Slipping Anchor?

Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity

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

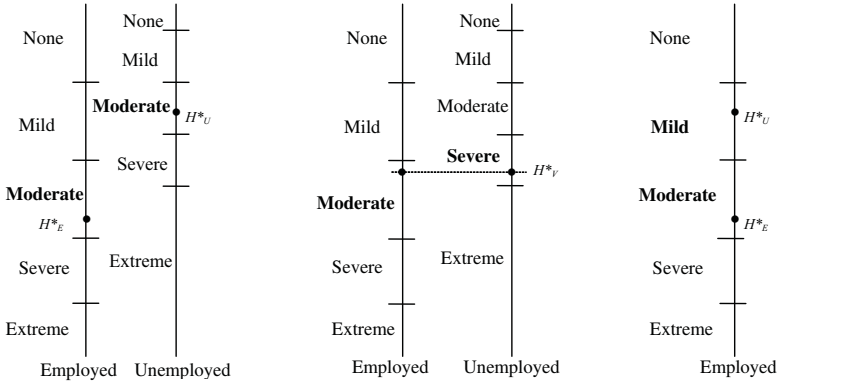
ABSTRACT

We propose tests of the two assumptions under which anchoring vignettes identify heterogeneity in reporting of categorical evaluations. Systematic variation in the perceived difference between any two vignette states is sufficient to reject vignette equivalence. Response consistency*—the respondent uses the same response scale to evaluate the vignette and herself—is testable given sufficiently comprehensive objective indicators that independently identify response scales. Both assumptions are rejected for reporting of cognitive and physical functioning in a sample of older English individuals, although a weaker test resting on less stringent assumptions does not reject* response consistency *for cognition.*

I. Introduction

Interpersonal comparability of subjective assessments of life satisfaction, health, political efficacy, etc. can be impeded by differences in reporting

Teresa Bago d'Uva is an assistant professor of applied economics at Erasmus University Rotterdam (the Netherlands). Maarten Lindeboom is a professor of economics at the Free University Amsterdam (the Netherlands). Owen O'Donnell is an associate professor of applied economics at the University of Macedonia (Greece) and Erasmus University Rotterdam. Eddy van Doorslaer is a professor of health economics at Erasmus University Rotterdam. *The authors wish to thank Patrick Hullegie and two anonymous referees for valuable comments on previous drafts. They are also grateful to Rob Alessie, Frank Windmeijer, and numerous seminar participants for comments. The research was funded by the NETSPAR theme "Health and income, work and care across the life cycle II", a VENI grant from the Netherlands Organisation for Scientific Research (NWO) (Bago d'Uva) and National Institute of Aging grant 1R01AG037398 (O'Donnell & Van Doorslaer). The ELSA data were made available through the U.K. Data Archive (UKDA). ELSA is funded by the US National Institute of Aging, and a consortium of U.K. government departments. The developers and funders of ELSA and the Archive do not bear any responsibility for the analyses or interpretations presented here. Researchers wishing to use the data analyzed in this article can apply to the UKDA (www.data-archive.ac.uk).*
[Submitted October 2009; accepted November 2010]
ISSN 022-166X E-ISSN 1548-8004



A: Hypothetical self-reported problems moving around for employed and unemployed individuals, with true latent mobility H^*_E and H^*_U respectively.

B: Hypothetical self-reported problems moving around for one vignette (Tom), with true latent mobility H^*_V.

C: Hypothetical self-reported problems moving around for employed and unemployed individuals using response scales of employed.

Figure 1
Hypothetical Heterogeneity by Employment Status in the Reporting of Physical Mobility and the Anchoring Vignette Correction

styles. A proposed solution is to anchor an individual's assessment of her own situation on her rating of a vignette description of a hypothetical situation (King et al. 2004). For example, respondents may be asked to rate, on an ordinal scale, the degree of difficulty in mobility experienced by the following hypothetical person: "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." Since the vignette is fixed for all respondents, variation in its rating is assumed to identify heterogeneity in reporting styles, which, it is argued, can then be purged from the individual's subjective assessment of her own situation. Identification of reporting behavior rests on the assumption that the vignette evokes the same picture of the underlying construct—mobility, in the example—for all respondents. Correction of reporting heterogeneity relies on the additional assumption that the same reporting behavior governs evaluation of both the vignette and the respondent's own situation. These two assumptions—labeled *vignette equivalence* and *response consistency* respectively by King et al. (2004)—have been subjected to very little formal testing. This paper introduces two tests of *response consistency* and one of *vignette equivalence*. It applies these tests to reporting of two distinct health domains—mobility and cognition—and finds evidence against the validity of the vignettes approach.

The problem of reporting heterogeneity and the vignette solution to it are illustrated in Figure 1, using hypothetical differences by employment status in the reporting of physical mobility. Panel A shows the mapping from latent true mobility (H^*) into categorical responses for an employed (E) and an unemployed (U) individual. All response thresholds are assumed higher for the unemployed person. She

is less constrained in mobility ($H_U^* > H_E^*$) but both report experiencing "moderate" problems moving around. Now suppose both individuals are confronted with the description of Tom given in the previous paragraph. Under the assumption of *vignette equivalence,* both interpret Tom as representing the same latent mobility functioning, H_V^* in Panel B. The unemployed person rates Tom as experiencing "severe" problems with mobility, while the employed person reports that Tom has only "mild" problems. This identifies reporting heterogeneity. If *response consistency* holds, the reporting thresholds identified from the rating of the vignette can be imposed on the reporting of own mobility. Standardizing on the thresholds of the employed person, the unemployed person's degree of difficulty in mobility is corrected to "mild" (Panel C).

If *vignette equivalence* does not hold, such that perceptions of the construct evoked by the vignette description vary with characteristics suspected of influencing reporting styles, then one cannot attribute systematic variation in vignette ratings to reporting heterogeneity. Respondents are then not reporting on the same state differently but are reporting on different perceived states. In relation to Figure 1, this means that the description of Tom evokes a different picture of mobility (H_V^*) for the employed and the unemployed person. This may happen if vignette descriptions are incomplete, and/or equivocal, and groups of individuals complement those descriptions in different ways.

If *response consistency* does not hold, such that response scales used to rate the vignettes and the individual's own situation differ, then the information obtained from the vignette responses is of no use in improving interpersonal comparability. This assumption will not hold if there are strategic influences on the reporting of the individual's own situation that are absent from evaluation of the vignette. For example, nonworking individuals may experience social pressure and/or financial incentives to understate their own health but not that of hypothetical individuals portrayed by the vignettes. The approach would not then correct fully the justification bias that has plagued estimates of the impact of health on labor market participation of older individuals (Stern 1989; Bound 1991; Kerkhofs and Lindeboom 1995; Benitez-Silva et al. 1999; Kreider 1999). To the extent that vignettes can detect employment related subconscious revisions to the general conception of work capacity, it may nonetheless shift estimates in the right direction.

Vignette equivalence has not previously been formally tested. We construct a test from the observation that if there is no systematic variation in perceptions of the state represented by each vignette, then there must be no such variation in the perceived difference between states corresponding to any two vignettes. This necessary condition can be tested with any dataset containing at least two vignettes for a given construct. With the same data requirement, but preferably with a greater number of vignettes, Murray et al. (2003) propose an informal check on the plausibility of *vignette equivalence* by examining whether there are systematic differences in the ranking of vignettes (see also Rice, Robone, and Smith 2011). This is likely to be successful in detecting extreme violations of the assumption, occurring when types differ greatly in their perceptions of the vignettes, but it will not be as powerful as our test in identifying less marked differences.

