Are anchoring vignettes ratings sensitive to vignette age and sex?

Hendrik Jürges
University of Wuppertal
Rainer-Gruenter-Str. 21 (FN.01)
42119 Wuppertal, Germany
juerges@uni-wuppertal.de

Joachim Winter
University of Munich
Ludwigstraße 28/ Rgb.
80539 München, Germany
winter@lmu.de

This version: September 2011

Abstract: Anchoring vignettes are commonly used to study and correct for differential item functioning and response bias in subjective survey questions. Self-assessed health status is a leading example. A crucial assumption of the vignette methodology is "vignette equivalence": The health status of the person described in the vignette must be perceived by all respondents in the same way. We use data from a survey experiment conducted with a sample of almost 5,000 older Americans to validate this assumption. We find weak evidence that respondents' vignette ratings may be sensitive to the sex, and for older respondents also to the age (implied by the first name) of the person described in the vignette. Our findings suggest that vignette equivalence may not hold, at least if the potentially subtle connotations of vignette persons' names are not fully controlled.

Keywords: anchoring vignettes; self-rated health; differential item functioning.

JEL-Codes: C81; J14; I19

Acknowledgements: This research was supported by the U.S. National Institute on Aging (NIA), grants P01 AG 05842-18 and R56 AG026622-01A1, with additional support from the E. Morris Cox Fund at the University of California, Berkeley. We thank two referees and the editor for helpful comments. We would also like to thank Daniel McFadden for supporting this project. Johanna Sophie Quis provided helpful research assistance.

1. Introduction

Subjective self-ratings reported by survey respondents are used frequently in the social sciences; examples of self-rated assessments are personal health, well-being, work ability, and job satisfaction. Unfortunately, self-ratings have been found to be subject to substantial reporting bias. The key problem is that subjective self-ratings involve respondents' evaluation of some domain of their own objective reality (such as their health) as well as their subjective thresholds for mapping their evaluation onto the response scale defined in the survey instrument (e.g., "excellent", "good", "fair", or "poor"). When these thresholds vary across respondents, their responses are not comparable any more – a phenomenon referred to as "differential item functioning" (see King et al., 2004; van Soest et al., 2011).

There is abundant evidence for differential item functioning in self-ratings reported by survey respondents. For example, when it comes to self-rated health, older respondents tend to have a milder view of their health, and there is also evidence for systematic heterogeneity in reporting styles across different countries (Sen, 2002; Lindeboom and Van Doorslaer, 2004; Jürges, 2007; Kapteyn et al., 2007). Methods based on "anchoring vignettes" are now commonly used to study and correct for differential item functioning in subjective survey responses, particularly in health domains (King et al., 2004; Salomon et al., 2004; Kapteyn et al. 2007, 2011; Bago d'Uva et al., 2008, 2011; Datta Gupta et al., 2010; Hopkins and King, 2010; Van Soest et al., 2011). In this paper, we evaluate a key assumption on response behavior required by this methodology.

Vignettes have been introduced in survey research and practice by Nosanchuck (1972) and Rossi et al. (1974). They are defined as "short descriptions of a person or a social situation which contain precise references to what are thought to be the most important factors in the decision-making or judgment-making process of respondents" (Alexander and Becker, 1978, p. 94). Statistical methods for adjusting self-rates for differential item functioning with the help of anchoring vignettes were developed by King et al. (2004). The key idea is to obtain not only a self-rating for some variable of interest, but also ratings for vignette persons whose descriptions keep the levels of that same variable fixed. The ratings for the vignette persons can then be used to adjust the self-rating, removing the effects of differential item functioning.

The methodology of adjusting responses using anchoring vignettes requires that two assumptions hold: response consistency and vignette equivalence. The purpose of this study is to test whether one of these – vignette equivalence – holds, using a randomized survey experiment in which we randomly assign different first names to otherwise identical vignette persons. According to King et al. (2004, p. 194), "vignette equivalence is the assumption that the level of the variable represented in any one vignette is perceived by all respondents *in the same way* and on the same unidimensional scale, apart from random measurement error" (our italics). Formally, this statement can be represented as

$$v_{ijk} = \alpha_{jk} + \varepsilon_{ijk}$$

where v_{ijk} is person *i*'s perceived level of vignette *j* in domain *k,* α_{jk} is the actual level of vignette *j* in domain *k,* and ε_{ijk} is person *i*'s random measurement error for this vignette. The key point is the relationship between α_{jk} and ε_{ijk}. Strictly speaking, a vignette description includes the first name used and its connotations, thus it is a part of α_{jk}. The vignette "Paul has problems walking about" can be viewed as different from the vignette "Richard has problems walking about." For practical reasons, however, first names and their connotations are treated as parts of ε_{ijk}, that is, they are not controlled for in standard regression analysis of vignette responses. This can result in some sort of omitted variables bias and vignette equivalence is violated.

