Anchoring Vignettes for Health Comparisons: The Validity of a Multidimensional IRT Model Approach and Design Improvements Using Visual Vignettes

by

Mengyao Hu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2018

Doctoral Committee:

Associate Research Scientist Sunghee Lee, Co-Chair
Professor Jacqui Smith, Co-Chair
Professor Arie Kapteyn, University of Southern California
Assistant Research Scientist Zeina Mneimneh
Research Assistant Professor Hongwei Xu

Mengyao Hu

maggiehu@umich.edu

ORCID iD: 0000-0002-9044-9009

DEDICATION

To my parents who have made who I am today, giving me all support and love.

To my parents-in-law who have been a great source of inspiration and support.

To my husband, Xiaotian for his unconditional love and support.

To my children, Aria and Eugene, for filling my days with joy and happiness.

To my brother, Qian, for his continued support and encouragement.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors, Sunghee Lee and Jacqui Smith. My PhD has been an amazing experience and I thank both Sunghee and Jacqui from the bottom of my heart. They have not only provided me tremendous academic support, but also constantly inspired me with their intelligence, motivation, and kindness.

I am also hugely appreciative to Arie Kapteyn, Zeina Mneimneh and Hongwei Xu for serving on my dissertation committee and providing me invaluable suggestions and feedbacks during the various stages of my PhD research.

Special mention goes to Beth-Ellen Pennell, the Director of the International Survey Operations Division of the Survey Research Center of the Institute for Social Research (ISR), who provided me many great learning opportunities at various cross-cultural survey conferences and providing me constant support throughout my PhD process. To Gina-Qian Cheung, for always encouraging me and providing me various conferences and networking opportunities. Many thanks also to those I have worked with at SRO for always being there for me: my heartfelt thanks to Nancy Bylica, Jennifer Kelley, Yu-chieh (Jay) Lin, Julie de Jong, Lisa Holland, Daniel Tomlin, and Jamal Ali.

I am grateful to Brady West and Ting Yan, for guiding me through the process of working on research projects and publishing papers. I truly learned a lot by observing how great researchers like you work and think and by collaborating with you. Also to Jim Lepkowski, for encouraging me when I am a master student at the School of Public Health many years ago and

introducing me into my beloved field of survey methodology. To Carrie Karvonen-Gutierrez, for providing me invaluable suggestions on the design of health-related visual vignettes.

Many thanks to the other faculty and staff at MPSM and to my fellow students, for their support, feedback, and friendship. Thanks in particular to Tuba Suzer Gurtekin, Kristen Cibelli, Chris Antoun, Chan Zhang, Raphael Nishimura, Yanna Yan, Brian Wells, Felicitas Mittereder, Micha Fischer, Colleen McClain, Ai Rene Ong, Jingwei Hu, Fan Guo, Shu Duan, Ben Duffey, Yuan Zhang and Yichen Wang. Special mentions to Edmundo Roberto Melipillán, for encouraging me at every stage of my PhD process and for discussing with me on different statistical methods and collaborating with me on various projects. To Sharan Sharma, for always encouraging me to think positively when I am confused and lost myself. To Mingnan Liu, for providing me invaluable suggestions through my PhD studying.

Special mention goes to the members of the Psychological Aging Lab, who over the past three years, have become a constant source of inspiration and encouragement. Many thanks to Lindsay Ryan, Marina Larkina, Jasmine Manalel, Elise Hernandez, Hannah Giasson, Shannon Mejía, Jennifer Sun, Liz Morris, Aneesa Buageila and Haena Lee. I cannot imagine getting here without your support.

Also many thanks to Chris Feak from the English Language Institute at the University of Michigan for the support over the past several years. Chris has been not only a great teacher but also a role model and friend to me. Thank you for always putting students at the first place and for all your suggestions and encouragements over the years.

This research would not have been possible without the generous support of the Rackham School of Graduate Studies at the University of Michigan, the Rensis Likert Fund, the CEW

Riecker Graduate Student Research Grant and the Michigan Program in Survey Methodology (MPSM).

Last but not least, I would like to thank my family for their unconditional love and support. My parents tried their best to give me the best education, and taught me how to love. Now since I have my own children, I can more deeply understand the great love of them. The sacrifices they have made and continue to make every day made it possible for me to be here today. I would also like to thank my parents-in-law for providing me constant support and love over the years. My husband has been very supportive and has always been my best friend. Thank you for always trying your best to be a good husband and a good father of our children. Thank you for your trust and confidence in me. I love you. During my 4.5 years of my PhD process, the happiest moments are the births of my children. Their love and trust in me have made me a more powerful person. Their births helped me to better understand the meaning of life, love and responsibility. Thank you both for filling my life with joy and happiness. I love you. Also many thanks to my beloved brother who always trusts me and has confidence in me. I am a better person because of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