Our test of *response consistency* is feasible when, in addition to the vignettes, data are available on objective indicators sufficiently rich such that they can be

presumed to capture all covariation between the construct of interest and the observable characteristics influencing reporting behavior. 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.

Van Soest et al. (2011) 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. Since, in some circumstances, even multiple objective indicators may be insufficient to absorb all this covariation, we introduce a second test that is valid even in the presence of such covariation and when *vignette equivalence* does not hold. This is a weaker test, in the sense that it tests a necessary condition for *response consistency*—that differences between adjacent reporting thresholds identified using objective indicators are equal to those identified from vignettes. Its robustness makes it a valuable additional tool for evaluating the validity of the vignettes approach.

Vignettes are being fielded in a growing number of household surveys, including the *English Longitudinal Study of Ageing* (ELSA), which we use in this paper, and the *Health and Retirement Study* (HRS). Applications of the methodology are increasing rapidly and now cover a wide range of topics including political efficacy (King et al. 2004), work disability (Kapteyn, Smith, and Van Soest 2007), job satisfaction (Kristensen and Johansson 2008), life satisfaction (Christensen et al. 2006), health (Bago d'Uva et al. 2008; Bago d'Uva, O'Donnell, and Van Doorslaer 2008) and health system responsiveness (Rice, Robone, and Smith 2010). These studies typically claim to 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, based on tests of its identifying assumptions, the appropriateness and accuracy of such 'corrections' remain in doubt. An informal check on the performance of the method can be made by assessing whether vignette corrections bring self-reports closer, in some sense, to an objective measure of the construct of interest (King et al. 2004; Van Soest et al. 2011; Vonková and Hullegie 2011). While helpful in assessing face validity, this does not establish whether the method succeeds in identifying reporting heterogeneity. The latter can only be determined by testing the veracity of the identifying assumptions.

We apply our tests of the validity of the methodology to a mental and a physical domain of health—cognitive functioning and mobility respectively. Importantly for our test of *response consistency*, well validated instruments exist for both dimensions of health and we observe these in the ELSA data. Available objective proxies for cognitive functioning include 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.

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 III presents the main contribution of the paper—the tests for *vignette equivalence* and *response consistency.* In section IV we describe the data, in particular the vignettes and the objective indicators for cognitive functioning and mobility. Results are presented in section V and the final section concludes.

II. Identification of reporting heterogeneity

For ease of exposition and given the application that follows, we will refer to the underlying concept of interest as 'health'.

A. 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, as in the example presented in Figure 1 for mobility. 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:

$$(1a) \quad H_i^* = \beta X_i + \varepsilon_i$$

$$(1b) \quad H_i^s = k \Leftrightarrow \tau_i^{k-1} \leq H_i^* < \tau_i^k$$

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

It is assumed that researchers are ultimately interested in the extent to which true health varies across populations or subgroups (the parameter vector β).[1] The problem is that the relationship between H^* and H^S may not be constant across populations, as was illustrated in Figure 1. 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:[2]

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

1. Consistent with all published applications of the vignettes methodology, we are concerned with correcting systematic error in the reporting of a variable the researcher wishes to compare across populations or groups, and not with correcting a mismeasured independent variable. The latter would introduce concern about random, in addition to systematic, error (Bound 1991).

2. An alternative is to define the first cut point as here but the following ones as: $\tau_i^k = \tau_i^{k-1} + \exp(X_i \gamma^k) \quad k = 2, \ldots, K-1$ (Kapteyn, Smith, and Van Soest 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.

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

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

where F(.) is the distribution function of the error term ε. It is apparent that it is not possible to identify simultaneously all γ^k and β.[3] 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.

In some circumstances an effect on either the latent construct or the reporting thresholds can be ruled out a priori. For example, work capacity depends not only on ability to perform selected tasks but also on the relevance of those tasks to the individual's occupation. It is then legitimate that reported work capacity varies with occupation for given measured ability in standardized tasks. In this case, there is no identification problem. Occupation can be excluded from the reporting thresholds and included in the latent index.

B. Identifying reporting heterogeneity: Anchoring with vignettes

Vignettes are descriptions of hypothetical health states, such as that provided in the first paragraph of this paper, 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 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,

$$\text{(A1)} \quad f(V_j^* \mid X) = f(V_j^*).$$

Then, the latent health of vignette j as perceived by individual i can be specified as an intercept (α_j) plus random measurement error (ξ_{ij}),[4]

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

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

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

3. Identification of a restricted model that arbitrarily excludes covariates from one cut point is possible (Terza 1985).

4. If, unlike in our application, gender varies across vignette descriptions, then one could allow the intercept to shift with gender, or any other background characteristic revealed in the vignette description.

$k=1,\ldots,K$, $\nu_i^0<\nu_i^1<\ldots<\nu_i^{K-1}<\nu_i^K$ and $\nu_i^0=-\infty$, $\nu_i^K=\infty$. As before, differential reporting behavior is reflected in differences in the cut points ν_i^k across individuals. Note that when, as in the applications that follow, individuals report on more than one vignette relating to a given construct, it is assumed they use the same thresholds for all vignettes. This follows from the *response consistency* assumption. So, the thresholds employed to report the mobility of Tom, described in the first paragraph of the paper, are assumed to be the same as those used to report on Robert, who 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."

By design, the vignettes are intended to describe different levels of the latent construct—Tom is less mobile than Robert. While sets of thresholds are assumed consistent across vignettes, the amount of information provided to identify any given threshold should vary across vignettes.

Like in Equation 1c, we can specify the cut points as linear functions of the individual characteristics:[5]

$$\text{(2c)} \quad \nu_i^k=\gamma_v^k X_i.$$

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

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

Under Assumptions A1 and A2, the vignettes ratings can be used to identify reporting behavior (γ^k) via Equations 2a–2c and so permit to test the null of reporting homogeneity:

$$\text{(RH)} \quad \tau_i^k=\tau^k \quad k=1,\ldots,K-1.$$

The reporting thresholds can 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 (see Table 1).

5. With evaluations of multiple vignettes it is possible to allow for unobserved heterogeneity in the response scale (Kapteyn, Smith, and Van Soest 2007). We have not done so both because identification of the random individual effect is weak in our application (possibly due to the limited number of vignettes) and because this effect is not identified in the proxy indicators model, making its introduction inappropriate within the context of our tests of response consistency.

Table 1
Models, Tests and Maintained Assumptions

			Maintained Assumptions	
Model	Test	Null	Objective component	Vignettes component
1:(2a)–(2c)	Reporting homogeneity (RH)	$\tau_i^k = \tau^k,\ k = 1,\ldots,K-1$	—	(A1): $f(V_j^* \mid X) = f(V_j^*)$ (A2): $\gamma^k = \gamma_v^k,\ k = 1,\ldots,K-1$
2:(3a)–(3c)	Reporting homogeneity (RH)	$\tau_i^k = \tau^k,\ k = 1,\ldots,K-1$	(A3): $h(H^* \mid H^O, X) = h(H^* \mid H^O)$	—
3:(3a)–(3c) (2a)–(2c)	Response consistency 1 (RC1)	$\gamma^k = \gamma_v^k,\ k = 1,\ldots,K-1$	(A3): $h(H^* \mid H^O, X) = h(H^* \mid H^O)$	(A1): $f(V_j^* \mid X) = f(V_j^*)$
3:(3a)–(3c) (2a)–(2c)	Response consistency 2 (RC2)	$\gamma^k - \gamma^{k-1} = \gamma_v^k - \gamma_v^{k-1}$ $k = 2,\ldots,K-1$	(A3′): $h(H^* \mid H^O, X)$ homoscedastic in X	(A1′): $f(V_j^* \mid X)$ homoscedastic in X
4:(2a′), (2b)–(2c)	Vignette Equivalence (VE)	$\lambda_j = 0 \quad \forall j$	—	(A4): $\gamma_j^k = \gamma_v^k \quad \forall j$

C. Identifying reporting heterogeneity: Objective proxy measures

An alternative approach is to consider a sufficiently comprehensive set of proxy indicators of health (H^O) that are believed to be insensitive to reporting behavior. These could include physical examinations, medical tests, scores from validated instruments, and even self-reported medical conditions and functioning in specified activities, provided the latter are sufficiently narrowly defined such that they can be presumed to be reported without systematic error. Let $h(.)$ be the density function of latent health, then reporting heterogeneity can be identified if:

$$\text{(A3)} \quad h(H^* \mid H^O, X) = h(H^* \mid H^O).$$

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 (Stern 1989; Kerkhofs and Lindeboom 1995; Kreider 1999). There is a potentially nonlinear relationship between latent true health and the proxy indicators as follows:

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

where $g(\,)$ is a sufficiently flexible function that is the same for all individuals 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 Equations 3a, 1b, and 1c, which we refer to as Model 2.