Our survey experiment allows us to detect such violations. The questionnaire is based as closely as possible on the WHO's World Health Survey mobility vignettes (see http://surveydata.who.int/index.html) that are being used, inter alia, in such surveys as ELSA or SHARE. After an introduction text, worded similar to the introductions to the vignette questions in those surveys, we randomly assign to respondents a sequence of standard "mobility vignettes". The treatments differ only with respect to the vignette persons' first names which are randomized along two dimensions, the implied sex and age. The former is straightforward to manipulate by using female and male first names. To manipulate the vignette person's age, we use a novel approach which is based on the observation that the popularity of first names varies by birth cohort. Our premise is that some names are perceived as being typical for older people while others have the connotations of younger people (see Lieberson, 2000 for a review of the literature on how first names signal socio-demographic characteristics of a person, and Bertrand and Mullainathan, 2004, for evidence from a recent field experiment in labour economics). In fact, psychological research has demonstrated that names typical of older cohorts activate age-related stereotypes (English, 1916; Rudman et al., 1999).

This experiment allows us to test whether respondents consciously or unconsciously rate a vignette person's health not only according to the described health state (as they should for the vignette methodology to work) but also by the connotative age and sex of the vignette person. If the vignette first name "signals" that the person described is most likely an older adult, judgments may be affected, for instance because some limitation might be considered as the norm. Similarly, if the vignette person has a female name and her condition is considered typical for women, respondents might tend to rate her limitation as less severe than that of a man. Typical vignette studies thus either give only male vignettes to men and female vignettes to women or control for sex in the statistical analysis (e.g. Kapteyn et al. 2007, Van Soest et al., 2007). To preview the results, our data suggest that vignette ratings are indeed somewhat sensitive to subtle connotations of first names, at least among the subgroup of older respondents. This casts doubts on the universal validity of vignette equivalence. We discuss the results and implications in detail in the concluding section of this paper.

One may of course wonder whether our results are relevant given current practice in applied survey design. Are vignette names and their connotations sufficiently different across surveys to warrant a study like ours? We believe they are. We have looked at SHARE, the Survey of Health, Ageing and Retirement in Europe (Börsch-Supan et al. 2005), which had administered vignettes in more than 10 European countries. We compared actual vignettes questionnaires administered in countries as diverse as Sweden, Germany, and Spain. For instance, one item reads:

> "Alice has pain in her knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, she feels uncomfortable when moving around, holding and lifting things."

Alice is the name given in one of the two generic (English-language) versions of the questionnaire. It was fairly popular in the 1910s, 20s and 30s in the US (rank 13 to 22), but fell out of the top 200 girl's names in the 70s and 80s. In Sweden, Alice was "translated" as Mikael, which has been very popular in Sweden in the 1960s to 1980s. The German translation uses Beate, popular until the 1940s to 1960s and again in very recent times. In Spain, Alice turned Ana, a name that reached the peak of its popularity first in the 1930s and then later another peak in the 1980s.[1] How much these differences have affected actual response behaviour is impossible to say on the basis of the SHARE survey data alone. A survey experiment is needed to test whether vignette name connotations matter.

Our study adds to an ongoing debate about the validity of the assumptions that underlie the anchoring vignettes methodology (see Chevalier and Fielding, 2011, for a recent account). In

[1] Lists of popular first names can be found at `http://www.ine.es/tnombres/` for Spain, `http://www.scb.se` for Sweden, and `http://beliebte-vornamen.de` for Germany; websites accessed on 24 November 2010.

our assessment, the evidence is not yet conclusive. In a study that uses objective measures along with subjective self-ratings and vignettes, Van Soest et al. (2011) document that anchoring vignettes work well in adjusting differential item functioning. However, several recent studies also report apparent violations of vignette equivalence or response consistency. Datta Gupta et al. (2010) study vignettes for work disability and mobility in SHARE and find violations of response consistency using combined self-rated and objective measurements (grip strength and gait speed) of health. They also present a version of the statistical model for anchoring vignettes introduced by King et al. (2004) that relaxes this assumption. Bago d'Uva et al. (2011) analyze data from ELSA and reject both response consistency and vignette equivalence for cognition and mobility vignettes (specifically, for mobility vignettes similar to the ones we use in our study). Similar to Datta Gupta et al. (2010), their test of response consistency hinges on the assumption that objective health indicators, such as the cognitive tests and gait speed tests embedded in ELSA, are alternative (and superior) ways to assess response styles in subjective health ratings. In contrast, a joint test of both response consistency and vignette equivalence that does not rely on objective health indicators is proposed by Peracchi and Rossetti (2010). Their test exploits the fact that—if both assumptions hold—the statistical model used to analyze vignettes is overidentified. Peracchi and Rossetti find that the overidentifying restrictions are almost always rejected in data from SHARE.

Kapteyn et al. (2011) test response consistency using repeated surveys conducted with the same respondent. Their key idea is to show respondents in the second survey vignettes that are based on their own self-assessed health in the first survey. The results are mixed: Whereas for some health domains (sleep, mobility, affect), the correspondence between self-assessed health and vignettes that resemble the respondent's own health is quite good, the correspondence appears problematic for other health domains (concentration and breathing). These problems may be an indication that response consistency is violated but they may also indicate that objective health questions are insufficient to capture the respondent's health in the given domain. Also, there is no clear pattern with respect to the domains in which the vignettes work better. Despite these apparent violations of response consistency, Kapteyn et al. conclude that "vignette equivalence is a much more fragile assumption than response consistency" (p. 20). Consistently with the results we present in this paper, they argue that "for vignette equivalence to hold, a description has to be complete, minimizing room for different interpretations by different respondents." (ibid.).