With the increasing popularity of cross-cultural research, researchers are facing a difficult problem: respondents with different backgrounds often use different standards when answering survey questions with ordinal response scales, resulting in incomparability of responses across groups (or Differential Item Functioning, DIF). Among the many techniques to ameliorate the problem, anchoring vignettes – a set of questions that describe hypothetical individuals' situations related to the measure of interest – become an increasingly popular tool for correcting for DIF in interpersonal and cross-cultural comparisons. The successful use of anchoring vignettes depends on two measurement assumptions: 1) response consistency (RC), which means that respondents rate vignette persons in the same way as they rate themselves; and vignette equivalence (VE), which means that the situations posed in vignettes are perceived similarly across respondents. Despite its widespread use, there are several practical challenges, and the fulfillments of the VE and RC assumptions are not always assured. This dissertation intends to fill three important gaps in the existing literature related to anchoring vignettes: specifically, I examine i) a potential statistical solution to evaluate and correct for DIF; ii) the utility of using visual versus verbal vignettes with respect to their efficiency and assumption fulfilments; and iii) question order effects as a source of response error.

The first study focuses on using anchoring vignettes to correct for DIF caused by differential response styles – respondents' systematic tendency to use the response scales in a certain way regardless of the question content. Recently, a multidimensional IRT model using anchoring vignette items is proposed by Bolt et al. (2014), which allows controlling for all types

of response styles. This present study uses several objective benchmarks to evaluate the validity of this IRT model approach by comparing the model with vignettes (to adjust response styles) and the model without vignettes. Data from the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) is used for analysis. Findings from this study indicate that the use of anchoring vignettes in this multidimensional IRT model does not effectively control for response styles in the SHARE data. This may be due to the violations of vignette measurement assumptions. These findings highlight the importance of constructing good vignettes that can meet the assumptions.

The second study examines the use of visual vignettes with images as an alternative design to current verbal vignettes for four health domains – pain, mobility, sleep and affect. To compare the performances of verbal and visual vignettes, this study conducts a web survey experiment, collecting data from various racial / ethnic groups. Their performances are compared using various approaches, including survey time and tests of assumptions. Despite the violations of VE found in both visual and verbal vignettes, the results show that, compared to verbal vignettes, visual vignettes can greatly reduce survey time without the loss of the DIF-adjusting quality. This study shows the great potential of visual vignettes in improving efficiency in the survey designs.

The third study evaluates the survey context effects (specifically question order of self-assessment question relative to vignettes) on self-assessed health. The results show that the question order effects could differ among racial / ethnic groups, suggesting that researchers and practitioners need to be cautious of the context effects when using the vignette first, self-assessment last order, especially in cross-cultural studies.

CHAPTER I

Overview

Self-assessment questions are routinely used in surveys for a wide range of topics, including health, disability, well-being, and life satisfaction. These questions often use Likert-type rating scales to measure respondents' attitudes, knowledge, perceptions, and behavior (Lee et al. 2002; Krosnick & Abelson 1992). Ideally, responses obtained from these questions reflect only the respondents' actual attitudes / perceptions. This, however, is not always the case. In fact, answers to self-assessment questions not only reflect respondents' true state, but also how they use the scales which could be determined by their cultural background. Such differences in the use of the scale can lead to response-category differential item functioning (DIF) (King et al. 2004; King & Wand 2007). As described in King and Wand (2007), DIF caused by differential response scale usage refers to the situation that when respondents map their state / perception onto the response scales, those from different backgrounds may understand and use the scales in different ways.

Figure 1.1 illustrates an example of DIF in a cross-cultural study. For a survey question with an ordinal scale and response categories ranging from "None" to "Extreme," indicating, for example, pain level, the cutpoints for the response categories may differ by cultural groups. Assume that 1) the study compares two cultures, A and B, 2) these two cultures perceive and/or use response scales differently, 3) there are two respondents, each from Culture A and B, and 4) these two respondents' true pain level falls on the vertical dashed line. Despite having identical

pain level, the respondent from Culture A will select mild and the respondent from country B will choose moderate. If this DIF is not accounted for, simple between-country mean comparisons will erroneously conclude that the Country B respondent experiences a higher level of pain.