Subject to Equation A3, the parameters of Equation 1c reflect only reporting heterogeneity. Otherwise, these parameters will reflect a mixture of reporting and true health effects. If these effects operate in the same direction—the covariate is associated positively (negatively) with true health and with a tendency to overstate (understate) health—then the estimated coefficient will be an upper bound on the magnitude of the reporting effect. On the other hand, if the effects offset one another, then a lower bound on the magnitude of the reporting effect will be obtained. In both cases, the bias will be smaller the greater is the association between the covariate and true health that is absorbed by the objective indicators.[6]

III. Tests of response consistency and vignette equivalence

A. 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* Equation A2 can then be tested by comparing the estimates of the cut points obtained from Model 1, which are identified without using subjective evaluations of own health status, 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:

6. We thank a referee for pointing out this bounding argument.

1. Response Consistency 1: Equality of cut points

(RC1) $\gamma^k = \gamma_v^k, \quad k = 1, \ldots, K-1.$

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 $\gamma^{'k}$ 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 Equations 3a and 2a, respectively. But even in that case, because the parameter vectors β_s and β_v are not cut point specific, we have $\gamma_v^k - \gamma_v^j = (\gamma_v^{'k} - \beta_s) - (\gamma_v^{'j} - \beta_v) = \gamma_v^{'k} - \gamma_v^{'j}$ and $\gamma_v^k - \gamma_v^j = (\gamma_v^{'k} - \beta_s) - (\gamma_v^{'j} - \beta_v) = \gamma_v^{'k} - \gamma_v^{'j}$. That is, Model 3 still identifies the distance between any two cut points. The equality of these distances in both approaches is a necessary condition for each cut point in the proxy indicators model to be the same as the corresponding one in the vignettes model, ie, for *response consistency*. Even if the combined Model 3 is too restrictive, in the sense that A1 and/or A3 is violated, it still permits testing of that condition. This leads to a second, more robust, test that is valid even when the identifying assumptions of RC1 do not hold. This is, however, less informative than the first in the sense that nonrejection of the null does not imply that *response consistency* holds.

2. Response Consistency 2: Equality of distances between cut points

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

Van Soest et al. (2011) 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. For health—even a single domain of health—this may be demanding. There is seldom 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 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.

B. Vignette equivalence

Vignette equivalence rules out any systematic differences in the perception of the health level described by any vignette. This is imposed in order that the covariates, *X,* can be excluded from Equation 2a and so their effects on the cut points Equation 2c are identified. We exploit a less restrictive specification of Equation 2a, which relaxes a necessary condition for *vignette equivalence*, while still being identified.

The necessary condition is that there is no systematic variation in the perceived difference between the levels of health represented by any two vignettes. This can be tested in the following specification that includes interactions between individual characteristics and all but one vignette:

$$(2a') \quad V^*_{i1} = \alpha_1 + \upsilon_{i1}$$
$$V^*_{ij} = \alpha_j + \lambda_j X^-_i + \upsilon_{ij} \qquad j \neq 1$$

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. Significantly nonzero 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* (VE): $\lambda_j = 0 \ \ \forall j$, which is tested in a model composed by Equations 2a′, 2b, and 2c, which we refer to as Model 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 shift, γ^k_v, depends on the particular vignette that is used as the reference in (2a′) and is therefore not meaningful.

The test rests on the (*response consistency*) 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 nonzero λ were driven by different cut points, rather than by vignette nonequivalence, that would still be evidence against the validity of using the HOPIT model.

C. Distributional assumptions and normalizations

The models, tests and the maintained assumptions required for the validity of each test are summarized in Table 1. All models are estimated by maximum likelihood. The tests of reporting homogeneity are not conducted from separate estimation of Models 1 and 2 but from estimates obtained from the combined Model 3. 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.

Estimation of Models 3 and 4 requires specification of the error distributions and normalization of location and scale parameters. The location parameters are normalized by excluding the constant terms from the first cut points (ν^1_i and τ^1_i). 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 normalized, which is usually done by setting them equal to one. Estimation of parameters of interest in Model 3, as well as results of the *vignette equivalence* test (Model 1 vs. Model 4), are not affected by these normalizations. 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 normalize only $\sigma_\xi = 1$ and maximize 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 maximize the likelihood with respect to γ^k and γ_v^k, normalizing both σ_ξ and σ_η. Response consistency is tested using likelihood ratio tests, and so test statistics do not depend on these normalizations.

IV. Data

The *English Longitudinal Study of Ageing* (ELSA) samples individuals aged 50 and over and their younger partners, living in private households in England. We use data taken mainly from the third wave, 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, which excluded proxy respondents. The vignettes questionnaire consisted of two sections: one which asked respondents to rate their own health on a five-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 five-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.

A. 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 (Reed, Jagust, and Seab 1989; Park 1999; Gill et al. 2001; Steel et al. 2004), 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. They are also two health domains for which anchoring vignettes have revealed reporting heterogeneity (Bago d'Uva et al. 2008; Bago d'Uva, O'Donnell, and Van Doorslaer 2008).

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:

- 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 meters without any problems but feels tired after walking one kilometer 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—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.
- Mobility 3—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.

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.

B. 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. We use all the tests implemented in Wave 3, and the numeracy and literacy tests performed on the same individuals in previous waves. These 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). 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 also has been used in HRS. It was taken from the Mini Mental Status Examination (MMSE), which is validated, widely used and considered as the "gold standard" of cognitive impairment screening tests (Folstein, Folstein, and McHugh 1975; Weuve et al. 2004).

(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

Table 2
Frequencies of Reported Difficulty with Cognitive Functioning and Mobility

	Cognition				Mobility			
		Respondent's Assessment of each Vignette				Respondent's Assessment of each Vignette		
Degree of Difficulty	Respondent's Assessment of Own Functioning	1	2	3	Respondent's Assessment of Own Functioning	1	2	3
At least moderate	316	419	1,318	1,646	269	637	1,155	1,198
Mild	735	1,078	403	96	254	545	77	48
None	731	285	61	40	757	98	48	34
Number of observations	1,782	1,782	1,782	1,782	1,280	1,280	1,280	1,280

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

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. They also are told that they will need to carry this out without being reminded of what they must do. The action (initialing a page on a clipboard) is based on a similar task used in the Medical Research Council (MRC) Cognitive Function and Ageing Study (MRC CFA Study 1998).

(5) Word-finding and 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, organization, and abstraction and set-shifting. This test was taken from the Cambridge Cognitive Examination (Huppert et al. 1995) 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 (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 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 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 (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). The problems were developed for ELSA and later used in HRS.