The paper proceeds as follows. In sections 2 and 3, we describe our survey experiment and the sample we obtained, respectively. We report our main results in section 4 and provide a detailed subgroup analysis in section 5. Section 6 concludes.

2. The Experiment

We used a split-ballot design to randomly assign one of four versions of WHO mobility vignettes to the respondents. The 2x2 design manipulates sex (male/female) and age (young/old) of the vignettes' first names. Three vignettes were used:

Vignette 1: <Name> has no problems with walking, running or using her hands, arms and legs. She jogs 3 miles twice a week.

Vignette 2: <Name> is able to walk distances of up to 200 yards without any problems but feels tired after walking one half mile or climbing up more than one flight of stairs. She has no problems with day-to-day physical activities, such as carrying food from the market.

Vignette 3: <Name> has a lot of swelling in her legs due to her health condition. She has to make an effort to walk around her home, as her legs feel heavy.

After vignettes are presented to the respondents, they are asked to state "overall, how much of a problem does <Name> have with moving around?" with five answer categories ranging from "none" to "extreme".

First names were taken from the "Top 1000 names by decade", published by the US Social Security Administration (http://www.ssa.gov/OACT/babynames/) and based on a 5% sample of social security records. Names are limited to births in the United States. As "old" names we chose the three best ranked names in the 1920s that were *not* in the top 1000 in the 1980s. As "young" names we chose the three best ranked names in the 1980s that were *not* in the top 1000 in the 1920s. Thus young names were typical for 20 year olds but virtually unknown for 80 year olds, and vice versa.

Table 1 lists the first names used in our experiment together with their ranks in the 1920s and 1980s. Female names are higher ranked in their respective decades than male names. Thus changes in the popularity of names are larger for women than for men. Indeed, three of the four most chosen names for girls in the 1980s had not been in the top 1000 in the 1920s. The "strength" of the stereotypes evoked by our experiment might hence be stronger for female vignettes than for male vignettes.

3. Data

We use data from a survey experiment conducted with a sample of almost 5.000 Americans aged 50 and over. The experiment was conducted as part of the Retirement Perspectives Survey (RPS 2005) on November 7–15, 2005.[2] The RPS survey was primarily concerned with older Americans' health and their decisions related to the introduction of the new Medicare Prescription Drug benefit. Winter et al. (2006) and Heiss et al. (2009, 2011) provide detailed discussions of the survey contents, sample composition, and the substantive findings related to Medicare Part D. The survey lasted about 22 minutes, and covered, in addition to questions about Medicare Part D, questions about health status and conditions, long-term care choices, prescription drug use and cost, and attitudes toward risk. Embedded within the survey were also a series of experiments on various aspects of survey response behavior; the experiment reported in this paper was one of those experiments.

The data collection for RPS 2005 was in the form of a self-administered online survey for all respondents. RPS 2005 used a panel of subjects enrolled by Knowledge Networks (the "KN Panel"), a commercial survey firm. The KN Panel was recruited from a random sample of the US population, and was provided by the survey firm with computers or TV set-top boxes that are used to respond to periodic interviews. Panel members are therefore representative in terms of demographics and socioeconomic status. The full sample consists of 4,738 respondents.[3] The sample is described in Table 2. As can be expected in a sample of older adults, it contains more women than men (54 vs. 46 percent). 20 percent of the respondents in our sample are non-white, 12 percent are high-school drop-outs, 35 percent have finished high-school, and more than one quarter have a college degree. The majority of sample members are married, but the proportion of widowed women is substantial (around 20 percent). Overall, about 80 percent rate their general health as good or better.

[2] The survey was conducted by Knowledge Networks under the protocol for human subjects protection that covers their ongoing data collection. Ethical approval for data analysis has been given by Committee for Protection of Human Subjects (CPHS), UC Berkeley, April 2006.

[3] For the RPS 2005 survey, 5879 randomly selected KN Panel members aged 50 or older were contacted, so the unit response rate is 80.6%. According to personal communication from Knowledge Networks, this response rate is well above average for similar studies; one potential explanation is that the topic (the introduction of Medicare Part D) was of particular interest at the time of the fieldwork in late 2005. See Table 5.1 in Heiss et al. (2011) and the surrounding discussion for more detail on the response rates achieved in the four waves of the Retirement Perspectives Survey.

4. Results

The first column of Table 3 shows unconditional vignette ratings. The ordering of the vignettes in terms of severity of the underlying problem as perceived by the respondents corresponds to our expectations. Vignette 3 (swelling in legs) is perceived as having severe mobility problems by more than 70 percent of the respondents. Nearly 90 percent rate vignette 2's problem as mild or moderate, and more than 90 percent think that vignette 1 has no mobility problem at all.