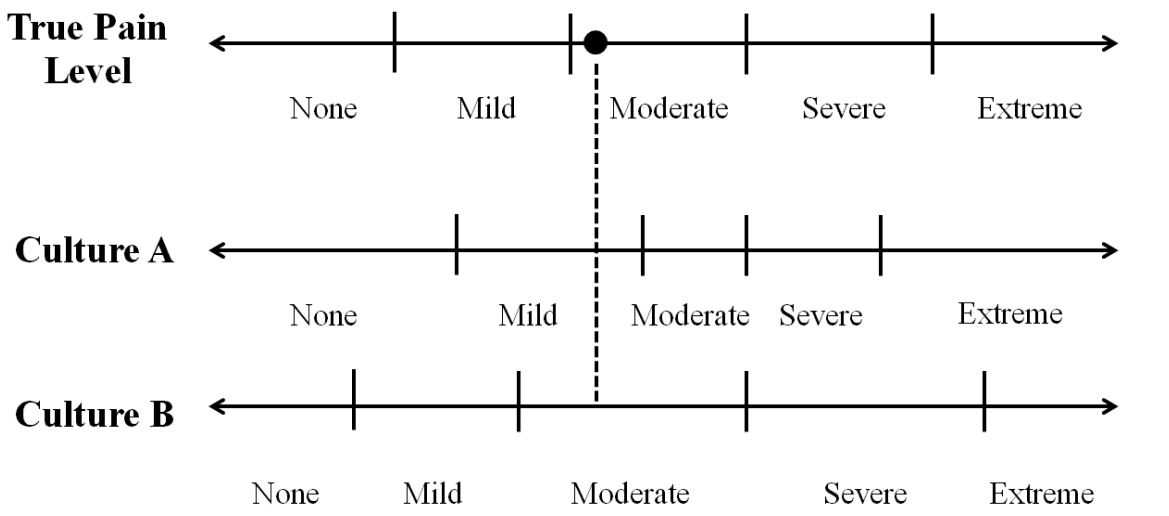


Figure 1. 1 Differential item functioning (DIF) for Cross-national Studies. The horizontal line with arrows indicate the continuum scales of the domain (pain level). The short vertical lines indicate the cutpoints respondents use to answer the self-assessment question. The vertical dashed line indicates respondents responses to self-assessment questions. Those whose pain level fall on that line indicates they have the same true pain level.

As can be seen in Figure 1.1, DIF can be particularly problematic when survey data are used in cross-cultural comparisons. Numerous studies have demonstrated the existence of DIF in cross-cultural surveys (e.g., Hirve, 2014; Lindeboom & Van Doorslaer, 2004). For example, Chinese respondents report higher political efficacy than Mexican respondents in their self-assessments, although all other indicators show the opposite (King et al., 2004). Respondents from different cultures are found to have systematically different understandings for each category in a widely-used "excellent" to "poor" response scale (Hsee & Tang 2007; Mojtabai 2015). For example, Jürges (2007) argued that although "excellent" is widely used in some countries like Sweden and Denmark, the word is used to express "ironic exaggeration" in German culture (Jürges 2007), which raises questions regarding the validity of cross-culture comparisons using this scale.

One innovative analytical approach to address DIF issues in cross-cultural comparisons is the use of anchoring vignettes. Anchoring vignettes are comprised of a set of questions, each describing in a few sentences a hypothetical person's situation related to the construct measured (e.g., pain). For example, a vignette used in the Health and Retirement Study (HRS) says "Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs". Respondents are then asked to rate the vignette person's pain level the same as they would for themselves. The vignette question can then serve as a benchmark for the actual unobserved pain level that researchers intend to measure.

Vignette questions are designed in a way to fulfill two key measurement assumptions. One is response consistency (RC), which means that respondents rate vignette persons in the same way as they rate themselves. The other one is vignette equivalence (VE), which means that

the situations posed in vignettes are perceived similarly across respondents. By assuming RC and VE, this method allows researchers to correct for differential response scale usage in cross-cultural or interpersonal comparisons (King et al. 2004; King & Wand 2007).

Despite the widespread use of anchoring vignettes in comparative studies (e.g., Bolt, Lu, & Kim, 2014; Paccagnella, 2014; Salomon, Tandon, & Murray, 2004), implementation of these vignettes introduces several limitations and raises questions regarding the validity of this method and how well they meet the assumptions. The research proposed here extends current work on the use of anchoring vignettes to mitigate DIF in cross-cultural research. Specifically, this dissertation has three research objectives:

- Objective 1: Evaluate how well a multidimensional IRT model approach using anchoring vignettes can statistically correct for both respondent and country-level response styles.
- Objective 2: Evaluate the use of visual vignettes as an alternative design to current verbal vignette designs.
- Objective 3: Evaluate the role of survey context, specifically I examine if the question order of self-assessment question relative to vignettes influences survey responses.

Each research objective corresponds to each of the three substantive chapters in this dissertation. Chapter 2 of the dissertation focuses on the use of anchoring vignettes to address response styles. Response styles, as a source of DIF, have been widely evaluated in the literature. Several well-known response styles are extreme response style (ERS, the tendency to select the two extreme endpoints of a scale), midpoint response style (MRS, the consistent selection of

middle or neutral category of the scale) and acquiescent response style (ARS, the tendency to agree with or to select the positive responses) (Harzing et al., 2012; He & van de Vijver, 2013; Vaerenbergh & Thomas, 2013). It is known that response styles can distort the reliability and validity of self-assessment measures. There is no standard approach to evaluate and control for response styles, and few methods can control for multiple response styles simultaneously. Recently, a multidimensional IRT model using anchoring vignette items is proposed, which allows controlling for all types of response styles. However, the literature is silent about its validity. This chapter aims to evaluate the concurrent validity of this approach, using several benchmarks constructed through objective measures.

