistency* - individuals use the same response scales to rate the vignettes and their own situation. We propose tests of these assumptions. For *vignette equivalence*, we test a necessary condition of no systematic variation with observed characteristics in the perceived difference in states corresponding to any two vignettes. To test *response consistency* we rely on the assumption that objective indicators fully capture the covariation between the construct of interest and observed individual characteristics, and so offer an alternative way to identify response scales, which can then be compared with those identified from the vignettes. We also introduce a weaker test that is valid under a less stringent assumption. We apply these tests to cognitive functioning and mobility related health problems using data from the English Longitudinal Survey of Ageing. *Response consistency* is rejected for both health domains according to the first test, but the weaker test does not reject for cognitive functioning. The necessary condition for *vignette equivalence* is rejected for both health domains. These results cast some doubt on the validity of the vignettes approach, at least as applied to these health domains.

JEL Classification*: C35, C42, I12*

Keywords: Reporting heterogeneity, Survey methods, Vignettes, Health, Cognition

***Correspondence to:** bagoduva@ese.eur.nl

Acknowledgements: This paper derives from the NETSPAR funded project "Income, health and work across the life cycle". Data were obtained from the ERSC Data Archive.

1 Introduction

Survey data are a valuable source of information on measures and determinants of individual welfare in the form of subjective assessments of life satisfaction, job satisfaction, health, political efficacy, quality of public services, etc. While these data are increasingly being used productively in economics and other social sciences, there is a persistent concern that reporting differences impede the interpersonal comparability of inherently subjective measures. A proposed solution to this measurement error problem is to anchor an individual's assessment of her own situation on her rating of a vignette description of a hypothetical situation that is fixed for all respondents (King et al, 2004). Since the vignette is fixed, variation in its rating identifies reporting heterogeneity and this can then be purged from the individual's subjective assessment of her own situation. For example, individuals may be asked to rate vignette descriptions of the functioning of hypothetical individuals on the same scale as their own health.

The anchoring vignettes approach relies on two identifying assumptions. First, *vignette equivalence* requires that all individuals perceive the vignette description as corresponding to a given state of the same underlying construct. So, for example, a health description must be perceived by all respondents as corresponding to a given level of functioning on the same unidimensional health scale. If this did not hold, then one could not attribute variation in ratings of a given vignette to reporting heterogeneity. The assumption may be violated if vignette descriptions are incomplete and/or equivocal and groups of individuals complement those descriptions in different ways. Second, *response consistency* requires that individuals use the same response scales to rate the vignettes and their own situation. If this did not hold, then information from the vignette responses would not be useful in identifying, and so correcting for, the reporting heterogeneity that confounds interpersonal comparability of the welfare indicator,

or determinant, of interest. This assumption will not hold, for example, if individuals have incentives to misreport their own health – perhaps as justification for not working - but not that of hypothetical individuals portrayed by the vignettes. In this case, there would be systematic under-reporting of own health by employment status, which would not be captured by vignette ratings. The approach would then be powerless with respect to correcting the justification bias that has plagued estimates of the impact of health on labor market participation of older individuals (Stern, 1989; Bound, 1991).

In this paper, we propose and apply tests of the two assumptions underpinning the anchoring vignettes approach. The *response consistency* test is feasible when, in addition to the vignettes, data are available on objective indicators that are presumed to capture all variation in the construct of interest that is associated with observed individual characteristics. Under this assumption, any systematic variation in subjective assessments that remains after conditioning on the objective indicators can be attributed to reporting heterogeneity (Kerkhofs and Lindeboom, 1995; Kreider, 1999). Since reporting heterogeneity is identified in this case without imposing *response consistency,* this assumption can be tested. This involves testing whether the thresholds used by the individual to report on her own situation, which are identified from the objective indicators, are equal to those used to report on the vignettes. For *vignette equivalence,* we test a necessary condition of no systematic variation with observed individual characteristics in the perceived difference in states corresponding to any two vignettes. This test can be performed with any dataset containing at least two vignettes for a given construct and does not require objective measures.

Vignette equivalence has not previously been formally tested. Van Soest et al (2007) introduced a test of *response consistency* that, like ours, is based on comparison between

reporting thresholds identified from vignettes and an objective measure. Our test differs in that it enables the use of a battery of objective measures which is desirable when a single indicator is unlikely to capture all association between covariates and the construct of interest. We also introduce a weaker test that is valid even if the objective indicators do not capture all of the association between covariates and the construct of interest.

Vignettes are being included in a growing number of household surveys, including the *English Longitudinal Study of Ageing* (ELSA), the *Health and Retirement Study* (HRS), the *Survey of Health, Ageing and Retirement in Europe* (SHARE) and the World Health Organisation's *World Health Surveys* (WHS). Not surprisingly, applications of the methodology are increasing rapidly and now include comparisons of political efficacy (King et al, 2004), work disability (Kapteyn et al, 2007), job satisfaction (Kristensen and Johansson, 2008), life satisfaction (Christensen et al, 2006), health (Bago d'Uva et al, 2008a, 2008b) and health system responsiveness (Rice et al, 2008). Many of these reveal substantial reporting heterogeneity and therefore important impacts of vignette corrections on the comparisons of interest. But in the absence of validation of the method, uncertainty remains about the appropriateness and accuracy of such 'corrections'.

The vignettes method has attracted most interest from researchers working on health and disability, a consequence of the heavy reliance on self-reported measures of these concepts in survey analyses conducted in economics, social science and epidemiology. While self-assessed health (SAH) has been repeatedly proven to be a good predictor of mortality (Idler and Benyamini, 1997), there are nonetheless concerns that differences in conceptions of health in general and in expectations for own health, as well as incentives created by disability insurance eligibility criteria may lead to systematic differences in the reporting of health. This would bias

any analysis relying on interpersonal comparability of the measure (Butler et al, 1987; Bound, 1991). For example, Van Doorslaer et al (2004) argue that evidence of apparent pro-poor utilisation of physician services in Europe may be attributable to under-reporting of morbidity among the socially disadvantaged. This would also lead to under-estimation of socio-economic inequalities in health (Lindeboom and Van Doorslaer 2004; Etilé and Milcent, 2006; Bago d'Uva et al, 2008a, 2008b). Considerable attention has been paid to the potential for under-reporting of health, in response to disability insurance entitlement rules, to upwardly bias the estimated impact of health and downwardly bias the estimated impact of financial incentives on labor force participation (Stern, 1989; Bound, 1991; Kerkhofs and Lindeboom, 1995; Benitez-Silva et al., 1999; Kreider, 1999; Disney et al., 2006). Estimated associations between socio-economic factors and self-reported health will reflect differences in both true health and reporting propensities, and these two effects cannot be separately identified without additional information on either true health or reporting behavior.

We apply our tests of the validity of the vignettes methodology to two domains of health – cognitive functioning and mobility. These represent mental and physical dimensions of health, allowing comparison of the performance of the vignettes method in relation to quite different concepts, and each is an important determinant of the welfare of older individuals. Important for our test of *response consistency, w*ell validated tests exist for both dimensions of health and we observe these in our dataset. The data are from the third wave (2006-07) of ELSA, which is a representative sample of the 50+ population in England. These data provide not only vignette ratings for cognitive functioning and mobility (and four other health domains), but also objective indicators of each health concept. For cognitive functioning, we have a battery of measured tests of retrospective and prospective memory, and of executive functioning. For mobility, we have a

measurement of walking speed, indicators of Activities of Daily Living (ADLs), and of motor skills and strength. If conditional on these objective proxies true cognition/mobility does not vary with socio-demographics, then any remaining systematic variation in reported cognition/mobility can be attributed to reporting behavior. *Response consistency* can then be tested by comparing response scales identified in this way with those identified by the vignettes approach. Systematic differences in the rating of a single vignette may derive either from differential perceptions of the health level described by it — a violation of *vignette equivalence* — or from differences in reporting thresholds, which is what the approach aims at identifying. When at least two vignettes describing different states within the same health domain are available, a necessary condition for *vignette equivalence* is that the perceived difference between the levels of health represented by any two vignettes does not vary systematically across individuals. This can be tested through the null of no interactions between vignette dummies and covariates in the perceived health level of vignettes.

The test results obtained from the application cast some doubt on the validity of the vignettes approach, at least as applied to the health domains of cognitive functioning and mobility. *Response consistency* is rejected for both health domains by the stronger test, but the weaker test does not reject for cognitive functioning. The necessary condition for *vignette equivalence* is rejected for both health domains.

The remainder of the paper is organized as follows. In the next section, we explain how reporting heterogeneity is identified by anchoring vignettes and by conditioning on objective indicators. Section 3 presents the tests for *vignette equivalence* and *response consistency.* In section 4 we describe the data, in particular the vignettes and the objective indicators for

cognitive functioning and mobility. Results are presented in section 5 and the final section concludes.

2 Identification of reporting heterogeneity

This section explains how reporting heterogeneity is identified by means of anchoring vignettes and proxy objective indicators, and the assumptions required in each case. For ease of exposition and given the application that follows, we will refer to the underlying concept of interest as 'health'.

2.1 The identification problem

The researcher has categorical data on self-reported health H^S obtained from a question inviting the respondent to choose which of a number of categories best describes her functioning in a particular health domain, such as cognition and mobility in our application. It is assumed that these responses are generated by a corresponding latent true health variable H^*. It is common practice to model ordered responses in the following way:

$$H_i^* = \beta X_i + \varepsilon_i \tag{1a}$$

$$H_i^S = k \Leftrightarrow \tau_i^{k-1} \leq H_i^* < \tau_i^k \tag{1b}$$

where X_i is a vector of observed characteristics, ε_i is a random error term, k=1,…,K is a categorical description of health, $\tau^0 < \tau_i^1 < ... < \tau_i^{K-1} < \tau_i^K$, τ_i^0=-∞ and τ_i^K=∞.