Table 3 also shows vignette ratings by vignette age and sex. The working hypothesis of our experiment was that old names elicit age-related stereotypes, which in turn affect vignette ratings. In particular, we hypothesized that health problems of presumably older vignettes would be rated more often as "none" or "mild" and less often as "severe" and "extreme" than health problems of young vignettes. An alternative hypothesis would assume that an old vignette signals all sorts of health problems that are correlated with the problems described in the vignettes. In that case respondents' ratings would be biased toward more negative ratings. However, none of the two hypotheses is clearly supported. In fact, the distributions of ratings for each vignette show hardly any difference when comparing vignettes with young and old, and female and male, first names.

To test whether vignette age induces a unidirectional shift in latent vignette health, we ran simple ordered probit regressions of vignette ratings on a vignette age dummy variable. The shift parameter (beta) along with its standard error is reported below each young/old comparison. None of the parameters is significantly different from zero. For different vignettes, we even find different signs. Older vignettes tend to get a milder rating when using vignette 1 but less mild ratings when using vignettes 2 and 3.

In contrast to vignette age, vignette sex appears to have an effect on respondents' vignette ratings. Male vignettes get significantly less mild health ratings than female vignettes when using vignettes 2 and 3 (at the 10 and 5 percent levels, respectively) and milder ratings when using vignette 1 (but the latter difference is statistically insignificant).

Above we have noted that the young and old female names that we use in our experiment are ranked higher in their respective decades, so that the "strength" of the stereotypes evoked by our experiment might be stronger for female vignettes than for male vignettes. Put differently, female first names might have been easier recognized as old and young than male first names. Vignette age effects could thus also be stronger for female vignettes. Restricting the analysis to female vignettes, however, does not change the nature of our results. All ordered probit coefficients of vignette age dummies remain statistically insignificant. Another concern might be that men and women have responded to vignettes that contained first names of the opposite sex, something one would usually try to avoid if possible. When we restrict the sample to men who got male vignettes and women who got female vignettes, results do not change either.

In sum, it appears as if mobility vignette ratings are largely insensitive to the implied vignette age, but not to the vignette sex. The introductory remark given to a respondent that she should consider the vignette person as someone of her own "age and background" seems to work reasonably well. We now test whether this result holds in more detailed analyses.

5. Robustness checks

We extend our analysis in two directions. First, we check the robustness of our bivariate results by controlling for a number of respondent characteristics (age, sex, race, marital status, education). Second, we perform several subgroup analyses, particularly by sex, race, and age. To the extent that randomization has worked (which it did – we find no significant differences in respondent characteristics across treatment arms), controlling for respondent characteristics should not substantially change the point estimates for the ordered probit shift parameters reported in Table 3. Yet it could increase the precision of our estimate if controlling for these covariates takes away individual variation in vignette ratings. Moreover, it is a question in its own right whether vignette ratings vary systematically by observed characteristics, i.e., whether there is any systematic differential item functioning.

Table 4 shows ordered probit regression results of vignette ratings controlling for covariates. As expected, the ordered probit coefficients (and their significance levels) for old and male vignettes do not change much compared to the results in Table 3. However, some interesting patterns emerge with respect to covariates. First, vignette 1 (although having not much variation at all) appears to be most susceptible to differential item functioning. Ratings are significantly affected by all included variables: age, sex, race, marital status, and education. For instance, vignette ratings "decrease" with age, i.e., older respondents have a more *negative* view of the vignette person's mobility (positive coefficients reflect a more negative view). This finding is contrary to expectations because older respondents are usually believed to "overstate" their health relative to younger respondents. Although going in the same direction as for vignette 1, effects are much weaker for vignette 2. For vignette 3, the individual characteristics included in the regression have even weaker effects and they are jointly significant only at the 10 percent level. Individually, only education has a significant effect at the 5 percent level.[4]

To see whether differences across vignettes are significant, we tested cross equation restrictions separately for each variable (for detailed results see Table A1 in the Appendix). Differences in the effect of respondent characteristics across vignettes potentially indicate violations of the overidentifying restrictions imposed jointly by response consistency and vignette equivalence (cf. Peracchi and Rossetti, 2010). The null hypothesis of equal parameters in all three vignette models is clearly rejected for each respondent characteristic (but not for vignette characteristics). When we test only vignette 2 against vignette 3, differences in the parameters for age, race and marital status are no longer statistically significant (education parameter differences are significant at the 10 percent level only). The conclusion one might draw is that vignette 1 is in fact not ideal for our purpose and that results based on vignettes 2 and 3 only are more convincing. Still, in order to increase the power of our test for vignette name effects, we also pooled the data (and additionally controlled for vignette level). Since vignette 1 shows a fairly different behavior than the other vignettes, we also pooled only vignettes 2 and 3. However, neither pooled model shows a significant effect of vignette age on vignette ratings.