Chapters 3 and 4 broadly focus on potential ways to improve the design of anchoring vignettes to better correct for DIF. Chapter 3 evaluates an alternative design option to current vignette designs, namely the use of visual vignettes. Despite the wide use of the verbal vignettes in comparative studies, research show that several critical challenges are associated with this method. Further, the fulfillment of the VE and RC assumptions behind this method is not always assured. To remedy the limitations of verbal vignettes, this chapter proposes the use of visual vignettes, consisting of carefully designed and pre-tested images. In this chapter, the performance of both verbal and visual anchoring vignettes are compared using various approaches, including survey time and tests of both VE and RC assumptions.

Chapter 4 investigates an overlooked source of measurement error related with both verbal and visual anchoring vignette designs, specifically, question order effects. This study is important for several reasons. First, research on the question order effects of anchoring vignettes have provided mixed recommendations. There is no clear guideline on the placement of self-

assessment relative to vignettes. This research can help researchers to make better question-order decisions. Second, prior studies have ignored the influence of potential question order effects for different racial / ethnic groups. Third, very few studies have evaluated question order effects in the use of health-domain vignettes, which are the most widely applied vignette questions worldwide. Finally, it remains unknown whether the same findings found in verbal vignettes can be applied to visual vignettes. This research aims to fill the research gap and evaluate the question order effects related to the placement of self-assessment for both verbal and visual vignettes on four health domains using a sample that includes multiple racial / ethnic groups.

Note that we treat Chapters 3 and 4 as two studies conceptually, but they are conducted in the same data collection period. Both studies are based on a web survey about vignettes in their use on four health domains, sleep, affect, mobility and pain. The sample of web survey includes respondents from four racial/ethnic groups[1]: non-Hispanic (NH) white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic[2].

By presenting analyses from a nationally representative cross-cultural survey and conducting an experiment comparing verbal and visual vignettes, this research advances previous work on vignette methodology, and provides the survey community with some of the first methodological examinations of this novel visual vignette method, along with insights on future cross-cultural survey questionnaire designs.

The contribution of this dissertation is twofold. First, from methodological perspective, the results of this research will create scientific knowledge regarding the impact of different

[1] Here, the four racial/ethnic groups are proxies for different cultural groups.
[2] Given that language can reflect and elicit cultural identification (Schecter & Bayley 2005), we include both English-speaking Hispanic and Spanish-speaking Hispanic in Chapters 3 and 4.

anchoring vignette designs on DIF correction and ultimately better comparability across groups. The study will also shed light on future design of visual vignettes. Second, from a practical perspective, products of this study include a series of validated visual vignettes for these domains, which will be an important contribution to not only survey methodology but also epidemiology, public health and clinical research, allowing researchers to better address racial/ethnic health disparities with a higher level of validity. The finding of this research can also be adapted to other fields, with the general implications for improving the design of anchoring vignette methods, which can help reducing survey time with improvements of data quality. The results for the four health domains will help practitioners from various fields to make better instrument design decisions, also allowing them to better address racial/ethnic health disparities with a higher level of validity.

References

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528.

Harzing, A. W., Brown, M., Köster, K., & Zhao, S. (2012). Response style differences in cross-national research. *Management International Review*, *52*(3), 341-363.

He, J., & Van de Vijver, F. J. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, *55*(7), 794-800.

Hirve, S. (2014). 'In general, how do you feel today?'self-rated health in the context of aging in India. *Global health action*, *7*(1), 23421.

Hsee, C. K., & Tang, J. N. (2007). Sun and water: on a modulus-based measurement of happiness. *Emotion*, *7*(1), 213.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health economics*, *16*(2), 163-178.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, *98*(1), 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 46-66.

Krosnick, J. A., & Abelson, R. P. (1992). The case for measuring attitude strength in surveys. *Questions about questions: Inquiries into the cognitive bases of surveys*, 177-203.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in nursing & health*, *25*(4), 295-306.

Lindeboom, M., & Van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported

health. *Journal of health economics*, *23*(6), 1083-1099.

Mojtabai, R. (2016). Depressed mood in middle-aged and older adults in Europe and the United States: a comparative study using anchoring vignettes. *Journal of aging and health*, *28*(1), 95-117.

Paccagnella, O. (2014). A New Tool for Measuring Customer Satisfaction: the Anchoring Vignette Approach. *Italian Journal of Applied Statistics*.

Salomon, J. A., Tandon, A., & Murray, C. J. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *Bmj*, *328*(7434), 258.

Guardado, M. (2003). Language as Cultural Practice: Mexicanos en el Norte. *TESOL Quarterly*, *37*(1), 192-193.

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195-217.

CHAPTER II

Using Anchoring Vignettes to Control for Response Styles: Validity of a Multidimensional IRT Model Approach

2.1 Introduction

This chapter focuses on using anchoring vignettes to address response styles in a cross-cultural context. Response style is defined as respondents' systematic stylistic tendency to use the response scales in a certain way regardless of the question content (Harzing et al. 2012; Yang et al. 2010; Vaerenbergh & Thomas 2013), which is a potential threat to the data reliability and validity. Studies have described various types of response styles. For example, extreme response style (ERS) is the tendency to select the two extreme endpoints of a scale; midpoint response style (MRS) refers to the consistent selection of middle or neutral category of the scale; and acquiescent response style (ARS) describes the tendency to agree with the given question or to select the positive responses (Vaerenbergh & Thomas 2013; Harzing et al. 2012; He & van de Vijver 2013).