Researchers are ultimately interested in the extent to which true health varies across populations or subgroups (the parameter vector β). The problem is that the relationship between H^* and H^S may not be constant across populations. For instance, an individual with a university

degree may report no problems remembering things, whereas another individual with exactly the same level of cognitive functioning but with only primary school education may report moderate problems remembering things. Unconditional comparison of H^S across populations would confound differences in true health with those in reporting behavior. A natural way to model reporting heterogeneity is by allowing the cut-points to be dependent on observed characteristics, adopting, for example, a linear specification: [1]

$$\tau_i^k = \gamma^k X_i \ . \qquad (1c)$$

Combining equations (1a), (1b) and (1c) results in the following probability of observing response category *k*, conditional on *X*:

$$\mathrm{P}\left[H_i^s = k \mid X_i \right] \ = \ \mathrm{F}\left[\left(\gamma^k - \beta \right) X_i \right] - \mathrm{F}\left[\left(\gamma^{k-1} - \beta \right) X_i \right],$$

where F(.) is the distribution function of the error term ε. From this it apparent that it is not possible to identify simultaneously all γ^k and β.[2] Identification of β separately from reporting heterogeneity can be achieved only with additional information either on reporting behavior (γ^k), which vignettes provide, or on true health (H^*) via proxy indicators.

2.2 Identifying reporting heterogeneity: Anchoring with vignettes

Vignettes are descriptions of hypothetical health states, which survey respondents are asked to rate on the same scale as they do their own health. Ratings are assumed to be generated by an unobserved latent variable corresponding to the perceived health state invoked by the vignette

[1] An alternative is to define the first cut-point as here but the following ones as: $\tau_i^k = \tau_i^{k-1} + \exp\left(X_i \gamma^k \right) \quad k = 2, \ldots, K-1$ (Kapteyn et al, 2007). This ensures increasing cut-points. In our application, this condition was always satisfied with the linear specification, which facilitates more direct interpretation of the effects on cut-points.

[2] Identification of a restricted model that arbitrarily excludes covariates from one cut-point is possible (Terza, 1985). This is limiting since it means that the estimated shift of the remaining cut-points is only identified relative to that from which covariates are excluded.

description. Crucial to the identification of reporting heterogeneity is the assumption that, apart from random measurement error, all individuals perceive a particular vignette *j* to be consistent with the same latent health level V_{ij}^*. If this holds, then all systematic association between individual characteristics and vignette ratings can be attributed to differential reporting of a given state of health. More formally, the *vignette equivalence* assumption implies that the density function *f*(.) of perceived latent health invoked by each vignette description is independent of *X*,

$$f\left(V_j^* \middle| X\right) = f\left(V_j^*\right). \qquad \text{(A1)}$$

Then, the latent health of vignette *j* as perceived by individual *i* can be specified as an intercept $\left(\alpha_j\right)$ plus random measurement error $\left(\xi_{ij}\right)$,[3]

$$V_{ij}^* = \alpha_j + \xi_{ij} \quad , \qquad \text{(2a)}$$

and the respective observed categorical rating is assumed to be determined as follows:

$$V_{ij} = k \Leftrightarrow \nu_i^{k-1} \le V_{ij}^* < \nu_i^k \qquad \text{(2b)}$$

k=1,…,*K*, $\nu_i^0 < \nu_i^1 < \ldots < \nu_i^{K-1} < \nu_i^K$ and ν_i^0=-∞ , ν_i^K=∞. As before, differential reporting behavior is reflected in differences in the cut-points ν_i^k across individuals. Like in (1c), we can specify the cut-points as linear functions of the individual characteristics:

$$\nu_i^k = \gamma_\nu^k X_i \ . \qquad \text{(2c)}$$

Response consistency requires the cut-points of the own health component (1c) to be the same as those identified by the vignette component (2c),

[3] One could also allow the intercept to shift according to the gender of the vignette since respondents' perceptions of the health state described may be influenced by the gender of the hypothetical person. This is not relevant in our application since there is no variation in vignette gender within health domain.

$$\gamma^k = \gamma_v^k \quad k=1,\ldots K\text{-}1. \qquad \text{(A2)}$$

Under assumptions (A1) and (A2), the vignettes ratings can be used to identify reporting behavior (γ^k) via equations (2a)-(2c). This can then be imposed on equation (1c), making it possible to identify the health effects β in equation (1a). This was proposed by King et al (2004), who refer to the combined model composed of equations (1a)-(1c) and (2a)-(2c), together with assumed normality of the errors, as the Hierarchical Ordered Probit (HOPIT) model. We refer to the model composed by equations (2a)-(2c) as Model 1.

2.3 Identifying reporting heterogeneity: Proxying with objectives measures

An alternative approach is to consider a sufficiently comprehensive set of proxy indicators of health (H^0) that are believed to be insensitive to reporting behavior. These could include physical examinations, medical tests and scores from validated instruments. Let $h(.)$ be the density function of latent health, then reporting heterogeneity can be identified if:

$$h\left(H^* \middle| H^O, X\right) = h\left(H^* \middle| H^O\right) . \qquad \text{(A3)}$$

This conditional independence assumption implies that after conditioning on the set of proxy indicators, any remaining systematic variation in self-assessed health with respect to observed characteristics X is solely attributable to differences in reporting behavior (Kerkhofs and Lindeboom, 1995; Kreider, 1999; Lindeboom and Van Doorslaer, 2004). There is a potentially nonlinear relationship between latent true health and the proxy indicators, as follows:

$$H_i^* = g\left(H_i^O\right) + \eta_i \ , \qquad \text{(3a)}$$

where $g()$ is a sufficiently flexible function and η_i is a random error term. Then, a model of the relationships between true health (H^*), objectively measured health (H^O), reported health (H^S) and covariates (X) is given by (3a), (1b) and (1c), which we refer to as Model 2.

3 Tests of response consistency and vignette equivalence

3.1 Response consistency

Under assumption (A3), Model 2 (see Table 1) identifies the response scales used by the individual in reporting her own health. *Response consistency* (A2) can then be tested by comparing the estimates of the cut-points obtained from Model 1 with those obtained from Model 2. To implement this, we estimate a joint model composed of Models 1 and 2 (which we call Model 3) and test the following condition:

Response Consistency 1: *Equality of cut-points*

$$\gamma^k = \gamma_v^k \text{ , } k = 1, \ldots, K\text{-}1. \qquad \text{(RC1)}$$

Besides assumption (A3) of Model 2, this test rests on the assumption of *vignette equivalence* (A1) in Model 1. Under these assumptions the *X*s enter neither (2a) nor (3a). If this were not true, then *RC1* would test $\gamma'^k - \beta_s = \gamma_v'^k - \beta_v$, where γ'^k and $\gamma_v'^k$ are the true cut-point parameters representing reporting behavior and β_s and β_v are vectors of coefficients on X that have been erroneously omitted from (3a) and (2a) respectively. However, there exists a second test that is valid even when the identifying assumptions of RC1 do not hold. This exploits the fact that a necessary condition for each cut-point in the proxy indicators model to be the same as the corresponding one in the vignettes model is that the distance between any two cut-points is the

same in both approaches. Even if the combined Model 3 is too restrictive, in the sense that (A1) and/or (A3) is violated, this condition can still be tested because the parameter vectors β_s and β_v are not cut-point specific and so the distance between any two true cut-points is identified. This leads to a second, more robust test that is, however, less informative than the first in the sense that non-rejection of the null does not imply that *response consistency* holds.

Response Consistency 2: *Equality of distances between cut-points*

$$\gamma^k - \gamma^{k-1} = \gamma_v^k - \gamma_v^{k-1}, \qquad \text{k} = 2, \ldots, \text{K-1} \qquad \text{(RC2)}$$

Van Soest et al (2007) also propose a direct test of response consistency (RC1). This requires a single measure of health that is assumed to be generated by the same latent index of true health that drives self-assessed health but free of the reporting heterogeneity that contaminates the latter. Under these assumptions, the parameter vector β of equation (1a) can be obtained by regressing the presumed objective measure of health on X and, conditional on these parameters, *RC1* can be tested. Unlike our approach, this requires a single measure that proxies the underlying construct of interest. In the context of the application made by Van Soest et al, which is to drinking behaviour, the assumption is plausible. But for health, even a single domain of health, it is less so. There is seldom, if ever, a single objective measure that captures all aspects of a health condition. If there were, then there would be less need to ask individuals about their health. With many proxy indicators of a health condition, one would expect each to relate differently to individual characteristics and no single one to respond to covariates exactly as does true health. It is more plausible that the information contained collectively in a battery of indicators is sufficiently rich such that assumption (A3) holds. Even if this is not the case, we still have the less informative test *RC2*.