The results of our subgroup analyses are shown in Table 5. This table contains regression coefficients for “old vignette” and “male vignette” dummy variables, obtained from ordered probit regressions that control for all covariates listed in Table 4. We first split the sample by race. Our name generating algorithm resulted in typical “white” names; hence it is possible that non-white respondents are unable to distinguish between young and old names. Put differently, the discriminatory power of “old” and “young” might be larger in case of white respondents. There is some evidence supporting this view. In the white-only sample, old vignettes receive less mild health ratings than young vignettes. The effect now becomes significant at the 10 percent level for vignette 2 and for the pooled vignette 2 and 3 model. In contrast, non-whites consistently give milder health ratings to older vignettes. None of the coefficients is significantly different from zero. Splitting the sample by sex shows that the effect of vignette sex is largely due to men giving significantly different ratings for male and female vignettes.

[4] We also estimated all models in Table 4 using two alternative specifications: one that allows for vignette age- and sex-dependent thresholds and one allowing for thresholds that depend on all variables included in the model (generalized ordered probit models). We then tested whether coefficients were different across thresholds (results available on request). Overall, we found no support for the hypothesis that coefficients on vignette age and sex differ at different thresholds – except for vignette sex in vignette 1. But we anyway do not consider vignette 1 to be very useful due to its lack of variation.

We obtain the most striking results when we split the sample by age. Here, we have chosen to separate the oldest old respondents (aged 80 and over) from younger respondents, and we repeated the analyses for both subsamples including only white respondents. Among "young" respondents, the coefficient of the old vignette dummy is positive (and when the sample is restricted to whites significant) in all specifications except for vignette 1. This means young respondents have a less mild view of the health states of old vignettes than of young vignettes. This finding reverses when we consider the 80+ year old respondents. They have not only significantly but also substantially milder view of the health states of old vignettes than of young vignettes.

6. Discussion

To our knowledge, this is the first study to analyze the effect of vignette names and their connotations on vignette ratings by survey respondents. Our study was motivated by the fact that although much effort is being spent on the translation of vignettes in cross-national studies, less attention seems to be given to the choice of first names for the vignette descriptions. However, first names are highly cohort specific. For instance, we found in the Survey of Health, Ageing and Retirement in Europe (SHARE) that one and the same item was associated with first names popular in the beginning, middle, and end of the 20th century depending on the country where the survey was administered. Psychological research clearly shows that names that were common some 80 years ago elicit age stereotypes, which can potentially bias vignette ratings. If in an international survey, translators in country A choose names that are common for 40 year olds while translators in country B choose names common for 80 year olds, cross-national differences in vignette ratings might also be survey artefacts.

We have used a survey experiment to test for such effects by randomly assigning vignettes with old and young, and male and female, first names to nearly 5,000 respondents. Our data reveal significant effects of vignette sex on vignette ratings. This finding is of practical importance mainly in situations in which the survey instrument cannot be adjusted to respondent characteristics such as sex (usually, men would be given male vignettes, and women would be given female vignettes). With regard to vignette age, we find no significant effects when we analyze our full sample. But, to conclude that survey researchers planning to include vignettes (in cross-national research) need not be concerned about the choice of names for the vignette descriptions would be premature. In particular, our subgroup analysis shows that vignette age appears to lead the oldest old (80+) to slightly change their response behaviour.

To illustrate the consequences in an actual application, we have run conventional ordered probit and vignette-corrected regressions (so called "chopit" models; see King et al. 2004) of self-rated mobility on age, sex, marital status, and education, restricting the sample to white respondents only. The self-rating questions were "because of a health or memory problem do you have any difficulty with (1) walking several blocks, (2) running or jogging about a mile, (3) climbing several flights of stairs without resting?" The results are shown in Table A2 in the Appendix. The main finding here is that the difference between ordered probit and chopit estimates varies by the implied vignette age. In the sample of respondents who were randomly assigned "young" vignettes, the comparison of age dummy coefficient implies that conventional ordered probit estimates significantly *overstate* the age-mobility gradient. Put differently, older respondent's mobility self-ratings appear to be downward biased (or older people are "health-optimistic"). In contrast, in the sample of respondents who were randomly assigned old vignette persons, the comparison of uncorrected and corrected estimates implies that uncorrected estimates significantly *understate* the age-mobility gradient. In other words, older respondents' self-ratings are upward biased (or older people are "health-pessimistic"). That such a fundamental difference in results is due to the first names chosen for vignettes should raise concerns among vignette designers.

Our study has two important strengths: the large sample size, which allows detecting even small effects, and the randomized setup. Limitations are, first, that our respondents are members of an internet panel who participate in surveys regularly and who thus may be more experienced and more vigilant when answering surveys than the average first-time respondent in a random sample drawn from the general population. Our results might thus not extend to other surveys. Second, we were able to experiment only on a single health domain using three vignettes. Our results need not hold for other health domains or other mobility vignettes. Third, although we are confident that the selection of vignette names for our experiment induced age connotations are as intended, these connotations might have been too subtle. Unfortunately, we had no opportunity to let our respondents "guess" the age of the vignette persons, which would have allowed us to verify whether our treatment worked as intended. Moreover, it is possible that names are more (or less) cohort specific in other countries than the US.