Research has shown that response style is associated with personality traits (He & van de Vijver 2013) and are stable across survey content and time (Weijters et al. 2010a), highlighting the widespread concern that response style is likely to be a source of systematic measurement error. For example, agree-disagree Likert scales are prone to ARS, which can lead to positive bias (Kam & Zhou 2014; Dolnicar & Grun 2009), where

the estimated agreement rating is likely to be artificially inflated(Baumgartner & Steenkamp 2001). ERS skews the response distributions toward to the endpoints of the scales, which increases the variance (Kieruj & Moors 2010; Baumgartner & Steenkamp 2001).

Recently, Bolt and colleagues (2014) proposed a multidimensional item response theory (IRT) model that evaluates and controls for response styles using anchoring vignette questions – a set of questions that describe hypothetical individuals' situations related to the domain of interest along with self-assessment questions (Bolt et al. 2014; King et al. 2004). The inclusion of anchoring vignette data in this approach allows researchers to control for all types of response styles by including latent response style factors and a latent substantive trait factor in the model. However, whether this multidimensional IRT model can effectively control for response styles remains an open question. To investigate this question, this study evaluates the concurrent validity of this approach by comparing the latent substantive trait variable with several criteria in a validation model.

2.1.1 Response Style Evaluation and Control

A key challenge in evaluating response styles is the difficulty of separating subjective traits from response styles (Bolt et al. 2014; Plieninger & Meiser 2014; Weijters et al. 2010a; Weijters et al. 2010b). For example, if a respondent selects "strongly agree" for an attitude question, it is hard to know whether the choice reflects a highly positive attitude (substantive trait) with no influence of ARS or a neutral or negative attitude influenced by a high level of ARS. A number of studies use heterogeneous, irrelevant items to evaluate response styles (Baumgartner & Steenkamp

2001; Greenleaf 1992; Weijters et al. 2010b) – e.g., ARS can be evaluated by investigating the level of agreement among many uncorrelated items. Similarly, ERS can be measured by evaluating the number or proportion of heterogeneous items where respondents provide extreme responses (Baumgartner & Steenkamp 2001). The rationale behind such evaluations is that "an average across a large number of heterogeneous items should contain no variance attributable to the content of the items and should thus reflect only the propensity of an individual to select the specific kind of category" (Plieninger & Meiser 2014).

While the use of heterogeneous items to evaluate response styles is straightforward and attractive, this approach has two potential limitations. The first limitation is that it often requires additional data collection leading to increased survey time and costs. Another important limitation of this approach is that since the method does not focus on one specific substantive trait, the effect of a specific substantive trait cannot be measured. As argued by Bolt et al. (2014), despite its usefulness in studying response styles, this method will be less useful when the measurement of a substantive construct is of interest (Bolt et al. 2014). However, in survey methodology, the validity of a substantive construct is usually a concern in surveys. Specifically, survey methodologists are particularly interested in how to effectively control for or lessen the impact of response styles in order to obtain less biased substantive trait measures. For example, Liu and colleagues evaluated whether different forms of response scales (e.g., item specific scales vs. agree-disagree scale) can be free of the effect of response style (Liu et al. 2015) and found both scale forms are subject to response styles.

One way to address the drawbacks of the heterogeneous item approach involves the use of latent trait models, which can control and evaluate response styles simultaneously. Note that there are multiple statistical methods available to evaluate response styles, including a method known as the representative indicators response styles means and covariance structure (Thomas et al. 2014), standardized rating scales (He et al. 2014) or scale heterogeneity models (Rossi et al. 2001). Among these methods, latent trait models are becoming increasingly popular (e.g., Johnson & Bolt 2010; Javaras & Ripley 2007; Billiet & McClendon 2000). By including a latent substantive trait variable and (a) latent response style factor(s) in the model, response styles can be evaluated and controlled at the same time (Bolt et al. 2014; Morren et al. 2011). Examples of such models include item response theory models (de Jong, Steenkamp, Fox, & Baumgartner, 2008; Johnson & Bolt, 2010), latent class models (e.g., Morren et al., 2011), and structural equation models (Billiet & McClendon 2000).

Despite the advantage of simultaneously evaluating and controlling for response styles, the traditional latent trait model can only be used to study a limited number, or certain types of response styles – usually just one type (e.g., Liu et al., forthcoming; Morren et al., 2011) or two specific types (e.g., ERS and MRS through multiprocess IRT model) (Plieninger & Meiser 2014; Böckenholt 2012). This suggests the need for new approaches to more comprehensively control for response styles.

2.1.2 The Use of a Multidimensional IRT Model Including Anchoring Vignette

The idea of Bolt et al. (2014)'s model is to include both latent substantive trait factor and multiple latent response style factors in the model, where the response style

factors are indicated by both self-reported measures and vignette questions (see Figure 2.1). Based on the vignette equivalence (VE) and reporting consistency (RC) assumptions of anchoring vignettes (see Chapter 1), the variability in anchoring vignette responses only reflects response styles (VE) (Bolt et al. 2014), and the style factors have the same influence on both self-reports and vignette questions (RC) (King et al. 2004; Bolt et al. 2014). See section 2.2 for model details.