3.2 Vignette equivalence

Vignette equivalence rules out any systematic differences in the perception of the health level described by any vignette. A necessary condition for this is no systematic variation in the perceived difference between the levels of health represented by any two vignettes. This can be tested by considering a less restrictive specification of equation (2a), as follows:

$$V_{i1}^{*} = \alpha_1 + \upsilon_{i1}$$
$$V_{ij}^{*} = \alpha_j + \lambda_j X_i^{-} + \upsilon_{ij} \qquad j \neq 1 \qquad \text{(2a')}$$

where X^{-} equals X with the constant term omitted and λ_j is a corresponding vector of parameters. Further extending the specification by allowing X^{-} to impact on perceptions of the first vignette (or another chosen reference vignette) would render the model unidentified, as explained above**.** Significantly non-zero elements of any λ_j indicate systematic differences in the perception of a vignette relative to the reference in contradiction with *vignette equivalence*. This gives the test:

Vignette Equivalence: $\lambda_j = 0 \quad \forall\, j$ (VE)

which is tested in a model composed by equations (2a'), (2b) and (2c), which we refer to as Model 4.[4]

Note that in a model with $\lambda_j \neq 0$ it is not possible to identify reporting heterogeneity since then the vector V^{*} does not represent the true latent health of vignettes but rather the result of different interpretations of vignette descriptions. Furthermore, the resulting cut-point

[4] Murray et al. (2003) conducted a partial, informal test of *vignette equivalence* by investigating whether there are systematic differences in the ranking of vignettes by socio-demographics and questionnaire characteristics.

shift, γ_v^k , depends on the particular vignette that is used as the reference in (2a') and is therefore not meaningful.

It should be noted that this test rests on the assumption that individuals use the same cut-points when rating all vignettes (see (A4) in Table 1). Differential cut-points across vignettes cannot be identified separately from λ. However, even if a non-zero λ were driven by different cut-points, rather than by vignette non-equivalence, that would still be evidence against the validity of using the HOPIT model for the purpose of correcting for reporting heterogeneity.

The model estimated and the assumptions required for the validity of each of the tests we perform are summarized in Table 1. The assumptions (A1') and (A3') are obviously weaker than (A1) and (A3) and require that the effect of each element of *X* on the respective latent index is constant at all levels of the latent health.

Table 1: Models estimated and assumptions required for validity of each test

Test	Model	Objective component	Vignettes component
Response consistency 1 (RC1)	*3: (3a)-(3c) & (2a)-(2c)*	(A3): $h\left(H^* \middle\| H^O, X\right) = h\left(H^* \middle\| H^O\right)$	(A1): $f\left(V_j^* \mid X\right) = f\left(V_j^*\right)$
Response consistency 2 (RC2)	*3: (3a)-(3c) & (2a)-(2c)*	(A3'): $h\left(H^* \middle\| H^O, X\right)$ homoscedastic wrt X	(A1'): $f\left(V_j^* \mid X\right)$ homoscedastic wrt X
Vignette Equivalence (VE)	*4: (2a'), (2b)-(2c)*	-	(A4): $\gamma_j^k = \gamma_v^k \quad \forall j$

3.3 Distributional assumptions and normalisations

All models are estimated by maximum likelihood. Estimation of Models 3 and 4 requires specification of the error distributions and normalisation of location and scale parameters. The location parameters are normalised by excluding the constant terms from the first cut-points (ν_i^1 and τ_i^1). The error terms ξ and η are assumed to be independent of each other and normally distributed with mean zero. Normality is also assumed for υ. The variances of these errors are not

identified and have to be normalised, which is usually done by setting them equal to one. The remaining coefficients are identified up to the respective scale parameters. Estimation of parameters of interest in Model 3 - such as effects of *X* on the probability distribution of vignette ratings and of self-reported health, before or after adjustment for reporting heterogeneity - as well as results of the *vignette equivalence* test (Model 1 *vs* Model 4) are not affected by these normalisations. Under the null hypotheses of the *response consistency* tests, it is possible to identify σ_η/σ_ξ in Model 3. For this reason, in the estimation of the respective restricted models, we normalise only $\sigma_\xi = 1$ and maximise the likelihood with respect to σ_η (and the restricted γ^k and γ_v^k). Under the alternative of no response consistency, the ratio σ_η/σ_ξ is not identified and so the value of the log-likelihood does not depend on either σ_ξ or σ_η. We then maximise the likelihood with respect to γ^k and γ_v^k, normalising both σ_ξ and σ_η. The response consistency hypotheses are tested using likelihood ratio tests, and so test statistics do not depend on these normalisations.

4 Data

The *English Longitudinal Study of Ageing* (ELSA) covers individuals aged 50 and over and their younger partners, living in private households in England. We use data taken mainly from the third wave of ELSA, collected in 2006-2007. In this wave, self-completion forms containing vignettes on six health domains were assigned to a (random) third of the ELSA sample, except for proxy respondents.[5] The vignettes questionnaire consisted of two sections: one which asked

[5] In the case of physically or cognitively impaired respondents, or those in hospital or temporary care, proxy interviews were allowed. The proxy respondents were responsible adults (aged 16 years or over) who are sufficiently aware of the respondent's characteristics and circumstances covered by the questionnaire, preferably

respondents to rate their own health on a 5-point scale, for the domains of cognition, mobility, breathing, pain, sleep and depression, and a second in which they were asked to rate three vignettes, on the same 5-point scale, for each of the health domains. Respondents were requested to assume that the hypothetical individuals described in the vignettes have the same age and background as they do.

4.1 Self-reported health and vignettes

We use self-reports and vignette ratings in a physical health domain (mobility) and a mental health domain (cognition). These two domains are selected because of their dissimilarity, allowing the vignettes approach to be tested with respect to two distinct concepts of health, their importance to the health and welfare of older individuals, and because the survey provides a rich set of objective measures of each of these dimensions of health, which increases the plausibility of assumption (A3). Independence in later life is determined by physical and cognitive functioning, which are key markers of the health of elderly populations (Steel et al, 2004) and strong determinants of the need for, and costs of, health and social care (Gill et al., 2001). Similar to physical impairment, cognitive decline may lead to inadequate functioning in daily life, including reduced ability to work and to deal with financial matters (Fillenbaum et al., 1988; Reed, Jagust, & Seab, 1989; Park, 1999; Ofstedal et al, 2006). Important decisions related to retirement and pension planning may be impeded by cognitive decline (Park, 1999; Banks and Oldfield, 2007). Physical and cognitive functioning are two health domains for which previous applications of the anchoring vignettes approach have revealed reporting heterogeneity (Bago d'Uva et al 2008a, 2008b).

---

their partner, son or daughter. The vignettes questionnaire was not assigned to proxy respondents, and so the vignette sample does not include individuals interviewed by proxy.

Self-reports are obtained from the questions "Overall in the last 30 days, how much difficulty have you had with concentrating or remembering things?" (cognitive functioning) and "Overall in the last 30 days, how much of a problem have you had with moving around?" (mobility). In each case, the categorical responses are: "Extreme", "Severe", "Moderate", "Mild" and "None". As a very low proportion of individuals reported "Extreme" or "Severe", we have collapsed the first three categories (also for the vignettes). The respondents are then asked to answer the same question regarding the functioning of three vignettes in each domain[6]:

- *Cognition 1 - Mary can concentrate while watching TV, reading a magazine or playing a game of cards or chess. Once a week she forgets where her keys or glasses are, but finds them within five minutes.*
- *Cognition 2- Sue is keen to learn new recipes but finds that she often makes mistakes and has to reread them several times before she is able to do them properly.*
- *Cognition 3 - Eve cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to her. When she starts a task, she never manages to finish it and often forgets what she was doing. She is able to learn the names of people she meets.*
- *Mobility 1- Robert is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing more than one flight of stairs. He has no problems with day-to-day activities such as carrying food from the market.*
- *Mobility 2 – Tom has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy.*

[6] Given each individual evaluates three vignettes within each health domain, in principle it is possible to allow for unobserved heterogeneity in the response scale, specified as a random individual effect. However, with only three vignettes available, identification can be expected to be weak, which we have confirmed.

- *Mobility 3 – David does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work.*

The response distributions for own functioning and each vignette are presented in Table 2. It is clear that the average rated degree of cognitive/mobility difficulties rises with vignette number, as would be anticipated, and is always higher than the average respondent's rated degree of difficulty with her own cognition/mobility.

Table 2: Frequencies of assessed degree of cognition and mobility of respondent and vignette

	Cognition				Mobility			
Degree of difficulty	Own	Vignette 1	Vignette 2	Vignette 3	Own	Vignette 1	Vignette 2	Vignette 3
At least moderate	316	419	1318	1646	269	637	1155	1198
Mild	735	1,078	403	96	254	545	77	48
None	731	285	61	40	757	98	48	34
N	1782	1782	1782	1782	1280	1280	1280	1280

Note: N is smaller for mobility since only respondents aged 60+ take the walking speed test, which is used as an objective indicator of mobility.

4.2 Cognitive functioning tests

The ELSA cognitive functioning module is administered to all respondents, except proxy respondents. This module assesses a range of cognitive processes, which in Wave 3 included memory (retrospective and prospective) and executive function (organization, verbal fluency, abstraction, attention, mental speed, etc) (Steel et al, 2004). In waves 1 and 2, basic numeracy and literacy respectively were tested. The cognitive measures administered aim at: (a) assessing cognitive processes relevant to the everyday function of older individuals; (b) using tasks that are known to be sensitive to age-related decline; (c) avoiding floor and ceiling effects; and (d) comparability with other studies, in particular, the HRS (*ibid*). The tests have been used extensively in gerontological, geriatric, medical, epidemiological, neurological and psychological studies (see below). The ELSA cognitive test data have been used in recent

geriatric (Lang et al, 2008), neurological (Llewellyn et al, 2008) and economic studies (Banks and Oldfield, 2006). Steel et al (2004) find that a global score based on the scores of memory, executive function and numeracy from wave 1 of ELSA covaries in the expected way with age, education and health. We use results from all cognitive tests available in Wave 3, and results of numeracy and literacy tests performed by the same individuals in previous waves. Memory (1-4), executive function (5-7), and basic skills (8, 9) were assessed using the following tests:

1. Orientation (in time): This test includes standard questions about the date (day, month, year) and the day of the week, and it has also been used in HRS. It was taken from the Mini Mental Status Examination (MMSE), which is widely used and considered as the "gold standard" of cognitive impairment screening tests (Lee et al, 2002; Weuve et al, 2004).[7] The ELSA dataset includes a score derived from the date questions, which is increasing with cognitive ability.