Overall, however, the present study provides some indication that the choice of vignette first names is not entirely innocuous. An implication for survey practice is that cross-national studies using vignettes have to be designed with special care since vignettes studies in different countries typically employ country-specific names. If one country uses vignette first names that are more likely to describe young persons while another country uses names that are more likely to describe older persons, cross-national differences in vignette ratings might also reflect differences in the vignette age. Statistical adjustment of differential item functioning that requires the assumption of vignette equivalence would then fail.

References

Alexander, C. S. and Becker, H. J. (1978): The use of vignettes in survey research. *Public Opinion Quarterly*, 42, 93–104.

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

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., and Van Doorslaer, E. (2011): Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, forthcoming.

Betrand, M., Mullainathan, S. (2004): Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labour market discrimination. *American Economic Review*, 94, 991–1013.

Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J., Siegrist, J., and Weber, G., eds. (2005): *Health, Ageing and Retirement in Europe – First Results from the Survey of Health, Ageing and Retirement in Europe.* Mannheim: MEA.

Chevalier, A. and Fielding, A. (2011): An introduction to anchoring vignettes. *Journal of the Royal Statistical Society: Series A*, 174, 569–574.

Datta Gupta, N., Kristensen, N., and Pozzoli, D. (2010): External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, 27, 854–865.

English, G. (1916): On the psychological response to unknown proper names. *American Journal of Psychology*, 27, 430–434.

Heiss, F., McFadden, D., and Winter, J. (2009): Mind the gap! Consumer perceptions and choices of Medicare Part D prescription drug plans. In: D. A. Wise, ed., *Research Findings in the Economics of Aging*, 413–481. Chicago and London: Chicago University Press.

Heiss, F., McFadden, D., and Winter, J. (2011): The demand for Medicare Part D prescription drug coverage: Evidence from four waves of the Retirement Perspectives Survey. In: D. A. Wise, ed., *Explorations in the Economics of Aging*, 159–182. Chicago and London: Chicago University Press.

Hopkins, D. J. and King, G. (2010): Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly*, 74, 201–222.

Kapteyn, A., Smith, J. P., and Van Soest, A. (2007): Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97, 461–473.

Kapteyn, A., Smith, J. P., Vonkova, H., and Van Soest, A. (2011): *Anchoring vignettes and response consistency*. Working Paper No. WR-840, RAND, Santa Monica.

Lieberson, Stanley (2000): *A Matter of Taste: How Names, Fashions, and Culture Change.* New Haven, CT: Yale University Press.

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

Jürges, H. (2007): True health vs. response styles: Exploring cross-country differences in self-reported health, *Health Economics*, 16, 163–178.

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

Nosanchuck, T. A. (1972): The vignette as an experimental approach to the study of social status: An exploratory study. *Social Science Research*, 1, 107–120.

Peracchi F. and Rossetti C. (2010). The heterogeneous thresholds ordered response model: Identification and inference. Working Papers No. 1012, Einaudi Institute for Economics and Finance (EIEF).

Rudman, L., Greenwald, A., Mellott, D., and Schwartz, J. (1999): Measuring the automatic components of prejudice: flexibility and generality of the implicit association test. *Social Cognition*, 17, 437–465.

Rossi, P. H., Simpson, W. A., and Bose, C. E. (1974): Measuring household social standing. *Social Science Research*, 2, 169–199.

Salomon, J. A., Tandon, A., and Murray, J. P. (2004): Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes. *British Medical Journal*, 328, 258–261.

Sen A. (2002): Health: Perception versus observation. *British Medical Journal*, 324, 860–861.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and Smith, J. (2011): Validating the use of vignettes for correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A*, 174, 575–595.

Winter, J., Balza, R., Caro, F., Heiss, F., Jun, B., Matzkin, R. L., and McFadden, D. (2006): Medicare prescription drug coverage: Consumer information and preferences. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 7929–7934.

Table 1: Men's and women's first names used in the experiment

Men's names			
Old Name	Rank 1920s	Young Name[a)]	Rank 1980s
Homer	115	Ryan	14
Wilbur	121	Brandon	18
Orville	161	Jeremy	28
Women's names			
Old Name	Rank 1920s	Young Name	Rank 1980s
Lois	21	Jessica	1
Florence	22	Jennifer	2
Thelma	35	Ashley	4

Note: [a)] Brian (rank 16) dropped because of similarity to Ryan.

Table 2: Sample description (column percentages, by respondent characteristic)

	Men	Women	Total
Age group			
50-59	45.5	40.0	42.5
60-69	27.6	32.2	30.1
70-79	20.0	20.9	20.5
80+	7.0	6.9	7.0
Race			
White	80.8	80.5	80.7
Non-White	19.2	19.5	19.4
Education			
Less than high school	12.0	12.0	12.0
High school	33.2	37.4	35.4
Some college	27.2	26.2	26.6
Bachelor's degree or more	27.7	24.5	26.0
Marital status			
Married	70.8	53.2	61.3
Single	7.7	7.0	7.3
Divorced	16.1	19.4	17.9
Widowed	5.4	20.3	13.4
Self-rated general health			
Excellent	9.3	9.5	9.4
Very good	35.5	34.1	34.7
Good	34.8	34.9	34.9
Fair	15.2	17.1	16.2
Poor	5.3	4.5	4.9
N	2,187 (46.2%)	2,551 (53.8%)	4,738