(9) Literacy: This test aimed at deriving a measure of prose literacy relevant to the lives of the elderly. Participants were shown a realistic label for a fictitious medicine and then asked questions to test understanding of the instructions on the label. 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.

C. Mobility indicators

We use results from a measured test of walking speed, administered within the ELSA survey to respondents aged 60 or over for whom the test is judged safe. Impaired

Table 3
Descriptive Statistics of Health Measures and Sociodemographic Variables

	Cognition sample		Mobility sample	
	Mean	Standard deviation	Mean	Standard deviation
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 Activity of Daily Living (ADL) limitation			0.110	0.313
2+ ADL limitations			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
Sociodemographic 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
Log 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
Number of observations	1,782		1,280	

Notes: All cognitive test scores are rescaled 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.

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 eight 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. 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 as objective health measures in, for example, Kreider (1999). We include dummy variables indicating the number of items with which the individual reports difficulties, collapsing those referring to five or more items as the respective estimated effects differed little and not significantly so.

The original scale of ADLs (Katz et al. 1963) includes activities which are likely to be part of the lives of most people. 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, or with two or more 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.

D. Sociodemographic 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. 2008; Bago d'Uva, O'Donnell, and Van Doorslaer 2008). ELSA provides a very accurate mea-

sure of wealth, which Banks et al. (2006) have found to be negatively associated with sickness, impaired functioning, and mortality. Cultural differences across ethnic groups may influence concepts and reporting of health. Testing *response consistency* by employment status is particularly interesting given the concern expressed in the introduction about the ability of vignettes to correct *justification bias* in the estimated effect of health on employment.

Age is represented by age-group dummies and ethnicity by a dummy to distinguish between *Whites* and ethnic minorities. The variable *Log Wealth* is the logarithm of total nonpension wealth, set to zero for individuals with nonpositive 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. An indicator of whether individuals are younger than 65 *and* are not working (*Not working* <65) aims to capture any effect of employment status for individuals below normal retirement age—those who may have an incentive to underreport health as a justification for not working. Because it is unlikely that individuals aged 65+ behave similarly and because the proportion older than 64 who work is small, the reference group includes individuals younger than 65 who are working and those aged 65 or older (regardless of working status). Because our age variables discriminate between individuals older and younger than 65, the effect of *Not working* <65 will actually represent, for those younger than 65, the effect of not working.

For the analysis of cognition, we drop observations with missing data on self-reported cognition (19), respective vignettes (49), the cognitive tests (159) and the sociodemographic variables (98). The resulting dataset contains 1,782 individuals aged 50 and older. In the case of mobility, we dropped individuals younger than 60, who did not perform the walking speed test, and those without full item response on the self-reports (16), vignettes (51), objective indicators (91) and covariates (62), but did not drop those with missing information for cognition, leading to a dataset with 1280 individuals. 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 (383), as part of the refreshment sample added to ELSA. For the cognition analysis, 15 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.

V. Results

A. Reporting heterogeneity

1. Cognition

Estimates of the combined vignettes and proxy indicators model (Model 3) for cognition are presented in Table 4. Estimates of the index function parameters in the top lefthand panel confirm that all test scores are positively correlated with cognitive

Table 4

Effects on Latent Cognition and Response Scales Estimated from Combined Vignettes and Proxy Indicator Model (Model 3)

Latent cognition as function of test scores			Latent cognition of vignettes		
Orientation	0.371	(0.262)			
Immediate memory	0.515**	(0.259)	Vignette 1	0.616**	(0.253)
Short-term memory	1.028***	(0.223)			
Prospective memory	0.067	(0.094)			
Word-finding & verbal fluency	0.781**	(0.327)	Vignette 2	−0.615**	(0.253)
Processing speed	0.252	(0.335)			
Search accuracy	0.138	(0.275)			
Numeracy	0.133	(0.184)	Vignette 3	−1.351***	(0.255)
Literacy	0.067	(0.145)			
Constant	−1.717***	(0.607)			
σ_η	1.132	(fixed)	σ_ξ	1	(fixed)
Response scales identified from test scores			Response scales from vignettes		
Cut point-moderate/severe/extreme					
Age 55 to 64	−0.092	(0.223)		0.000	(0.087)
Age 65 to 74	0.208	(0.218)		0.072	(0.089)
Age 75+	0.567**	(0.222)		0.132	(0.092)
Female	−0.127	(0.090)		−0.056	(0.041)
White	−0.189	(0.377)		0.663***	(0.178)
Log Wealth	−0.084***	(0.025)		−0.050***	(0.014)
No wealth	−0.664*	(0.351)		−0.440**	(0.192)
Qualifications 2	−0.147	(0.112)		−0.156***	(0.052)
Qualifications 1	−0.134	(0.108)		−0.061	(0.052)
Not working < 65	0.398***	(0.149)		−0.059	(0.062)
Cut point-mild					
Age 55 to 64	−0.105	(0.152)		0.153	(0.118)
Age 65 to 74	0.090	(0.155)		0.206*	(0.120)
Age 75+	0.422**	(0.167)		0.361***	(0.128)
Female	0.017	(0.077)		0.006	(0.059)
White	−0.223	(0.341)		1.057***	(0.191)
Log Wealth	−0.013	(0.024)		−0.021	(0.020)
No wealth	0.208	(0.343)		−0.580**	(0.276)
A-level or above	0.127	(0.094)		0.000	(0.074)
Qualification < A-level	0.081	(0.093)		0.051	(0.076)
Not working < 65	0.320***	(0.107)		0.095	(0.088)
Constant	0.529	(0.527)		0.437	(0.334)
Log likelihood	−5,178.03				
Number of observations	1,782				

Notes: Standard errors in parenthesis. *, ** and *** indicate significance at 10 percent, 5 percent and 1 percent respectively. Coefficients are identified up to the scale parameters σ_η and σ_ξ. Estimates presented are with normalization $\sigma_\xi = 1$ and σ_η equal to the estimate under the null hypothesis of RC1. Inference does not depend on the chosen normalization.

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).[7]

The lower panels of Table 4 give effects on the reporting cut points identified from both the test scores and the vignettes. A positive coefficient indicates a greater probability of reporting difficulty concentrating or remembering things. There is evidence of reporting heterogeneity using both means of identification. This is confirmed in Table 6 by the strong rejection of reporting homogeneity jointly for all variables and for certain categories. Both approaches indicate that reporting differs by wealth and age, while employment status only affects the cut points identified from test scores. Only the vignettes reveal reporting differences by education and ethnicity. Wealth lowers the first cut point—greater wealth reduces the likelihood that a given level of cognitive functioning will be reported as corresponding to at least a moderate degree of difficulty. Using the vignettes, there is evidence of a nonlinear effect—those with no wealth are also less likely to report mild or moderate difficulties with concentration or memory. In line with findings from other data (Bago d'Uva et al. 2008), the oldest individuals rate a given level of cognitive functioning as corresponding to a greater degree of difficulty. Also consistent with other evidence (Bago d'Uva et al. 2008; Bago d'Uva, O'Donnell, and Van Doorslaer 2008), the vignettes approach indicates 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. It could be that educated individuals are less willing to admit cognitive impairment. Whites tend to rate the vignettes as more cognitively impaired, which would suggest that observed ethnic differences in cognitive functioning understate true differences. However, this is not confirmed when reporting behavior is identified from the test scores. Nonworking (< 65 years old) individuals are more likely to declare difficulty with cognitive functioning, given measured test scores, but they do not apply the same strict criteria to rating of the vignettes. This is consistent with our hypothesis that nonemployment may introduce a justification bias to the reporting of health that is not captured by the vignettes approach.