2. Immediate memory and 3. Short-term memory (verbal learning and recall): Participants are presented orally with 10 common words and asked to remember them. Word recall is tested both immediately and after a short delay, during which other cognitive tests are performed. ELSA uses the word lists developed for HRS. These tests are very commonly used. The derived measures are the number of words recalled correctly immediately and after delay.

4. Prospective memory (memory for future actions): Early in the cognitive module, respondents are told about an action that they will be asked to carry out later.[8] They are also told that they will need to carry this out without being reminded of what the action is. The action is based on a

[7] The MMSE (Foldstein et al, 1975) is the most commonly used instrument for cognitive function, it has been validated and extensively used in both clinical practice and research. It provides measures of orientation, registration (immediate memory), short-term memory (but not long-term memory) as well as language functioning, from 11 questions. The MMSE is effective as a screening tool for cognitive impairment with older, community dwelling, hospitalized and institutionalized adults. It has however been considered limited to detect subtle memory loss in educated people, Small (2002). Immediate and delayed word list recall can be added to identify subtle memory loss (Small, 2002).

[8] Respondents are asked to write their initials in the top left-hand corner of a page that is attached to a clipboard, when they are later handed the clipboard.

similar task used in the Medical Research Council Cognitive Function and Ageing Study (MRC CFA Study, 1998; Huppert et al, 2000).

5. Word-finding & verbal fluency: This test assesses how quickly individuals can think of words from a particular category (in this case animals) in one minute. It tests self-initiated activity, organisation and abstraction and set-shifting. This test was taken from the Cambridge Cognitive Examination - CAMCOG (Huppert et al, 1995; Roth et al, 1986) and it has been used in many studies including the MRC National Study of Health and Development (Richards et al, 1999) and the Nurses' Health Study (Lee et al, 2002; Weuve et al, 2004). The result of this test is the number of animals mentioned.

6. Processing speed and 7. Search accuracy (attention, visual search and mental speed): The respondent is handed a clipboard to which is attached a page of random letters of the alphabet set out in (26) rows and (30) columns, and is asked to cross out as many target letters (65 in total) as possible in a minute. The total number of letters searched[9] provides a measure of speed of processing. The proportion of correctly identified target letters among all those scanned is a measure of search accuracy. This test was taken from the MRC National Study of Health and Development (1946 Birth Cohort Study, Richards et al, BMJ 2008, Richards et al, 1999).

8. Numeracy: Respondents are asked to solve up to six problems requiring simple mental calculations based on real-life situations. They are first tested using three moderately easy items. Those who fail on all these items are then asked an easier question, while those who answer correctly at least one of those questions are asked two progressively more difficult questions (and given credit for the easiest one). A score of 1 is given to correct answers on each question. The

[9] This can be identified because respondents are asked to work across and down the page as though they were reading, and to mark the last letter that they have checked at the end of the test.

participant can obtain a score between 0 and 6. The problems were developed for ELSA and later used in HRS.

9. Literacy: This test aimed at deriving a measure of prose literacy relevant for the lives of the elderly. Participants were shown a realistic label for a fictitious medicine called Medco Aspirin. The test then consisted of assessing their understanding of the instructions on the label, by asking them: *i)* the maximum number of days for which this medication should be taken; *ii)* to name three situations in which a doctor should be consulted (out of six situations mentioned on the label); and, *iii)* to name one condition for which the tablets can be taken (out of six). Scores on this test range from 0 to 3. This test has been used in the International Adult Literacy Survey (IALS) (OECD & Statistics Canada, 2000) and the Adult Literacy and Life Skills Survey (Statistics Canada & OECD, 2005).

All tests scores were rescaled to the [0,1] interval, increasing in cognitive functioning resulting in the variables summarized in Table 3.

4.3 Mobility indicators

We use results from a measured test of walking speed, administered to respondents aged 60 or over for whom the test is judged safe. Impaired mobility measured by functional tests such as walking speed is predictive of future disability, nursing-home entry and mortality (Guralnik et al., 1994) and such tests may be used in clinical assessments of older people (Guralnik and Ferrucci, 2003; Studenski et al., 2003). Eligible ELSA respondents were asked to walk a distance of 8 feet (244 cm) at their usual walking pace. They were asked to do this twice and the interviewer recorded the time taken in each walk, using a stopwatch. Our measure (Walking speed) equals the average of the two measurements, for participants with two valid measurements (as in Banks et al, 2008). This gives an objective, but perhaps not sufficiently

comprehensive, measure of mobility. We complement it with a battery of indicators of physical functioning, in particular, difficulties with activities of daily living (ADL) and problems with motor skills and strength summarized in Table 3.

The existence of problems with motor skills and strength is assessed through questions about any difficulty in: walking 100 yards; getting up from a chair after sitting for long periods; climbing several flights of stairs without resting; climbing one flight of stairs without resting; stooping, kneeling or crouching; pulling or pushing large objects like a living-room chair; lifting or carrying weights over 10 pounds, like a heavy bag of groceries; reaching or extending arms above shoulder level; sitting for about two hours; and, picking up a small coin from a table. Similar items are included in the HRS (Wallace and Herzog, 1995) and have been used, for example, as objective health measures in Kreider (1999). We include dummy variables indicating the number of items with which the individual reports difficulties, collapsing the ones referring to 5 or more items as the respective estimated effects differed little and not significantly so.

The original scale of ADLs includes activities which are likely to be part of the lives of most people and was developed by Katz et al (1963). This scale is widely used for professional assessments of the needs of older people (Banks et al, 2008) and versions of it have been widely used in the gerontological, medical, epidemiological, and health economics literature. The activities covered in ELSA are: dressing (including putting on shoes and socks); walking across a room; bathing or showering; eating (such as cutting up food); getting in or out of bed; and, using the toilet. We include indicators of whether individuals have difficulty with one ADL (1 ADL), or with two or more ADLs (2+ ADLs). The reference is no difficulty with any ADL. Similar to

motor problems, further discrimination of the number of ADLs with which individuals have difficulty was not informative.

While both the indicators of motor skills and ADLs are self-reported, the precise definition of each task and the dichotomous nature of the responses (is/isn't restricted) make it unlikely that they are subject to any substantial systematic reporting heterogeneity. Conditioning on these indicators, as well as walking speed, should therefore be effective in controlling for systematic variation in true mobility, leaving any residual variation in reported mobility attributable to differences in reporting thresholds.

4.4 Socio-demographic variables

We examine reporting heterogeneity in cognitive functioning and mobility with respect to age, gender, ethnicity, wealth, education and employment status. Age, gender and education have been shown to influence reporting of several health domains, including cognition, in previous vignette studies (Bago d'Uva et al 2008a, 2008b). Income has also been shown to influence vignette ratings in Bago d'Uva et al (2008a). For an older sample, wealth is a more appropriate measure of economic status than income, and ELSA provides a very accurate measure of wealth. Using ELSA data, Banks et al (2006) have found less wealth to be associated with more sickness, less functionality and a greater likelihood of dying. Cultural differences across ethnic groups may influence concepts and reporting of health. Use of self-reported health will bias estimated associations between health and labor force participation if non-working individuals under-report their health to justify their status. However, the vignette approach may be powerless to correct such *justification bias* since respondents do not have a similar incentive to misclassify the health of hypothetical individuals. *A priori* one may therefore doubt the validity of the *response consistency* assumption in relation to work status.

Age is represented by age-group dummies and ethnicity by a dummy to distinguish between *Whites* and ethnic minorities. The variable *ln(Wealth)* is the logarithm of total non-pension wealth, set to zero for individuals with non-positive wealth, who are distinguished by a dummy (*No wealth*). Since wave 3 wealth data are not yet available, we used those from wave 2, which, in any case, may be preferable in order to minimize potential endogeneity to health. Education is represented by dummies for the highest qualification, with those with no qualifications being the reference. We include an indicator of whether individuals are younger than 65 *and* are not working (*Not working <65*) to capture the effect of employment status for individuals below the normal retirement age, i.e., those who may have an incentive to under-report health as a justification for not working. Because it is unlikely that individuals aged 65+ behave similarly and because the proportion above 64 who work is very small, the reference group includes individuals younger than 65 who are working and those aged 65 or older (regardless of working status). Since our age variables discriminate between individuals above and below 65, the effect of *Not working <65* will actually represent, for those below 65, the effect of not working.

The dataset used for the analysis in the domain of cognition results from deletion of individuals with missing data on self-reported cognition, respective vignettes, the cognitive tests and the socio-demographic variables. The resulting dataset contains 1782 individuals aged 50 and over. In the case of mobility, we dropped individuals younger than 60, who did not perform the walking speed test, but did not drop those with missing information for cognition (vignettes, self-reports or measured tests), leading to a dataset with 1280 individuals. Table A1 in the appendix documents the number of observations lost to item non-response in each domain. Since we use information on wealth and the literacy test from Wave 2 and on the numeracy test from

Wave 1, our samples do not include respondents who have entered the sample only in Wave 3, as part of the refreshment sample added to ELSA (Nunn et al, 2008), and for the cognition analysis some individuals who joined in Wave 2 (mainly new partners) are excluded.