As proposed by Bolt et al. (2014), the inclusion of anchoring vignettes in a multidimensional IRT model enables researchers to solve the limitations of prior methods. First, unlike the heterogonous-item approach, the presence of anchoring vignette questions in this model avoids the frequent confounding effects of response styles with substantive traits, and allows the evaluation of the substantive trait (Bolt et al. 2014). Second, traditional latent trait models can control for only one or two response styles, this model, with the inclusion of anchoring vignette items, controls for all types of response styles at the individual level (Bolt et al. 2014; Billiet & McClendon 2000; Moors 2008; Johnson & Bolt 2010). Third, it can evaluate the assumptions of anchoring vignettes, as shown in Bolt et al. (2014). Last, this method, by using a multi-group version of this multidimensional IRT model (see 2.3.1), can control for response styles at both individual level and at group or country level.

2.1.3 Goal of this Study

The multidimensional IRT model proposed by Bolt et al. (2014) separates the latent substantive trait and the latent response style factors through the use of anchoring vignettes. Importantly, in their paper, Bolt et al. (2014) call for further research to

evaluate how well response styles can be controlled by the model. To answer this call, this proposal has the following objective, as described in detail below.

Given that this model can separate the latent substantive trait and response styles, one can reasonably assume that after controlling for all different types of response styles, the latent substantive trait will be less biased and closer to the true score. However, as noted by Plieninger and Meiser (2014), the statistical model itself is "silent" about the interpretation of the latent substantive trait variable. In other words, the model itself does not inform us whether the latent substantive trait after controlling for response styles is closer to the true score. This highlights the need for future studies on how well this model controls for response styles. The overall goal of this study is thus to evaluate the concurrent validity of this multidimensional IRT model approach, with respect to how well it controls for response styles. Specifically, this study has three research questions:

- Overall, is the latent substantive trait closer to true score, when controlling for response styles using anchoring vignettes?
- Does the concurrent validity of this approach vary by the choice of objective benchmarks? (In this study, I use several objective benchmarks as proxies for the true score.)
- Does the correction of response styles vary by country?

I apply this validation approach (described in Section 2.3) using existing data from the Survey of Health, Ageing and Retirement in Europe (SHARE) in this study (Dataset is described in Section 2.4).

2.2 Specification of Bolt et al. (2014) Model

Figure 2.1 provides a conceptual illustration of this multidimentional IRT model proposed by Bolt et al. (2014).

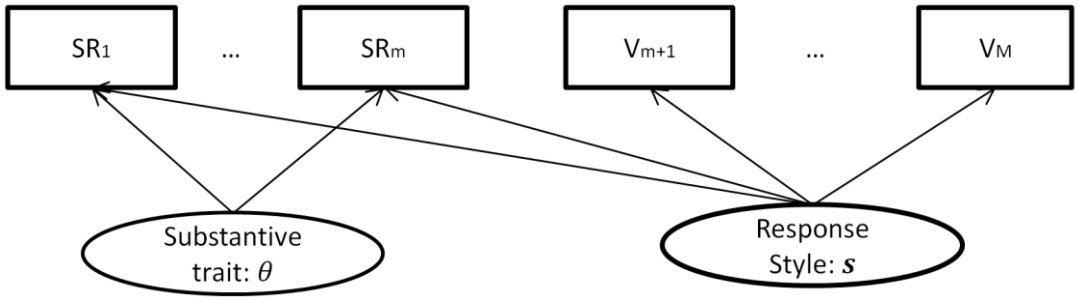


Figure 2. 1 Multidimensional IRT model controlling for response styles. SR_1 to SR_m are self-assessment measures of the same construct (e.g., sub-domains of one construct), and V_{m+1} to V_M are anchoring vignette questions designed to correspond to the self-assessment questions. There are in total M questions (self-assessments and vignettes combined).

The statistical model for Figure 1 is presented as Equation 1. The probability that respondent r selects response category k on item i ($i = 1, \dots m$ for self-reports and $i = m + 1, \dots M$ for vignettes) can be represented as the multinomial logistic model below:

$$\mathrm{P}(Y_{ri} = \mathrm{k} \mid \theta_r, \boldsymbol{s}_r) = \frac{\exp(a_{ik}\theta_r + s_{rk} + c_{ik})}{\sum_{h=1}^{K} \exp(a_{ih}\theta_r + s_{rh} + c_{ih})} \qquad (1)$$

θ denotes the content / substantive trait level.

h denotes all response categories where $h = 1, \dots, \mathrm{K}$. For a five-point Likert scale, K=5.