2. Mobility

Walking speed and each of the indicators of ADL and motor skills are significantly correlated with latent mobility (Table 5, top left panel). Homogeneity in the reporting of mobility across all covariates is also strongly rejected (Table 6). The nature of the heterogeneity differs from that observed for cognition in several respects (Table 5, bottom panel and Table 6). The proxy indicators approach reveals cut point shift by gender, ethnicity, and wealth, while differential rating of vignettes is observed only by ethnicity and education. Females and the less wealthy are more likely to rate their own mobility positively but not that of the vignettes. Better educated individuals are less likely to consider the mobility level of a vignette as correspond-

7. We experimented with quadratic specifications for the test scores. The square terms were not jointly significant and the test outcomes are not affected by their exclusion.

Table 5
Effects on Latent Mobility and Response Scales Estimated from Combined Vignettes and Proxy Indicator Model (Model 3)

Latent mobility as function of proxy indicators			Latent mobility of vignettes		
Walking speed	−0.364***	(0.061)			
Walking speed squared	0.009***	(0.003)	Vignette 1	0.610**	(0.279)
1 ADL limitation	−0.353**	(0.143)			
2+ ADL limitations	−0.606***	(0.189)			
1 motor problem	−0.693***	(0.135)	Vignette 2	−0.528*	(0.279)
2 motor problems	−1.294***	(0.156)			
3 motor problems	−1.586***	(0.168)			
4 motor problems	−1.878***	(0.188)	Vignette 3	−0.755***	(0.280)
5+ motor problems	−2.316***	(0.169)			
Constant	4.166***	(0.739)			
σ_η	1.182	(fixed)	σ_ξ	1.000	(fixed)
Response scales identified from proxy indicators			**Response scales from vignettes**		
Cut point-moderate/severe/extreme					
Age 65 to 74	0.318	(0.270)		0.041	(0.090)
Age 75+	0.367	(0.275)		−0.061	(0.094)
Female	−0.304**	(0.121)		0.018	(0.051)
White	−0.717	(0.521)		0.614***	(0.200)
Log wealth	0.106***	(0.037)		0.005	(0.016)
No wealth	0.747	(0.521)		0.325	(0.240)
A-level or above	−0.010	(0.150)		0.056	(0.065)
Qualification < A-level	0.031	(0.147)		−0.026	(0.062)
Not working < 65	0.308	(0.298)		0.060	(0.104)
Cut point-mild					
Age 65 to 74	0.052	(0.187)		0.139	(0.129)
Age 75+	0.204	(0.197)		0.066	(0.135)
Female	−0.281***	(0.105)		0.048	(0.075)
White	−1.075**	(0.443)		1.152***	(0.217)
Log wealth	0.016	(0.035)		−0.016	(0.024)
No wealth	0.493	(0.509)		−0.245	(0.349)
A-level or above	0.055	(0.129)		0.337***	(0.099)
Qualification < A-level	0.002	(0.126)		0.203**	(0.095)
Not working < 65	0.047	(0.213)		0.294*	(0.156)
Constant	2.676***	(0.769)		0.523	(0.320)
Log likelihood	−2,948.89				
Number of observations	1,280				

Notes: As for Table 4.

ing to no difficulty, while there is no evidence of heterogeneous reporting of own mobility by education. The disparity between the approaches in the identification of cut point shift by ethnicity that was observed for cognition is confirmed for mobility. There is a clear tendency for *Whites* to be optimistic in reporting their own mobility,

Table 6
Likelihood Ratio Tests of Reporting Homogeneity (RH)

	Degrees of freedom	Proxy Indicators Model		Vignettes Model	
		Test statistic	*p*-value	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	2	12.36	0.002	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	2	1.15	0.562	3.52	0.172

while being pessimistic in reporting that of the vignettes. Unlike for cognition, estimated differences in the reporting of mobility by employment status are not consistent with the hypothesis that vignettes will fail to identify justification bias.

B. Global tests of response consistency and vignette equivalence

There is clearly evidence of heterogeneity in the reporting of both cognition and mobility. But the nature of this heterogeneity by covariates appears to differ depending upon whether response scales are identified directly from vignettes, or indirectly through proxy indicators. This suggests that *response consistency* does not hold, which is confirmed, for both health domains, using the stricter test of the assumption (RC1) applied jointly across all covariates (Table 7). The weaker test (RC2) applied to all covariates jointly also rejects the null for mobility, but not for cognition. Given that RC2 is a valid test even if Assumption A1 or A3 does not hold, this is strong evidence against *response consistency* in the domain of mobility. The discrepancy between the outcomes of RC1 and RC2 for the reporting of cognition may be because: (1) RC2 only tests a necessary condition; or, (2) the assumptions required for RC1 to be valid do not hold. In Case 1, the response scales used for reporting own cognition and that of the vignettes do indeed differ. *Response consistency* does not hold, although the distance between any two cut points is equal across the scales. In Case 2, the response scales are the same—*response consistency*

Table 7
Likelihood Ratio Tests of Response Consistency and Vignette Equivalence

	Cognition			Mobility		
Test	Degrees of freedom	Test statistic	*p*-value	Degrees of freedom	Test statistic	*p*-value
Response consistency 1 (RC1)						
All variables	21	43.26	0.003	19	55.18	<0.001
Age	6	15.14	0.019	4	4.31	0.365
Female	2	0.66	0.718	2	8.83	0.012
White	2	11.47	0.003	2	20.17	<0.001
Wealth	4	10.84	0.029	4	14.48	0.006
Education	4	1.80	0.773	4	4.31	0.366
Not working	2	8.85	0.012	2	2.42	0.298
Response consistency 2 (RC2)						
All variables	11	13.10	0.287	10	20.35	0.026
Age	3	2.32	0.510	2	1.62	0.445
Female	1	0.84	0.358	1	<0.01	0.990
White	1	0.67	0.413	1	3.07	0.080
Wealth	2	5.52	0.063	2	9.70	0.008
Education	2	1.64	0.441	2	2.88	0.237
Not working	1	1.33	0.248	1	2.40	0.121
Vignette equivalence (VE)						
All variables	20	105.56	<0.001	18	44.66	0.001
Age	6	16.84	0.010	4	8.83	0.066
Female	2	4.43	0.109	2	4.35	0.114
White	2	8.06	0.018	2	4.38	0.112
Wealth	4	20.58	0.000	4	7.52	0.111
Education	4	23.98	0.000	4	14.56	0.006
Not working	2	0.32	0.852	2	0.48	0.786

holds—but Equations A1 and/or A3 are too restrictive such that covariates should appear in the index for latent cognition. Strictly, it is not possible to distinguish between these explanations but testing *vignette equivalence* (Equation A1) can help determine which is more plausible.

We first follow Murray et al. (2003) in examining consistency across respondents in their ordering of the vignettes. In both domains, Vignette 1 (3) is intended to represent the least (greatest) degree of difficulty. On average, respondents concur with this ordering. We define a respondent's ordering as consistent if it does not involve Vignette 3 (2) being rated as experiencing less difficulty than Vignette 2 (1). The degree of consistency is very high: 93.4 percent for cognition and 94.5 percent for mobility. While it is reassuring that variation in the rankings is limited, it would be problematic for the vignettes method if that which exists is systematic. To check this, we estimate logit regressions of an indicator of whether a respondent's ordering

is inconsistent on the covariates included in our models. There is no evidence of systematic variation in the ordering of the mobility vignettes (p-value of test of joint significance = 0.530), but ordering of the cognition vignettes do vary, at least to some degree, with covariates (p-value < 0.001).[8] Greater systematic variation in the interpretation of descriptions of cognitive, as opposed to physical, functioning is probably to be expected.