Table 3: Vignette ratings by vignette sex and age (column percentages, by vignette)

		vignette age		vignette sex	
	total	young	old	female	male
	(N=4,703)	(N=2,340)	(N=2,263)	(N=2,377)	(N=2,326)
Vignette 1 (No problems)					
None	92.4	92.0	92.7	92.1	92.7
Mild	3.5	3.3	3.6	3.1	3.8
Moderate	2.3	2.8	1.9	2.7	2.0
Severe	1.0	1.1	1.0	1.1	1.0
Extreme	0.8	0.9	0.7	1.0	0.6
beta		-0.056 (0.053)		-0.066 (0.053)	
Vignette 2 (Tired after walking)					
None	5.0	5.2	4.8	4.8	5.2
Mild	36.7	37.0	36.4	38.1	35.2
Moderate	52.5	52.1	52.9	51.8	53.3
Severe	5.4	5.2	5.6	4.8	5.9
Extreme	0.5	0.5	0.4	0.5	0.4
beta		0.028 (0.032)		0.054 (0.032)+	
Vignette 3 (Swelling in legs)					
None	1.3	1.5	1.1	1.5	1.0
Mild	1.0	0.9	1.2	1.0	1.1
Moderate	8.3	8.8	7.8	9.1	7.5
Severe	71.8	71.3	72.3	71.5	72.2
Extreme	17.6	17.5	17.7	17.0	18.1
beta		0.032 (0.034)		0.068 (0.034)*	

Note: “beta” is the coefficient estimate from an ordered probit regression as explained in the text; standard errors are in parentheses. + significant at 10%; * significant at 5%; ** significant at 1%

Table 4: Ordered probit regression for vignette ratings

	Vignette 1	Vignette 2	Vignette 3	Pooled (1,2,3)	Pooled (2,3)
Vignette characteristics					
Old Vignette	-0.050	0.027	0.030	0.008	0.029
	(0.055)	(0.032)	(0.034)	(0.023)	(0.027)
Male Vignette	-0.066	0.052	0.070*	0.023	0.060*
	(0.055)	(0.032)	(0.035)	(0.023)	(0.027)
Respondent characteristics					
Age 60 to 69	0.004	0.048	-0.002	0.016	0.022
	(0.072)	(0.039)	(0.042)	(0.027)	(0.033)
Age 70 to 79	0.322**	0.022	0.060	0.088**	0.035
	(0.075)	(0.046)	(0.049)	(0.032)	(0.037)
Age 80+	0.770**	0.134+	0.009	0.259**	0.071
	(0.100)	(0.071)	(0.076)	(0.060)	(0.061)
Female	-0.116*	-0.129**	0.020	-0.056*	-0.048+
	(0.058)	(0.034)	(0.036)	(0.023)	(0.028)
White	-0.357**	0.077+	0.052	-0.028	0.074*
	(0.067)	(0.042)	(0.045)	(0.032)	(0.037)
Never married	0.211+	0.160*	0.035	0.108*	0.100+
	(0.110)	(0.064)	(0.068)	(0.047)	(0.055)
Divorced	0.311**	-0.029	-0.044	0.034	-0.037
	(0.072)	(0.044)	(0.047)	(0.031)	(0.037)
Widowed	0.041	-0.020	0.056	0.016	0.025
	(0.086)	(0.054)	(0.058)	(0.037)	(0.042)
Less than high school	0.840**	-0.018	-0.076	0.123**	-0.071
	(0.097)	(0.057)	(0.060)	(0.043)	(0.050)
High school	0.499**	0.007	0.083+	0.101**	0.033
	(0.086)	(0.042)	(0.045)	(0.027)	(0.034)
Some college	0.436**	0.046	0.004	0.083**	0.016
	(0.089)	(0.045)	(0.048)	(0.030)	(0.037)
Vignette 2				2.389**	
				(0.048)	
Vignette 3				4.577**	2.425**
				(0.088)	(0.044)
N	4703	4700	4701	14104	9401

Note: Standard errors in parentheses; Standard errors in pooled models are clustered on individual level; + significant at 10%; * significant at 5%; ** significant at 1%; positive coefficients indicate *worse* vignette health ratings