$\boldsymbol{s}_r = [s_{r1}, s_{r2}, \ldots\ s_{rk}, \ldots, s_{rK}]$ is a respondent-specific reporting heterogeneity vector, which reflects respondent differences in using the response scales. For a five-point Likert scale, $\boldsymbol{s}_r = [s_{r1}, s_{r2}, \ldots\ s_{r3}, s_{r4}, s_{r5}]$.

c_{ik} is the intercept parameter, which represents item differences in selecting each category (e.g., item "difficulty").

In this model, a_{ih} is the slope parameter for θ. $a_{ih} = 0$ for $h = 1, \ldots,$ K (or, $a_i = [0, 0, 0, 0, 0]$) is set for all vignette items, so that only self-assessments are used to define the content latent variable, θ. As described in Bolt et al. (2014), $a_i = [-2, -1, 0, 1, 2]$ is set for all self-rating items, which is a common assumption in analyses of polytomous IRT models (See Bolt et al. (2014)).

The VE assumption is made by having only the latent response style factors influence the vignette items (e.g., only s_{rk} and c_{ik} relates with the vignette items, as shown in Figure 2.1 and Model 1). The RC assumption is made by fixing the response style parameters the same across self-assessments and vignette items (e.g., s_{rk} does not differ by items in Equation 1).

To estimate this model, following Bolt et al. (2014), a Bayesian framework is used as it is highly flexible and can deal with the high dimensionality of this model (Bolt et al. 2014). Unlike the frequentist approach, Bayesian estimation considers the model parameters to be random, instead of fixed. Therefore, Bayesian methods use both the prior distributions for the model parameters and the data to estimate the model results – the posterior probability distributions, which reflect the probability distribution of the

unknown random parameters, conditional on the evidence obtained from the survey data (Kaplan 2012).

2.3 The Proposed Validation Approach

2.3.1 Multi-group multidimensional IRT model

First I use the multi-group multidimensional IRT model on a cross-cultural dataset following Bolt et al. (2014). The model is specified as below (Model 2), which adds group or country-level effects into the multidimensional IRT model (Model 1), where j represents country.

$$\mathrm{P}\left(Y_{ri(j)} = \mathrm{k} \mid \theta_{r(j)}, \boldsymbol{s}_{r(j)}\right) = \frac{\exp(a_{ik}\theta_{r(j)} + s_{rk(j)} + c_{ik})}{\sum_{h=1}^{K} \exp(a_{ih}\theta_{r} + s_{rh} + c_{ih})} \qquad (2)$$

where the distribution of the latent θ and response style traits s_{rk} are assumed to differ both at the individual and the country levels. Specifically, this model assumes "invariance" of the item category intercepts (c_{ik}) across countries and a "constant within-country covariance matrix of the latent traits" (Bolt et al. 2014). This is reflected in the Bayesian approach (see section 2.3.3) by estimating the mean vectors for the latent θ and response style factors at the country level.

2.3.2 Validation model approach

To validate the use of this model in response style-control, I use several benchmarks to validate the latent substantive trait variable in the multidimensional IRT model (a validation model), specifically I compare the concurrent validity between the model with vignette questions (see Figure 2.2b) and the model without vignette

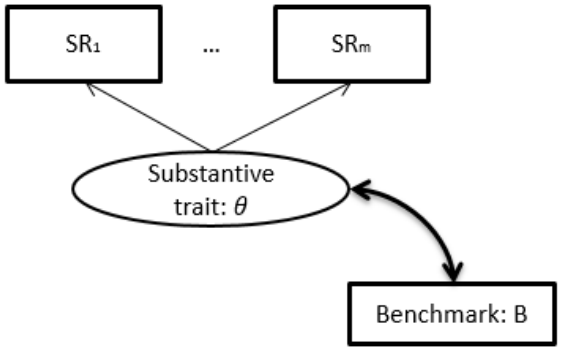
questions [3] (see Figure 2.2a). This approach of including benchmark for validity measurement has been applied in many previous studies (Kam & Zhou 2014; Billiet & McClendon 2000). Different from prior work, the novelty of this research lies in the validity of response-style control using this multidimensional IRT model approach.

As shown in Figure 2.2, the overall validation model contains equation 2 and the correlation between the substantive trait (θ) and the benchmarks (B), as indicated by equation 3.

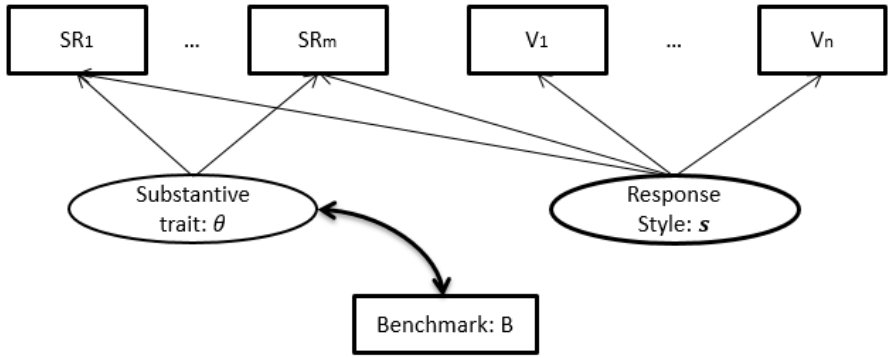
$$corr(\theta, B) = \frac{cov(\theta, B)}{\sigma_\theta \sigma_B} \qquad (3)$$

My hypothesis is that the model fit is better in the model with anchoring vignettes. The correlation between θ and the benchmark in the model with vignettes is expected to be higher than in the model without vignettes.