Even if vignette orderings were entirely consistent, this would not guarantee that *vignette equivalence* holds. Two individuals may differ in their interpretations of each vignette, resulting in different sets of ratings, while remaining consistent in their rankings across the vignettes. We therefore turn to our test of a necessary condition for *vignette equivalence*: no systematic variation in the perceived latent health of two of the three vignettes. 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 7).[9] 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 cognitive functioning 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. This suggests that there are systematic differences in the interpretation of the mobility vignettes that do not result in differences in their ordering. The evidence against the vignette approach in the domain of mobility is compelling—all three null hypotheses are decisively rejected.

C. Response consistency and vignette equivalence by covariate

We now test *vignette equivalence* and *response consistency* in relation to each covariate in order to assess whether the vignette approach may adequately correct for reporting heterogeneity in relation to a particular characteristic, even if it fails in general. Relative to the general unrestricted model that allows cut point coefficients of all covariates to differ between Model 1 and Model 2, we impose restrictions, defined by the null of each test, on the parameters corresponding to each group of covariates. The discussion here concentrates on the test outcomes. In the next subsection we illustrate implications for the magnitude and direction of adjustment for reporting heterogeneity using both methods.

In the case of cognition, at least one test rejects its null for all covariates with the singular exception of gender (Table 7). The stronger test of *response consistency* (RC1) rejects this assumption with respect to age, wealth, work status, and ethnicity, while the weaker test (RC2) rejects only for wealth, and then only at the 10 percent level. Since neither test rejects the null for education, the vignettes approach would appear to appropriately correct for differences by education in the reporting of cognition. However, *vignette equivalence* is rejected for education and all other factors, except for sex (although it is marginal) and employment status. RC1 strongly rejects the null for ethnicity, a consequence of the large differences in the ethnicity specific

8. Estimates from these logit regressions are available on request.
9. The coefficient estimates from Model 4, on which the test is based, are available on request.

cut points identified from the vignettes and the test scores observed in Table 4. The RC2 test does not show evidence against *response consistency* by ethnicity, suggesting that it is *vignette nonequivalence* that is the problem.

The vignette and proxy indicator approaches also do not concur with respect to reporting of cognition by employment status, resulting in rejection of *response consistency* by RC1. This is consistent with our *a priori* expectation that reporting on vignettes would not be helpful in correcting for justification bias. However, RC2 does not reject *response consistency*. Since *vignette equivalence* is also not rejected, it is possible that the assumptions of the vignettes approach hold but that of the proxy indicators approach fails. That is, employment remains correlated with true cognition even after conditioning on all cognitive functioning test scores. The comprehensiveness of these scores 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 favorable to working individuals, which would then be reflected as positive cut point shift in the proxy indicators model.

For mobility, across the three tests there is evidence against at least one null for every covariate except employment status. The exception is interesting since mobility-related problems are an important reason given for labor force withdrawal and, unlike for cognition, 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 5).

RC1 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 (Table 5). Unlike for cognition, RC2 also rejects *response consistency* by ethnicity (10 percent), as well as wealth. *Vignette equivalence* is rejected for age (10 percent) and education and the *p*-value lies only just above 10 percent for all the other factors except for employment status.

D. Quantitative impact of adjusting for reporting heterogeneity

The test results presented in the previous two subsections cast considerable doubt on the validity of the vignettes method of identifying reporting heterogeneity. It might be, however, that while the identifying assumptions are rejected, the method does a reasonable job of bringing subjective assessments closer to the truth, and correcting bias in estimated associations between the construct of interest and covariates. We now examine whether this is the case and, in so doing, illustrate the quantitative impact of adjusting for heterogeneity in the reporting of both cognition and mobility using both the vignettes and proxy indicators methods.

To the extent that vignette adjustments purge subjective assessments of differential reporting styles, they should bring those assessments closer to a reasonably comprehensive and objective measure of the construct of interest. The predicted latent index (Equation 1a) estimated from an ordered probit model provides a measure of cognition (mobility) that is potentially contaminated by reporting heterogeneity. To obtain a measure that is purged of reporting heterogeneity identified from vignette ratings, we reestimate Equation 1a using a model consisting of Equation 1a–1c with

the individual specific cut points in Equation 1c set equal to those obtained from the vignettes model (Equation 2c, Tables 4 and 5). The resulting estimates are used to predict vignette-adjusted latent health. We obtain an objective measure of cognition (mobility) from the latent index of Model 2 predicted on the basis of the proxy indicators. The correlation between the index that is not adjusted for reporting styles and the index predicted on the basis of the objective proxies is 0.508 for cognition and 0.421 for mobility. It is not at all encouraging that adjustment using vignettes actually reduces the correlation with the index derived from objective indicators to 0.459 for cognition and to 0.401 for mobility. Using data from the *Survey of Health Aging and Retirement in Europe,* Vonková and Hullegie (2011) find a similarly disappointing effect of vignette adjustment for cognition, but not for mobility. Van Soest et al. (2011) find that vignettes do help in bringing self-reports closer to an objective measure of drinking behavior, which may reflect improved performance of vignettes when applied to a more narrowly defined concept.

On top of the rejection of the identifying assumptions, the correlations suggest that, overall, vignettes do not do a good job of correcting for reporting heterogeneity. We now examine the direction and magnitude of the adjustment by covariate. Figure 2 presents partial effects on the probability of reporting at least some difficulty in cognition and in mobility unadjusted for reporting heterogeneity and adjusted using both vignettes and proxy indicators.

Partial effects unadjusted for reporting heterogeneity are obtained from ordered probit models. To obtain vignette-adjusted partial effects, we first estimate the parameters of Equation 1a as explained above and then predict latent health for each sociodemographic group for which partial effects will be computed, setting the remaining variables to their sample means. Finally, we predict the vignette-adjusted probability that a group has some difficulty in cognition (mobility) using the predicted latent health and the cut points of a reference individual (again using Equation 2c). Given the cut points are fixed, these probabilities vary with the impact of sociodemographics on latent health only. The same procedure is repeated with vignette-identified cut points replaced by those estimated from proxy indicators model to give an alternative set of probabilities purged of reporting heterogeneity. Finally, partial effects are computed by taking the difference in predicted probabilities between groups differentiated by age, gender, ethnicity, education, and employment. In the case of wealth, we compute the change in probability arising from movement from the 80th percentile of the distribution to the 60th, 40th, and 20th percentiles.

The unadjusted partial effects indicate that the probability of reporting at least some difficulty in cognition is greater for individuals who are older, male, from an ethnic minority, less wealthy, poorly educated and not working. The same is true for mobility, except that women are more likely to report a difficulty in this domain. In general, the vignettes adjustment decreases the partial effects. Exceptions are that it has no impact on the effect of wealth on mobility and it actually increases the effect of ethnicity in both domains as well as that of education on mobility. In most, but not all, cases, the direction of the vignette adjustment is consistent with that achieved using the proxy indicators. If one is willing to accept the assumption necessary for the latter to be valid (Equation A3), then this indicates that the vignette method mostly shifts the estimated effects in the right direction. But this is not always true. The two adjustments go in opposite directions for the effect of ethnicity