Descriptive statistics for the covariates are given in Table 3. The distribution of covariates is similar in the two samples, except that the mobility sample is obviously older (60+) and for that reason is on average less educated.

Table 3: Descriptive statistics of health measures and socio-demographic variables

Variable	Cognition sample		Mobility sample	
	Mean	Std dev	Mean	Std dev
Cognitive tests				
1. Orientation	0.947	0.120		
2. Immediate memory	0.582	0.173		
3. Short-term memory	0.457	0.206		
4. Prospective memory	0.748	0.350		
5. Word-finding & verbal fluency	0.363	0.114		
6. Processing speed	0.380	0.107		
7. Search accuracy	0.813	0.131		
8. Numeracy	0.694	0.204		
9. Literacy	0.865	0.230		
Mobility indicators				
Walking speed			3.362	1.906
1 ADL			0.110	0.313
2+ ADLs			0.081	0.273
1 motor problem			0.195	0.396
2 motor problems			0.111	0.314
3 motor problems			0.083	0.276
4 motor problems			0.067	0.250
5+ motor problems			0.170	0.376
Socio-demographic variables				
Age 55 to 64	0.392	0.488		
Age 65 to 74	0.308	0.462	0.427	0.495
Age 75+	0.238	0.426	0.321	0.467
Female	0.574	0.495	0.559	0.497
White	0.989	0.105	0.988	0.108
ln(Wealth)	11.446	2.772	11.459	2.631
No wealth	0.038	0.192	0.032	0.176
A-level or above	0.341	0.474	0.282	0.450
Qualification < A-level	0.263	0.440	0.266	0.442
Not working <65	0.190	0.392	0.157	0.364
N	1782		1280	

Notes: All cognitive test scores are re-scaled to 0,1 and are increasing with cognitive functioning. 'A-level or above' includes National Vocational Qualification (NVQ) level >=3 and higher education. A-level is roughly equivalent to high school graduation; 'Qualification<A-level' includes O level, NVQ 2, CSE, NVQ 1, or other (including foreign). No qualifications is the reference.

5 Results

We first use Model 3 to assess the similarities and differences in reporting heterogeneity estimated from the vignettes approach and the proxy indicators approach. We then implement the proposed tests of *response consistency* (RC1 and RC2) in this model. Finally, we test the necessary condition for *vignette equivalence* (VE) in the vignettes part of the model (i.e. test Model 1 against Model 4).

5.1 Reporting heterogeneity

Results from the estimation of Model 3 for cognition are presented in Table 5.[10,11] As expected, all cognitive test scores are positively correlated with cognitive functioning. Due to collinearity, only the scores from the immediate and short-term memory tests, and the verbal fluency test are individually significant, but each is significant when no control is made for the others and they are jointly significant (p-value <0.000).

Both the proxy indicators and vignettes approaches reveal evidence of reporting heterogeneity (Tables 5 & 7). Wealth and age affect the cut-points in both models, while work status affect the cut-points only with the proxy indicators approach and significant cut-point shift by educational level and ethnicity is revealed only with the vignettes approach. Wealth lowers

[10] As noted before, the coefficients in this model are identified up to the scale parameters σ_η and σ_ξ. For illustration, we present here results with normalisation $\sigma_\xi = 1$ and σ_η equal to the estimate under the null hypothesis of RC1. The presented z-scores do not depend on the chosen normalisation.

[11] We experimented with quadratic specifications for the test scores but found that square terms were not jointly significant (only one was individually significant) and the effects of the test scores and their squares were imprecisely estimated. Estimated cut-point shift according to the proxy indicators approach and results of response consistency tests are not affected by the exclusion of the squared terms.

the first cut-point indicating that the more wealthy are less likely to declare a given level of functioning as at least a moderate degree of limitation (as opposed to none or mild). But there is evidence of a non-linear effect in wealth, with those with no wealth also being less likely to report mild/moderate difficulties in functioning (except using the proxy indicator approach). The cut-points tend to be highest for the oldest individuals, indicating they rate a given level of cognitive functioning as corresponding to a greater degree of limitation. This is confirmed by both approaches and is in line with previous results (Bago d'Uva et al, 2008a). The vignettes approach finds that the better educated are more likely to consider a given level of cognitive functioning as corresponding to mild or no difficulty, as opposed to at least moderate difficulty. Similar effects are obtained with the proxy indicator approach but these are estimated less precisely. It may seem somewhat surprising that the higher educated have lower expectations regarding cognitive functioning but this finding is in line with existing evidence (Bago d'Uva et al, 2008a; Bago d'Uva et al, 2008b). It could be that educated individuals are less willing to admit cognitive impairment. Whites tend to rate vignettes as more cognitively impaired, which would suggest that observed ethnic differences in cognitive functioning understate true differences. However, this is not confirmed by the proxy indicator approach. According to the latter, individuals under 65 who are not working are more likely to declare a limitation in their own cognitive functioning, but they do not apply the same strict criteria to rating of the vignettes. This is consistent with our hypothesis that non-employment may introduce a justification bias to the reporting of health that is not captured by the vignettes approach.

Walking speed and each of the ADL and motor skills dummies are significantly correlated with latent mobility (Table 6). The evidence of heterogeneity in the reporting of mobility differs from what is observed for cognition in several respects (Tables 6 and 7). The

proxy indicators approach reveals evidence of cut-point shift according to gender, ethnicity and wealth, while evidence of differential rating of vignettes is found only according ethnicity and education. Females and the less wealthy are more likely to rate their own mobility more positively but not that of the vignettes. Better educated individuals are less likely to consider the mobility level of the vignettes as corresponding to no difficulty, while there is no evidence of heterogeneous reporting of own health by education. The disparity in cut-point shift by ethnicity across the two approaches that was observed for cognition is confirmed for mobility. In fact, there is a clear tendency of Whites to be optimistic in reporting their own mobility, while being pessimistic in reporting that of the vignettes. The estimated effects of employment status on the reporting of mobility are less supportive than those for cognition of the justification bias hypothesis.

Table 5: Estimation results of Model 3 for cognition

	Coef.	Std. Err.	z		Coef.	Std. Err.	Z
Latent cognition as function of test scores				**Latent cognition of vignettes**			
1. Date	0.371	0.262	1.420				
2. Words immediate	**0.515**	0.259	1.990	Vignette 1	**0.616**	0.253	2.440
3. Words delay	**1.028**	0.223	4.610				
4. Prospective memory	**0.781**	0.327	0.710				
5. Animals	0.067	0.094	2.390	Vignette 2	**-0.615**	0.253	-2.430
6. Processing speed	0.252	0.335	0.750				
7. Search accuracy	0.138	0.275	0.500				
8. Numeracy	0.067	0.145	0.720	Vignette 3	**-1.351**	0.255	-5.310
9. Literacy	0.133	0.184	0.460				
Constant	**-1.717**	0.607	-2.830				
σ_η	1.132	(fixed)		σ_ξ	1.000	(fixed)	
Response scales identified from test scores				**Response scales identified from vignettes**			
Cut-point – moderate/severe/extreme				*Cut-point – moderate/severe/extreme*			
Age 55 to 64	-0.092	0.223	-0.410		0.000	0.087	0.000
Age 65 to 74	0.208	0.218	0.950		0.072	0.089	0.810
Age 75+	**0.567**	0.222	2.560		0.132	0.092	1.440
Female	-0.127	0.090	-1.410		-0.056	0.041	-1.370
White	-0.189	0.377	-0.500		**0.663**	0.178	3.720
ln(Wealth)	**-0.084**	0.025	-3.340		**-0.050**	0.014	-3.690
No wealth	-0.664	0.351	-1.890		**-0.440**	0.192	-2.290
Qualifications 2	-0.147	0.112	-1.310		**-0.156**	0.052	-3.020
Qualifications 1	-0.134	0.108	-1.240		-0.061	0.052	-1.170
Not working <65	**0.398**	0.149	2.670		-0.059	0.062	-0.950
Cut-point – mild				*Cut-point - mild*			
Age 55 to 64	-0.105	0.152	-0.690		0.153	0.118	1.300
Age 65 to 74	0.090	0.155	0.580		0.206	0.120	1.720
Age 75+	**0.422**	0.167	2.520		**0.361**	0.128	2.820
Female	0.017	0.077	0.220		0.006	0.059	0.100
White	-0.223	0.341	-0.660		**1.057**	0.191	5.530
ln(Wealth)	-0.013	0.024	-0.560		-0.021	0.020	-1.040
No wealth	0.208	0.343	0.610		**-0.580**	0.276	-2.100
Qualifications 2	0.127	0.094	1.350		0.000	0.074	0.000
Qualifications 1	0.081	0.093	0.880		0.051	0.076	0.670
Not working <65	**0.320**	0.107	2.980		0.095	0.088	1.080
Constant	0.529	0.527	1.000		0.437	0.334	1.310
Log likelihood		-5178.03					
N		1782					

Note: Bold indicates significance at 5%.