Table 5: Ordered probit regressions for vignette ratings; subgroup analyses

	Vignette 1	Vignette 2	Vignette 3	Pooled	Pooled
				(1,2,3)	(2,3)
Whites					
Old Vignette	-0.044	0.060+	0.042	0.030	0.053+
	(0.064)	(0.036)	(0.039)	(0.025)	(0.030)
Male Vignette	-0.095	0.045	0.086*	0.025	0.065*
	(0.064)	(0.036)	(0.039)	(0.025)	(0.030)
N	3795	3789	3790	11374	7579
Non-whites					
Old Vignette	-0.065	-0.114	-0.022	-0.062	-0.068
	(0.112)	(0.073)	(0.078)	(0.050)	(0.060)
Male Vignette	0.017	0.094	0.009	0.024	0.046
	(0.112)	(0.073)	(0.078)	(0.050)	(0.059)
N	908	911	911	2730	1822
Men					
Old Vignette	-0.142+	0.018	0.044	-0.004	0.033
	(0.080)	(0.048)	(0.051)	(0.034)	(0.040)
Male Vignette	0.022	0.105*	0.068	0.065+	0.083*
	(0.080)	(0.048)	(0.051)	(0.034)	(0.040)
N	2168	2170	2170	6508	4340
Women					
Old Vignette	0.041	0.031	0.017	0.017	0.022
	(0.078)	(0.044)	(0.047)	(0.031)	(0.036)
Male Vignette	-0.147+	-0.000	0.068	-0.018	0.033
	(0.078)	(0.044)	(0.047)	(0.030)	(0.036)
N	2535	2530	2531	7596	5061
Aged 50-79					
Old Vignette	-0.041	0.040	0.056	0.026	0.047+
	0.059)	(0.034)	(0.036)	(0.023)	(0.028)
Male Vignette	-0.086	0.060+	0.072*	0.025	0.066*
	0.059)	(0.034)	(0.036)	(0.024)	(0.028)
N	4380	4372	4374	13126	8746
Aged 80+					
Old Vignette	-0.057	-0.116	-0.300*	-0.159+	-0.190+
	0.162)	(0.124)	(0.134)	(0.082)	(0.098)
Male Vignette	-0.002	-0.096	0.039	-0.027	-0.040
	0.161)	(0.124)	(0.133)	(0.084)	(0.098)
N	323	328	327	978	655
Aged 50-79 (only Whites)					
Old Vignette	-0.040	0.072+	0.070+	0.049+	0.073*
	0.070)	(0.038)	(0.040)	(0.027)	(0.032)
Male Vignette	-0.115	0.058	0.089*	0.031	0.074*
	0.070)	(0.038)	(0.040)	(0.027)	(0.032)
N	3502	3492	3494	10488	6986
Aged 80+ (only Whites)					
Old Vignette	-0.008	-0.078	-0.293*	-0.139	-0.172+
	0.171)	(0.131)	(0.142)	(0.087)	(0.103)
Male Vignette	-0.035	-0.117	0.060	-0.040	-0.038
	0.170)	(0.131)	(0.141)	(0.088)	(0.106)
N	293	297	296	886	593

Note: Standard errors in parentheses; Standard errors in pooled models are clustered on individual level;
+ significant at 10%; * significant at 5%; ** significant at 1%, Control variables: see Table 4.
Positive coefficients indicate *worse* vignette health ratings

Table A1: Cross-equation tests of ordered probit coefficients (Chi-squared test statistics)

	Vignette 1=Vignette 2=Vignette 3			Vignette 2=Vignette 3		
	Chi-squared	df	p-value	Chi-squared	df	p-value
Vignette characteristics						
Old Vignette	1.68	2	0.431	0.01	1	0.940
Male Vignette	4.25	2	0.119	0.23	1	0.630
Respondent characteristics						
Age	53.82	6	0.000	4.68	3	0.197
Sex	14.72	2	0.001	14.70	1	0.000
Race	38.50	2	0.000	0.23	1	0.630
Marital status	25.17	6	0.000	4.97	3	0.174
Education	76.87	6	0.000	7.18	3	0.066

Table A2: Comparison of age-gradients in mobility problems estimated by ordered probit and by vignette corrected

		Young vignettes			Old vignettes	
	Ordered Probit	Chopit	p-value difference	Ordered Probit	Chopit	p-value difference
Walking several blocks						
Age 60-69	0.039 (0.069)	0.028 (0.071)	0.539	0.163** (0.069)	0.153* (0.071)	0.602
Age 70-79	0.353** (0.078)	0.316** (0.080)	0.081	0.430** (0.077)	0.440** (0.079)	0.625
Age 80+	0.736** (0.108)	0.669** (0.112)	0.020	0.654** (0.110)	0.696** (0.116)	0.182
N	1886			1921		
Running or jogging about a mile						
Age 60-69	0.292** (0.061)	0.295** (0.063)	0.837	0.448** (0.060)	0.428** (0.062)	0.212
Age 70-79	0.658** (0.072)	0.621** (0.074)	0.040	0.848** (0.071)	0.847** (0.073)	0.949
Age 80+	1.122** (0.121)	1.036** (0.126)	0.013	0.945** (0.120)	1.033** (0.126)	0.002
N	1885			1919		
Climbing several flights of stairs without resting						
Age 60-69	0.178** (0.060)	0.166** (0.064)	0.594	0.270** (0.061)	0.249** (0.064)	0.327
Age 70-79	0.531** (0.072)	0.504** (0.075)	0.228	0.476** (0.069)	0.473** (0.072)	0.892
Age 80+	0.883** (0.106)	0.816** (0.111)	0.041	0.742** (0.110)	0.810** (0.116)	0.027
N	1887			1928	.	

Note: Standard errors in parentheses; + significant at 10%; * significant at 5%; ** significant at 1%, Control variables: see Table 4. Vignettes 2 and 3 used in Chopit model. Positive coefficients indicate *worse* health.