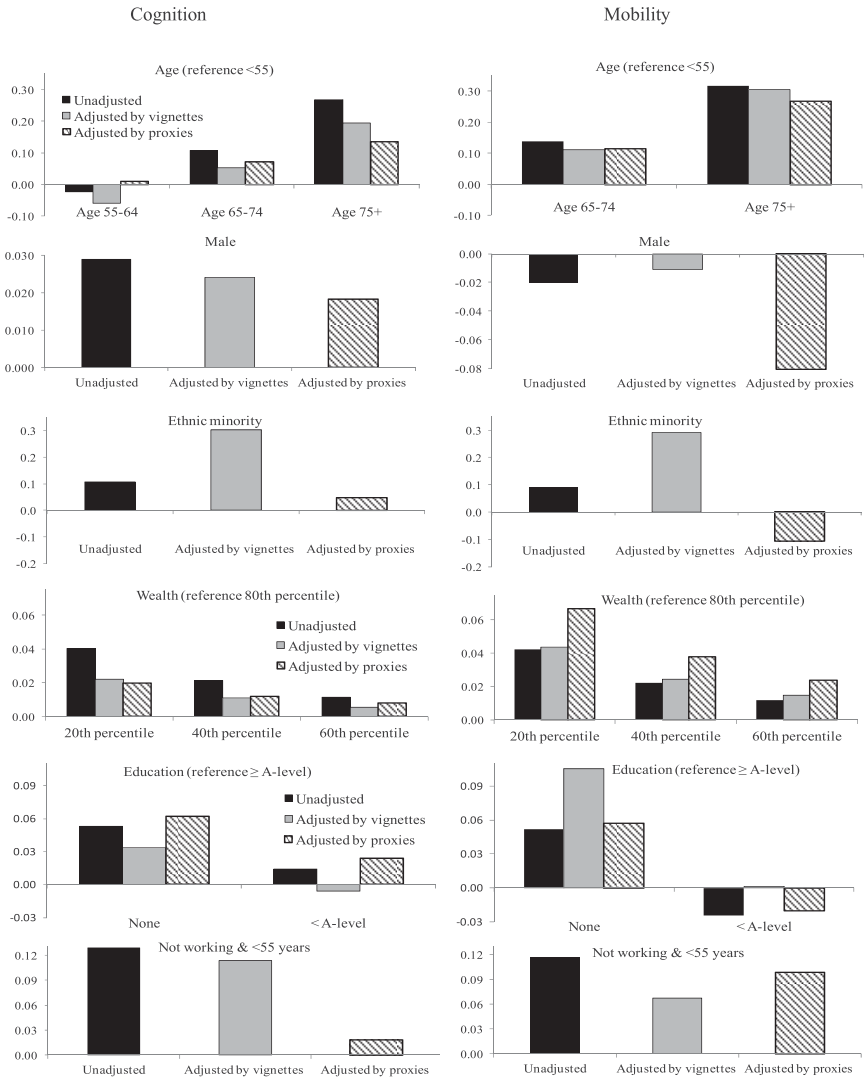


Figure 2
Partial Effects on Probability of Reporting Any Difficulty in Cognition and Mobility with and without Adjustment for Reporting Heterogeneity using Vignettes and Proxy Indicators

on both health domains. Using the vignettes suggests large disadvantages of ethnic minorities in both health domains, while this is not apparent when adjustment is made using the proxy indicators. Minorities are more likely to rate the health of a vignette positively but they are not more likely to rate their own health positively at given levels of objective indicators. Divergence in the directions of the adjustments also occurs for the effect of education on cognition and of gender on mobility.

With respect to the estimated effects of age, gender, wealth and, in particular, employment on cognition, the vignettes do not adjust by as much as is achieved using the objective indicators. For mobility, this is true only for age. Adjustment by proxy indicators increases the partial effect of wealth on the probability of having some difficulty in mobility, while the vignette adjustment has no impact. The vignettes do adjust the effects of education and employment on mobility by more than is achieved using the proxy indicators.

How sure can we be that differences in the magnitudes of the adjustments made by the two methods are attributable to violation of *response consistency* and so are indicative of mistakes of the vignettes approach? Is violation of Assumption A3—the available objective indicators are insufficient to absorb all association between covariates and true health—not an alternative explanation? We can partially answer these questions by considering the direction in which the proxy indicators adjustment is likely to be biased if Assumption A3 does not hold. For example, it is safe to assume that both cognitive functioning and mobility deteriorate with age. Both methods also suggest that the elderly understate cognitive functioning and mobility. In these circumstances, if Assumption A3 is not satisfied in the sense that the objective indicators cannot completely capture all age-related declines in health, then the magnitude of the reporting effect estimated using the proxy indicators will be biased upward. This could explain why proxy indicators make a larger adjustment to the age effect (at least for 75 +) than do vignettes. There is support for this in the fact that for cognition *response consistency* by age is rejected by RC1, which relies on Equation A3, but not by RC2, which does not require this assumption (Table 7). The same argument applies to the other three cases in which the proxy indicators adjustment shifts the effect in the same direction as the vignettes adjustment but to a greater extent—the effects of male, wealth, and not working on cognition. In two of these three cases, RC1 is rejected at 5 percent significance but RC2 is not. Following the same logic, the effect of no qualifications on mobility may be downward biased by the proxy indicators approach. Conditioning on more indicators that absorbed any residual covariance between true mobility and education would then bring the effect closer to that estimated using the vignettes adjustment.

Bias in estimated reporting effects due to insufficient conditioning on objective indicators cannot explain cases in which the two methods adjust in opposite directions. Males appear to have less difficulty with mobility but for given objective indicators they are more likely to report being constrained. If the available objective indicators do not sufficiently control for the mobility advantage to males, then the tendency for men to understate their mobility is underestimated. Better control would result in an even greater mobility advantage to males and a greater discrepancy from the effect estimated using the vignettes, according to which males slightly overstate mobility. Similarly, violation of Equation A3 cannot explain the different directions of the adjustments made to the effects of ethnicity on both health domains and to

that of education on cognition. Lower wealth is associated with an increased probability of experiencing a mobility problem, which, according to the proxy indicators approach, is partly obscured by a tendency of the less wealthy to report their mobility more positively. If the objective indicators do not sufficiently capture the greater mobility of the more wealthy, then the estimate of the reporting effect obtained is a lower bound and better control for mobility would result in an even greater discrepancy from the wealth effect estimated using the vignettes. Finally, the nonemployed are more likely to have a mobility problem but also more likely to report this. If the proxy indicators are insufficiently comprehensive, their application will give an upper bound on the estimated reporting effect and an over adjustment to the impact of work status on mobility. If there is such a bias, then correcting it would increase the inconsistency between the reporting thresholds (and so the partial effects) estimated using objective indicators and vignettes.

VI. 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 response to this challenge. The method relies on two identifying assumptions that hitherto have seldom been tested. We propose tests of both assumptions. Our test of *response consistency* requires data on objective indicators of the construct of interest that allow response scales to be identified and compared with those obtained from vignettes. Unlike Van Soest et al. (2011), we do not require that there exists a single objective measure that relates to individual sociodemographic 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 sociodemographics 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 method—*vignette equivalence*—that does not require data on objective indicators.

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. At least one null hypothesis is rejected for all factors but for age in the case of cognition and all but employment in the case of mobility.

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 and so bringing subjective assessments closer to the truth. Unfortunately, in this application, using vignettes to purge reporting heterogeneity does not increase the correlation between subjectively assessed cognition (mobility) and a measure derived

from objective indicators. Vignettes do adjust the estimated effects of age, wealth and employment on both health domains in the same direction as is achieved using proxy indicators to identify reporting scales. This is also true for the effects of gender on cognition and education on mobility. In the other cases examined, in particular for ethnicity, vignettes suggest reporting behavior that is contradictory to that revealed using proxy indicators.

We hypothesized that the vignettes approach would fail to fully correct for any tendency of the nonemployed to understate their health as a justification of their inactivity. The evidence on this is mixed. While vignettes do not succeed in reducing the association between inactivity and reported difficulties with cognition by as much as is achieved using objective indicators, the reverse is true for mobility.