Table 6: Estimation results of Model 3 for mobility

	Coef.	Std. Err.	z		Coef.	Std. Err.	z
Latent mobility as function of proxy indicators				**Latent mobility of Vignettes**			
Walking speed	**-0.364**	0.061	-6.000				
Walking speed squared	**0.009**	0.003	2.770	Vignette 1	**-0.755**	0.280	-2.690
1 ADL	**-0.353**	0.143	-2.470				
2+ ADLs	**-0.606**	0.189	-3.200				
1 motor problem	**-0.693**	0.135	-5.150	Vignette 2	-0.528	0.279	-1.890
2 motor problems	**-1.294**	0.156	-8.290				
3 motor problems	**-1.586**	0.168	-9.460				
4 motor problems	**-1.878**	0.188	-10.010	Vignette 3	**0.610**	0.279	2.190
5+ motor problems	**-2.316**	0.169	-13.680				
Constant	**4.166**	0.739	5.640				
σ_η	1.182	(fixed)		σ_ξ	1.000	(fixed)	
Response scales identified from proxy indicators				**Response scales identified from vignettes**			
Cut-point – moderate/severe/extreme				*Cut-point – moderate/severe/extreme*			
Age 65 to 74	0.318	0.270	1.180		0.041	0.090	0.460
Age 75+	0.367	0.275	1.330		-0.061	0.094	-0.650
Female	**-0.304**	0.121	-2.510		0.018	0.051	0.350
White	-0.717	0.521	-1.370		**0.614**	0.200	3.060
Ln(Wealth)	**0.106**	0.037	2.850		0.005	0.016	0.330
No wealth	0.747	0.521	1.430		0.325	0.240	1.360
Qualifications 2	-0.010	0.150	-0.070		0.056	0.065	0.870
Qualifications 1	0.031	0.147	0.210		-0.026	0.062	-0.430
Not working <65	0.308	0.298	1.030		0.060	0.104	0.580
Cut-point - mild				*Cut-point – mild*			
Age 65 to 74	0.052	0.187	0.280		0.139	0.129	1.080
Age 75+	0.204	0.197	1.040		0.066	0.135	0.490
Female	**-0.281**	0.105	-2.670		0.048	0.075	0.640
White	**-1.075**	0.443	-2.430		**1.152**	0.217	5.320
Ln(Wealth)	0.016	0.035	0.460		-0.016	0.024	-0.680
No wealth	0.493	0.509	0.970		-0.245	0.349	-0.700
Qualifications 2	0.055	0.129	0.430		**0.337**	0.099	3.390
Qualifications 1	0.002	0.126	0.020		**0.203**	0.095	2.150
Not working <65	0.047	0.213	0.220		0.294	0.156	1.880
Constant	**2.676**	0.769	3.480		0.523	0.320	1.630
Log likelihood		-2948.89					
N		1280					

Note: Bold indicates significance at 5%

Table 7: Tests of no reporting heterogeneity

		Vignettes Model 1		Proxy Indicators Model 2	
	Degrees of freedom	**LR test statistic**	**p-value**	**LR test statistic**	**p-value**
Cognition					
All variables	20	68.17	<0.001	109.11	<0.001
Age	6	32.50	<0.001	11.88	0.065
Female	2	2.72	0.257	2.09	0.352
White	2	0.51	0.776	33.10	<0.001
Wealth	4	15.81	0.003	24.54	<0.001
Education	4	6.60	0.158	10.36	0.035
Not_working <65	2	12.36	0.0021	2.74	0.255
Mobility					
All variables	18	33.47	0.015	52.06	<0.001
Age	4	3.33	0.505	3.94	0.414
Female	2	9.37	0.009	0.42	0.810
White	2	6.12	0.049	25.12	<0.001
Wealth	4	13.83	0.008	5.49	0.241
Education	4	0.43	0.980	14.45	0.006
Not_working <65	2	1.15	0.562	3.52	0.172

Note: LR – Likelihood Ratio.

5.2 Global tests of response consistency and vignette equivalence

Results of the tests of *response consistency* are presented in Table 8. Using the stricter test (RC1), *response consistency* (RC1) is rejected in all cases. This could be anticipated from the results in Tables 5 and 6, which show that estimates of reporting heterogeneity by covariates differ depending upon whether identification is achieved through the vignettes or proxy indicators of health. The second, weaker test of *response consistency* (RC2) does *not* reject the null for cognition, but does so for mobility. With respect to cognition, the discrepancy between RC1 and RC2 may be either (i) because the latter only tests a necessary condition or (ii) because the assumptions required for RC1 to be a valid test do not hold. In the case of (i), the cut-points do in fact differ for own and vignette cognition, i.e. RC does not hold, but the distances between

them do not differ. In case (ii), the cut-points are in fact the same, i.e. RC holds, but (A1) and/or (A3) is too restrictive such that covariates should appear in the health index.

Strictly, it is not possible to distinguish between these explanations but the test of the necessary condition for *vignette equivalence* can help identify the more plausible of the two. This test is performed on Model 4, estimation results of which can be found in Tables A2 and A3 in the Appendix. This model allows the perceived latent health of two of the three vignettes to vary with covariates, which is implemented through interaction terms. For cognition, there are significant interactions between all factors (except working status and age) and at least one of the vignette dummies and these are jointly significant (Table 8). This suggests that violation of (A1) may be driving the conflicting results given by RC1 and RC2. In this case, the vignette approach would not be appropriate to correct for cut-point shift as this cannot be identified separately from systematic differences in the perceived latent health level of the vignettes. In the case of mobility, there are fewer significant interactions (mainly due to lack of precision of the estimates in this smaller sample) but they are jointly significant. The evidence against the use of the vignette approach in the domain of mobility is even more compelling, as the null hypothesis is decisively rejected in all three tests.

Table 8: Tests of response consistency and vignette equivalence

Test	Degrees of freedom	LR test statistic	p-value
Cognition			
RC1	21	43.26	0.003
RC2	11	13.10	0.287
VE	20	105.56	<0.001
Mobility			
RC1	19	55.18	<0.001
RC2	10	20.35	0.026
VE	18	44.66	0.001

Note: LR – Likelihood Ratio.

5.3 Tests by covariate

We now examine all tests separately by each covariate in order to assess the extent to which the vignette approach may adequately correct for reporting heterogeneity in relation to a particular characteristic, even if it fails in general (Table 9). In the case of cognition, the first test of *response consistency* (RC1) rejects the assumption with respect to age, wealth, work status and ethnicity, while the second test (RC2) rejects it only for wealth, and then only at the 10% level. Since neither test rejects the null for education, the vignettes approach appears to appropriately correct for differences by education in the reporting of cognition, resulting in smaller education disparities. However, *vignette equivalence* is rejected for education and all other factors, except for sex (although it is marginal) and working status. This suggests that the better educated tend to have lower perceptions of the latent health level of vignettes. Nevertheless, the vignettes approach does provide estimates of cut-point shift by education similar to the ones given by the proxy indicators approach (Table 5). So, while the vignettes approach may not be strictly valid as a correction for reporting heterogeneity by education, it may nonetheless succeed in revising the estimated disparities by education in the correct direction.[12]

For each of the remaining variables, at least one of the tests is rejected in the domain of cognition (except for female, but there is no evidence of reporting heterogeneity by gender - Table 5). The vignettes approach would lead to overestimation of white *vs* non-white differences in cognition. The *vignette equivalence* test suggests that white individuals tend to have lower perceptions of the cognitive ability of the vignettes (see Table A2 in the appendix), which may be erroneously translated into positive cut-point shift by the vignettes model. The RC2 test,

[12] It should be borne in mind, however, that the direction of the adjustment by education found here for English elderly is the opposite to that found by Bago d'Uva (2008) for most continental European countries, with the exceptions of Spain and Sweden.

which does not impose *vignette equivalence*, does not show evidence against *response consistency*, consistent with a situation of *response consistency* but no *vignette equivalence*. But, in any case, this is evidence against the validity of the vignettes approach.

The anchoring and proxy indicator approaches also do not concur with respect to the reporting of cognition by employment status, resulting in rejection of *response consistency* by the first test. This is consistent with our *a priori* expectation that reporting on vignettes would not be helpful in correcting for justification bias. However, the second test, which relaxes *vignette equivalence* and the assumption of the proxy indicators approach that the cognitive scores capture all association between cognitive ability and work status, does not reject *response consistency*. Since *vignette equivalence* is not rejected, it is possible that the assumptions of the vignettes approach hold but that of the proxy indicators approach fails. The comprehensiveness of the indicators lead us to believe that this is not the case, but we cannot rule out the possibility that the tests do not sufficiently pick up some aspects of cognitive ability favourable to working individuals, which would then be reflected as positive cut-point shift in the proxy indicators model.

For mobility, the first test rejects *response consistency* with respect to gender, ethnicity and wealth. Rejection is strongest for ethnicity, a reflection of the fact that the two approaches show opposite and significant cut-point shift by that factor. The weaker test of *response consistency* is rejected for wealth and ethnicity (10%). *Vignette equivalence* is rejected for age (10%) and education and the p-value lies only just above 10% for all the other factors except for employment status. Across all three tests there is evidence against at least one null hypothesis for each variable, again with the sole exception of work status. This exception is interesting since mobility related problems are an important reason given for labor force withdrawal and the

finding goes against our expectation that the vignettes approach would not perform well in the identification of reporting heterogeneity by employment status. Admittedly, the impact of employment status on the response scale for mobility is only marginally significant (Table 6).

Table 9: Tests of response consistency and vignette equivalence by variable

		Cognition			Mobility		
Test	**Variable**	**Degrees of freedom**	**LR test statistic**	**p-value**	**Degrees of freedom**	**LR test statistic**	**p-value**
RC1							
	Age	6	15.14	0.0192	4	4.31	0.3654
	Female	2	0.66	0.7178	2	8.83	0.0121
	White	2	11.47	0.0032	2	20.17	<0.0001
	Wealth	4	10.84	0.0285	4	14.48	0.0059
	Education	4	1.80	0.7733	4	4.31	0.3655
	Not_working <65	2	8.85	0.0120	2	2.42	0.2978
RC2							
	Age	3	2.32	0.5095	2	1.62	0.4451
	Female	1	0.84	0.3580	1	<0.01	0.9897
	White	1	0.67	0.4132	1	3.07	0.0799
	Wealth	2	5.52	0.0634	2	9.70	0.0078
	Education	2	1.64	0.4409	2	2.88	0.2372
	Not_working <65	1	1.33	0.2482	1	2.40	0.1214
VE							
	Age	6	16.84	0.0099	4	8.83	0.0655
	Female	2	4.43	0.1091	2	4.35	0.1138
	White	2	8.06	0.0177	2	4.38	0.1119
	Wealth	4	20.58	0.0004	4	7.52	0.1109
	Education	4	23.98	0.0001	4	14.56	0.0057
	Not_working <65	2	0.32	0.8522	2	0.48	0.7862

Note: LR – Likelihood Ratio.

6 Conclusion

Improving the interpersonal comparability of subjective indicators is an important challenge for survey research. Anchoring individuals' responses on evaluations of vignette descriptions is an intuitively appealing idea that could be effective in meeting this challenge. But the method relies on two identifying assumptions that hitherto have seldom been tested. We propose tests of both assumptions. Like Van Soest et al (2007), our test of *response consistency* requires data on

objective indicator(s) of the construct of interest that allows response scales to be identified and compared with those obtained from the vignettes approach. Unlike Van Soest et al, we do not require that there exists a single objective measure that relates to individual socio-demographic characteristics in exactly the same way as the latent construct. Rather, we require that a battery of proxy indicators contains sufficient information such that there is no residual covariance between socio-demographics and the construct. We argue that this is a more plausible assumption in the context of health measurement. We introduce a weak test of *response consistency* that rests on a less strong assumption about the information content of the objective indicators. In addition, we propose a test of a necessary condition for the second assumption of the vignettes approach – *vignette equivalence.*

Application of these tests to the reported cognition and mobility of a sample of older English males and females provides evidence against the validity of the vignettes approach. *Response consistency* and *vignette equivalence* are rejected for both health domains. The weaker test does not reject *response consistency* for cognition but does so for mobility. By factor, *response consistency* is rejected by the stronger test for all but gender and education in the case of cognition and for all but age, education and employment status for mobility. Using the weaker test, the assumption is clearly rejected only for wealth for both health domains, and more marginally for ethnicity in the case of mobility. *Vignette equivalence* is rejected for all factors other than employment status and (marginally) gender for both domains, and for mobility there is also no clear evidence against *vignette equivalence* for ethnicity and wealth.

An arguably legitimate defense of the vignettes approach against these findings is that the tests are very demanding. While *response consistency* and *vignette equivalence* are required to identify the parameters of reporting behavior, researchers may be satisfied with uncovering the

direction of bias induced by reporting heterogeneity. For example, one may settle for knowing whether the higher educated over or under state their health, without having an unbiased estimate of the extent to which they do so. Our results suggest that the vignettes approach is sometimes able to satisfy this less ambitious objective. For both health domains, it reveals reporting tendencies by age and education in the same direction as those estimated from the objective indicators approach, and for most of the other covariates the two approaches do not give significant effects of opposite sign. But this is not always the case. The vignettes approach indicates significant differences in reporting of health by ethnicity in the opposite direction to those found by the objective proxies method. On the whole, these results are a warning that caution should be exercised in using the vignettes approach to identify and to correct reporting heterogeneity, at least with respect to the reporting of health and more specifically cognitive functioning and mobility.

Rejection of *vignette equivalence* may be attributable to a lack of objectivity in the wording of the vignette descriptions. For example, expressions such as "often makes mistakes", "has difficulty", and "some light household work" are frequently found in vignette descriptions and may be prone to variable interpretation in much the same way as the category labels of the variables the approach aims at correcting. Researchers should aim to make the vignette descriptions as objective as possible, making reference to specific activities that can and cannot be done and the precise frequency with which problems arise. Admittedly, this is more feasible for some health domains (those related to physical functioning) than it is for others (that derive from mental health problems and the experience of pain).

Another practical measure that could improve implementation of the approach would be to switch the usual question order so that self-assessments follow the vignettes. This would be a

way of priming respondents to define the response scale in a common way. Hopkins and King (2008) show that asking the vignette questions first leads to significant improvements in the estimation of expected relationships between socio-demographic variables and vignette-corrected political efficacy and economic class.

While our results do cast some doubt on the validity of the vignettes approach, they are not sufficient to reject a promising and potentially effective survey instrument. The proposed tests should be applied in other domains of health and, where possible, to other subjective welfare indicators that have been anchored on vignette evaluations. Different tests, perhaps based on experiments, need to be developed for constructs for which it is difficult to obtain objective measures. If future research confirms the negative results found here, then researchers willing to use vignettes should consider exploiting nonparametric approaches that rest on weaker assumptions (e.g., Wand, 2007).

References

Bago d'Uva, T., E. van Doorslaer, M. Lindeboom, O. O'Donnell. Does Reporting Heterogeneity Bias Health Inequality Measurement? *Health Economics* 2008, 17(3): 351-375.

Bago d'Uva, T., O O'Donnell, E van Doorslaer. Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans, *International Journal of Epidemiology* 2008, 37:1375–1383.

Banks J and Oldfield Z. Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving. *Fiscal Studies* 2007. Volume 28 Issue 2, Pages 143 – 170.

Banks J, Breeze E, Lessof C and Nazroo J *eds.* (2008) Living in the 21st century: older people in England THE 2006 ENGLISH LONGITUDINAL STUDY OF AGEING (Wave 3). Institute for Fiscal Studies.

Banks, J., Breeze, E., Lessof, C. and Nazroo, J. (eds) (2006), Retirement, Health and Relationships of the Older Population in *England: The 2004 English Longitudinal Study of Ageing*, London: Institute for Fiscal Studies.

Benitez-Silva, H., M. Buschinski, H. M. Chan, S. Cheidvasser, et al. (1999). "How large is the bias in self-reported disability?" *Journal of Applied Econometrics* **19**(6): 649-670

Bound, J. (1991). "Self reported versus objective measures of health in retirement models." *Journal of Human Resources* **26**: 107-137

Christensen, K., A.M. Herskind, J.W. Vaulpel (2006), "Why Danes are Smug: Comparative study of life satisfaction in the European Union," *British Medical Journal* 333,1289-1291.

Disney, R., C. Emerson and M. Wakefield (2006). "Ill-health and retirement in Britain: A panel data based analysis." *Journal of Health Economics* **25**(4): 621-649

Etilé, F. and C. Milcent (2006). "Income-related reporting heterogeneity in self-assessed health: evidence from France." *Health Economics* **15**(9): 965-981

Fillenbaum, G.G., Hughes, G., Heyman, A., George, L.K., & Blazer, D.G. (1988). Relationship of Health and Demographic Characteristics to Mini-Mental State Examination Score among Community Residents. *Psychological Medicine*, 18, 719-726.

Gill, T.M., Desai, M.M., Gahbauer, E.A., Holford, T.R. and Williams, C.S. (2001), 'Restricted activity among community-living older persons: incidence, precipitants, and health care utilization', *Annals of Internal Medicine*, 135(5): 313–21.

Guralnik, J.M. and Ferrucci, L. (2003), 'Assessing the building blocks of function: utilizingmeasures of functional limitation', *American Journal of Preventive Medicine*, vol. 25, pp. 112–121.

Guralnik, J.M., Simonsick, E.M., Ferrucci, L., Glynn, R.J., Berkman, L.F., Blazer, D.G., Scherr, P.A. and Wallace, R.B. (1994), 'A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission, *Journal of Gerontology*, vol. 49,pp. M85–M94.

Hopkins, D. and G. King, 2008. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Comparability," *Kennedy School of Government, Harvard University*,mimeo http://gking.harvard.edu/files/implement.pdf

Huppert, F. A., Brayne, C., Gill, C., Paykel, E. S. & Beardsall, L.(1995). CAMCOG - a concise neuropsychological test to assist dementia diagnosis : socio-demographic determinants in an elderly population sample. *British Journal of Clinical Psychology* 34, 529-541.

Idler, E. and Y. Benyamini (1997). "Self-rated health and mortality: a review of twenty-seven community studies." *Journal of Health and Social Behavior* **38**(1): 21-37

Kapteyn, A., J. Smith and A. van Soest. "Vignettes and self-reports of work disability in the US and the Netherlands." *American Economic Review* 2007.

Kerkhofs, M. J. M. and M. Lindeboom (1995). "Subjective health measures and state dependent reporting errors." *Health Economics* **4**: 221-235

King, G., C. J. L. 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): 184-91

Kreider, B. (1999). "Latent work disability and reporting bias." *Journal of Human Resources* **34**(4): 734-769

Kristensen, N. & Johansson, E. "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes". *Labor Economics 2008; 15: 96-117.*

Lindeboom, M. and E. van Doorslaer (2004). "Cut-point chift and index shift in self-reported health." *Journal of Health Economics* **23**(6): 1083-1099

Lang IA, Llewellyn DJ, Langa KM, et al. Neighborhood deprivation, individual socioeconomic status, and cognitive function in older people: analyses from the English Longitudinal Study of Ageing. *J Am Geriatr Soc* (2008) 56:191–8.