While our results do cast serious doubt on the validity of the vignettes approach, they are obviously not sufficient to dismiss it. The proposed tests should be applied in other domains of health and to other subjective indicators that have been anchored on vignette evaluations. The opportunity for survey respondents to slip this anchor could be reduced by better implementation of the method. 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 concepts (such as health domains related to physical functioning) than it is for others (such as mental health problems and life/job satisfaction).

A potential way of increasing *response consistency* would be to switch the usual question order so that self-assessments follow the vignettes, thus priming respondents to define the response scale in a common way. Hopkins and King (2008) show that asking the vignette questions first significantly raises the likelihood of estimating expected relationships between sociodemographic variables and vignette-corrected political efficacy and economic class.

References

Bago d'Uva, Teresa, Eddy van Doorslaer, Maarten Lindeboom, and Owen O'Donnell. 2008. "Does Reporting Heterogeneity Bias Health Inequality Measurement?" *Health Economics* 17(3):351–75.

Bago d'Uva, Teresa, Owen O'Donnell, and Eddy van Doorslaer. 2008. "Differential Health Reporting by Education Level and Its Impact on the Measurement of Health Inequalities among Older Europeans." *International Journal of Epidemiology* 37(6):1375–83.

Banks, James, and Zoë Oldfield. 2007. "Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving." *Fiscal Studies* 28(2):143–70.

Banks, James, Elizabeth Breeze, Carli Lessof, and James Nazroo, 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, Hugo, Moshe Buschinski, Hiu Man Chan, Sofia Cheidvasser, and John Rust. 1999. "How Large Is the Bias in Self-Reported Disability?" *Journal of Applied Econometrics* 19(6):649–70

Bound, John. 1991. "Self Reported Versus Objective Measures of Health in Retirement Models." *Journal of Human Resources* 26(1):107–37

Christensen, Kaare, Anne Maria Herskind, James Vaulpel. 2006. "Why Danes Are Smug: Comparative Study of Life Satisfaction in the European Union." *British Medical Journal* 333: 1289–91.

Folstein, Marshall, Susan Folstein, and Paul McHugh. 1975. "'Mini-mental state': A Practical Method for Grading the Cognitive State of Patients for the Clinician." *Journal of Psychiatric Research* 12(3):189–98.

Gill, Thomas, Mayur Desai, Evelyne Gahbauer, Theodore Holford, and Christianna Williams. 2001. "Restricted Activity among Community-Living Older Persons: Incidence, Precipitants, and Health Care Utilization." *Annals of Internal Medicine* 135(5):313–21.

Guralnik, Jack, and Luigi Ferrucci. 2003. "Assessing the Building Blocks of Function: Utilizing Measures of Functional Limitation." *American Journal of Preventive Medicine* 25(3):112–21.

Guralnik, Jack, Eleanor Simonsick, Luigi Ferrucci, Robert Glynn, Lisa Berkman, Dan Blazer, Paul Scherr and Robert Wallace. 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* 49(2): M85–M94.

Hopkins, Daniel, and Gary King (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly* 74(2):201–22.

Huppert, Felicia, Carol Brayne, Caroline Gill, Eugene Paykel, and Lynn Beardsall. 1995. "CAMCOG—A Concise Neuropsychological Test to Assist Dementia Diagnosis: Socio-Demographic Determinants in an Elderly Population Sample." *British Journal of Clinical Psychology* 34(4):529–41.

Kapteyn, Arie, James Smith, and Arthur van Soest. 2007. "Vignettes and Self-Reports of Work Disability in the US and the Netherlands." *American Economic Review* 97(1):461–73.

Kerkhofs, Marcel, and Maarten Lindeboom. 1995. "Subjective Health Measures and State Dependent Reporting Errors." *Health Economics* 4(3):221–35

King, Gary, Christopher Murray, Joshua Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1):184–91

Kreider, Brent. 1999. "Latent Work Disability and Reporting Bias." *Journal of Human Resources* 34(4):734–69.

Kristensen, Nicolai, and Edvard Johansson. 2008. "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes." *Labor Economics* 15(1):96–117.

Lang, Iain, David Llewellyn, Kenneth Langa, Robert Wallace, and David Melzer. 2008. "Neighborhood Deprivation, Individual Socioeconomic Status, and Cognitive Function in Older People: Analyses from the English Longitudinal Study of Ageing." *Journal of the American Geriatrics Society* 56(2):191–98.

Llewellyn, David, Iain Lang, Jing Xie, Felicia Huppert, David Melzer and Kenneth Langa. 2008. "Framingham Stroke Risk Profile and Poor Cognitive Function: A Population-Based Study." BMC *Neurology* 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(2):319–35.

Murray, Christopher, Emre Özaltin, Ajay Tandon, Joshua Salomon, Ritu Sadana, and Somnath Chatterji. 2003. "Empirical Evaluation of the Anchoring Vignettes Approach in Health Surveys." In *Health Systems Performance Assessment: Debates, Methods and Empiricism*, ed. Christopher Murray and David Evans, 369–400. Geneva: World Health Organization.

Park, Denise. 1999. "Cognitive Aging, Processing Resources, and Self-Report." In *Cognition, Aging, and Self-reports*, ed. Norbert Schwarz, Denise Park, Bärbel Knauper, and Seymour Sudman, 45–70. 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.

Reed, Bruce, William Jagust, and J. Philip Seab. 1989. "Mental Status as a Predictor of Daily Function in Progressive Dementia." *Gerontologist* 29(6):804–807.

Richards, Marcus, Diana Kuh, Rebecca Hardy, and Michael Wadsworth. 1999. "Lifetime Cognitive Function and Timing of the Natural Menopause." *Neurology* 53:30814.

Rice, Nigel, Silvana Robone, and Peter Smith. 2010. "International Comparison of Public Sector Performance: The Use of Anchoring Vignettes to Adjust Self-Reported Data." *Evaluation* 16(1):81–101.

Rice, Nigel, Silvana Robone, and Peter Smith. 2011. "Analysis of the Validity of the Vignette Approach to Correct for Heterogeneity in Reporting Health System Responsiveness." *European Journal of Health Economics*. Forthcoming.

Stern, Steven. 1989. "Measuring the Effect of Disability on Labor Force Participation." *Journal of Human Resources* 24(3):361–95

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, Nicholas, Felicia Huppert, Brenda McWilliams and David Melzer. 2004. "Physical and Cognitive Function." In *Health, Wealth and Lifestyles of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, ed. Michael Marmot, James Banks, Richard Blundell, Carli Lessof, and James Nazroo, 249–71. London: Institute for Fiscal Studies.

Studenski, Stephanie, Subashan Perera, Dennis Wallace, Julie Chandler, Pamela Duncan, Earl Rooney, Michael Fox, and James Guralnik. 2003. "Physical Performance in the Clinical Setting." *Journal of the American Geriatric Society* 51(3):314–22.

Terza Joseph. 1985. "Ordinal Probit: A Generalization." *Communications in Statistics* 14(1):1–11.

Van Soest, Arthur, Liam Delaney, Colm Harmon, Arie Kapteyn, and James Smith. 2011. "Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjective questions." *Journal of the Royal Statistical Society Series A* 174(3):575–95.

Vonková, Hana, and Patrick Hullegie. 2011. "Is the Anchoring Vignettes Method Sensitive to the Domain and Choice of the Vignette?" *Journal of the Royal Statistical Society Series A* 174(3):597–620.

Wallace, Robert, and A. Regula Herzog. 1995. "Overview of the Health Measures in the Health and Retirement Study." *Journal of Human Resources* 30(5): S84–S107.

Weuve, Jennifer, Jae Hee Kang, JoAnn Manson, Monique Breteler, James Ware, and Francine Grodstein. 2004. "Physical Activity, Including Walking, and Cognitive Function in Older Women." *Journal of the American Medical Association* 292(12):1454–61.