Lee S., Kawachi I., Berkman LF, Grodstein F. Education, Other Socioeconomic Indicators, and Cognitive Function. *American Journal of Epidemiology* 2003; 157.

LlewellynDJ, Lang IA, Xie J, Huppert FA, Melzer D and Langa KM. Framingham Stroke Risk Profile and poor cognitive function: a population-based study. *BMC Neurology* 2008, 8:12.

MRC CFA Study (1998), 'Cognitive function and dementia in six areas of England and Wales: the distribution of MMSE and prevalence of GMS organicity level in the MRC CFA Study', *Psychological Medicine*, 28: 319–35.

Murray, C. J. L., E. Ozaltin, A. Tandon, J. Salomon, et al. (2003). Empirical evaluation of the anchoring vignettes approach in health surveys. *Health systems performance assessment: debates, methods and empiricism.* C. J. L. Murray and D. B. Evans. Geneva, World Health Organization.

Nunn S, Cox K, Wood N and Scholes S. (2008) *English Longitudinal Study of Ageing* (ELSA), Wave 3 Core Dataset, Phase 2 Deposit User Guide. Version 1.

Park, D. (1999). Cognitive aging, processing resources, and self-report. In N. Schwarz, D.C.Park, B. Knauper, & S. Sudman (Eds.), *Cognition, aging, and self-reports*. Philadelphia:Psychology Press.

Organisation for Economic Cooperation and Development (OECD) and Statistics Canada (2000), Literacy in the Information Age: Final Report of the International Adult Literacy Survey, Paris: OECD and Statistics Canada.

Ofstedal MB, Fisher GG Regula Herzog A (2005). Documentation of Cognitive Functioning Measures in the Health and Retirement Study. *HRS/AHEAD Documentation Report DR-006*. Survey Research Center. University of Michigan.

Reed, B.R., Jagust, W.J., & Seab, J.P. (1989). Mental status as a predictor of daily function in progressive dementia. *The Gerontologist*, 29, 804-807.

Richards M, Kuh D, Hardy R, Wadsworth M. Lifetime cognitive function and timing of the natural menopause. *Neurology* 1999; 52:308-14.

Roth M, Tym E, Mountjoy CQ, Huppert FA, Hendrie H, Verma S, Goddard R: CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *Br J Psychiatry* 1986, 149:698-709.

Rice, N, Robone, S, Smith, P.C. International Comparison of Public Sector Performance: The Use of Anchoring Vignettes to adjust Self-Reported Data. *University of York HEDG WP* 08/28. 2008.

Stern, S. (1989). "Measuring the effect of disability on labor force participation." *Journal of Human Resources* **24**: 361-395

Small GW. What we need to know about age related memory loss. *Br Med J* 2002; 324: 1502–5.

Statistics Canada and Organisation for Economic Cooperation and Development (OECD) (2005), *Learning a Living: First Results of the Adult Literacy and Life Skills Survey*, Ottawa and Paris: Statistics Canada and OECD.

Steel, N., F. Huppert, B. McWilliams and D. Melzer, 2004, 'Physical and cognitive function', in Health, wealth and lifestyles of the older population in England: THE 2002 ENGLISH LONGITUDINAL STUDY OF AGEING, Marmot, M., Banks, J., Blundell, R., Lessof, C. and Nazroo, J (eds.). Institute for Fiscal Studies. London.

Studenski, S., Perera, S., Wallace, D., Chandler, J.M., Duncan, P.W., Rooney, E., Fox, M. and Guralnik, J.M. (2003), 'Physical performance in the clinical setting', *Journal of the American Geriatric Society*, vol. 51, pp. 314–322.

Terza, J. V. (1985). "Ordinal probit: a generalization." *Communications in Statistics* **14**(1): 1-11

Van Doorslaer, E., X. Koolman and A. M. Jones (2004). "Explaining income-related inequalities in doctor utilization in Europe." *Health Economics* **13**(7): 629-647

Van Soest, A., L. Delaney, C.P. Harmon, A. Kapteyn, J. Smith. Validating the Use of Vignettes for Subjective Threshold Scales. *IZA Discussion Paper* No. 2860, 2007.

Wallace, R.B. and Herzog, A.R. (1995), 'Overview of the health measures in the Health and Retirement Study', *Journal of Human Resources*, 30 (5): S84–S107.

Wand, J. "Credible Comparisons Using Interpersonally Incomparable Data: Ranking self-valuations relative to anchoring vignettes or other common survey questions". *Department of Political Science*, *Stanford University*, mimeo. 2007.

Weuve, JH Kang, JE Manson, MMB Breteler, JH Ware and F Grodstein, Physical activity, including walking, and cognitive function in older women, *JAMA* 292 (2004), pp. 1454–1461.

Appendix Tables

Table A1: Sample sizes and item non-response

Full health self-completion sample	3,088	
Questionnaire not received	633	
Quest. received but resp. new in wave 3 (missing wealth, numeracy and literacy tests)	383	
Quest. received but resp. new in wave 2 (missing literacy test)	15	
Quest. received but resp. <60 (missing walking speed test)	914	
Observations lost due to item non-response	Cognition sample	Mobility sample
Self-reported measure	19	16
Vignettes	49	51
Covariates	98	62
Objective measures	159	91
Final samples	1782	1280

Table A2: Estimation results of Model 4 for cognition

	Coeff	**Index Std. Err.**	**z-stat**				
Vignette 1	-0.522	0.380	-1.370				
Vignette 2	-0.226	0.392	-0.580				
Vignette 2 *interacted with*				**Cut-point 1**	Coef.	Std. Err.	Z
Age 55 to 64	-0.252	0.181	-1.390		-0.152	0.135	-1.120
Age 65 to 74	-0.158	0.185	-0.860		0.028	0.136	0.210
Age 75+	-0.210	0.194	-1.080		0.126	0.141	0.900
Female	-0.112	0.088	-1.280		**-0.146**	0.062	-2.340
White	**-0.777**	0.361	-2.160		0.082	0.279	0.290
ln(Wealth)	-0.036	0.029	-1.230		**-0.081**	0.020	-4.110
No wealth	-0.197	0.410	-0.480		-0.498	0.274	-1.820
Qualifications 2	-0.207	0.110	-1.890		**-0.374**	0.079	-4.720
Qualifications 1	-0.173	0.111	-1.560		**-0.207**	0.078	-2.650
Not working <65	0.073	0.132	0.560		-0.028	0.099	-0.280
Vignette 3	-0.106	0.470	-0.230				
Vignette 3 *interacted with*				**Cut-point 2**			
Age 55 to 64	-0.269	0.239	-1.120		0.065	0.134	0.490
Age 65 to 74	0.081	0.238	0.340		0.177	0.136	1.300
Age 75+	0.322	0.244	1.320		**0.355**	0.144	2.470
Female	**-0.217**	0.106	-2.040		-0.039	0.065	-0.600
White	**-1.040**	0.385	-2.700		**0.586**	0.266	2.200
ln(Wealth)	**-0.084**	0.034	-2.510		-0.041	0.022	-1.850
No wealth	-0.069	0.456	-0.150		-0.590	0.308	-1.920
Qualifications 2	**-0.646**	0.136	-4.760		-0.117	0.082	-1.430
Qualifications 1	**-0.375**	0.133	-2.820		-0.045	0.084	-0.530
Not working <65	0.020	0.173	0.120		0.120	0.098	1.220
Constant					0.136	0.325	0.420
Number of obs = 1782							
Log likelihood = -3424.52							

Table A3: Estimation results of Model 4 for mobility

	Coeff	Index Std. Err.	z-stat				
Vignette 1	-0.495	0.552	-0.900				
Vignette 2	-0.080	0.492	-0.160				
Vignette 2 *interacted with*				**Cut-point 1**	Coef.	Std. Err.	Z
Age 65 to 74	0.251	0.274	0.920		0.102	0.201	0.510
Age 75+	0.068	0.278	0.240		-0.233	0.202	-1.160
Female	0.063	0.145	0.430		0.162	0.109	1.480
White	-0.201	0.481	-0.420		**0.845**	0.359	2.350
ln(Wealth)	-0.013	0.045	-0.300		-0.001	0.034	-0.030
No wealth	-0.472	0.642	-0.740		-0.287	0.478	-0.600
Qualifications 2	-0.091	0.193	-0.470		0.262	0.144	1.820
Qualifications 1	0.101	0.176	0.570		0.140	0.134	1.040
Not working <65	0.133	0.317	0.420		0.072	0.234	0.310
Vignette 3	0.138	0.402	0.340				
Vignette 3 *interacted with*				**Cut-point 2**			
Age 65 to 74	-0.007	0.232	-0.030		0.182	0.230	0.790
Age 75+	-0.365	0.236	-1.540		-0.152	0.233	-0.650
Female	0.236	0.128	1.840		0.211	0.127	1.660
White	0.673	0.464	1.450		**1.407**	0.375	3.750
ln(Wealth)	-0.006	0.040	-0.140		-0.022	0.040	-0.540
No wealth	-0.973	0.578	-1.680		-0.860	0.553	-1.560
Qualifications 2	**0.419**	0.166	2.520		**0.593**	0.168	3.520
Qualifications 1	0.274	0.157	1.740		**0.384**	0.157	2.440
Not working <65	-0.043	0.270	-0.160		0.297	0.272	1.090
Constant					0.496	0.319	1.560
Number of obs = 1280							
Log likelihood = -2065.33							