[3] This model in Figure 2.2a differs from Model 1 (which assumes no response style heterogeneity across respondents) in Bolt et al. (2014). Although it does not include latent response style factors, Bolt et al. (2014)'s Model 1 still includes anchoring vignettes and estimates the item category intercept associated with the vignettes. The current model in Figure 2.2a is a more parsimonious model which excludes anchoring vignettes.



(2.2a)



(2.2b)

Figure 2. 2 Validation models to evaluate the concurrent validity of the multidimensional IRT model. The correlation coefficients between θ and B in the two models are compared based on their 95% credible intervals.

2.3.3 A Bayesian Estimation Algorithm

To estimate the multidimensional IRT model, following Bolt et al. (2014), a Bayesian estimation approach was used with Markov chain Monte Carlo (MCMC) simulation through Gibbs sampling.

For the multidimensional IRT model where anchoring vignettes are used to control for all types of response styles, the priors for the parameters are set the same as Bolt et al. (2014)'s paper. $c_{ik} \sim Normal(0,5)$ is assigned as priors for the item category intercept.

For response scales with five response categories, following Bolt et al. (2014),

$$\theta_{r(j)} = [\theta_{r(j)}, s_{r(j)1}, s_{r(j)2}, s_{r(j)3}, s_{r(j)4}, s_{r(j)5}]$$

$$\theta_{r(j)} \sim MultiNormal(\mu_j, \sum)$$

where

$$\mu_j \sim MultiNormal(0, I_{6\times6}), \text{ and } \sum^{-1} \sim Wishart(41I_{3\times3}, 40)$$

which indicates that each country differs in terms of their mean latent substantive trait, θ, and the propensity to select each category (five in total).

Parameters were estimated using the software WinBUGS 1.4. Correlations between models were compared based on the 95% credible intervals. For each model, I simulated two chains with different initial values for each of the parameters. Simulated chains were carried out using at least 10,000 iterations following 5000 burn-in iterations. Multiple diagnostics were used to examine the convergence conditions, including Gelman and Rubin's diagnostics.

2.4 Application of the Validation Approach Using Existing Dataset

To apply the validation approach, I use an existing cross-cultural health survey dataset from the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE; Alcser, Avendano, et al., 2005; Alcser, Benson, et al., 2005; Börsch-Supan et al., 2013) collected in 2004–2005. SHARE is selected for the following two reasons: 1) it not only collects self-assessments and vignettes data on different health domains, but also includes a comprehensive set of objective measures on these domains (as given in detail below), and 2) with data from multiple countries available, it provides a good data source for cross-country response-style evaluations and control. Controlling for country-level response style is important because response styles have been found to vary across cultures, and ignoring this phenomena may confound substantial cross-cultural differences, thus leading to biased results (Harzing et al. 2012; Mottus et al. 2012). In summary, with SHARE data, this study can examine whether the latent substantive trait variable is closer to the benchmark measure (obtained from objective measures in the survey) when controlling for response style than when not.

2.4.1 Self-Assessments and Vignettes Data in SHARE

SHARE is a cross-national biennial survey of individuals in Europe who are 50 and over. The SHARE vignette questionnaire was administered to a subsample [4] through a self-administered paper and pencil questionnaire (see Appendix 2.1 for a list of anchoring vignette questions asked in SHARE). In the vignette questionnaire, respondents were asked both self-assessment questions and vignettes on six health domains: mobility, cognition, pain, sleep, breathing and emotional health. For both self-assessments and vignette questions, a five-point

[4] Respondents who completed vignette questionnaire also completed the main SHARE questionnaire.

response scale from none to extreme was used to rate problems and difficulties they experienced in last 30 days. After each self-assessment question, respondents are asked three vignette questions on the same health domain. Respondents are asked to rate the vignette questions as if they were rating for themselves. In this analysis, I use five domains in analysis – mobility, pain, sleep, breathing and emotional health, leaving to a total of five self-assessment questions and fifteen vignette questions. A test of the measurement model using the self-assessment questions shows that these domains load on the same latent variable.

In total, respondents from eight countries were assigned to complete the vignette questionnaire. The total sample size is 4544, with respective sample sizes by country as below: Germany (n = 508), Sweden (n = 417), the Netherlands (n = 538), Spain (n = 464), Italy (n = 445), France (n = 885), Greece (n = 720) and Belgium (n = 567). In the analysis, to make sure all models being compared have the same sample size, I removed the observations with missing data on any analyzed variables. The sample size for each country used in the analysis is as below: Germany (n = 472), Sweden (n = 355), the Netherlands (n = 483), Spain (n = 428), Italy (n = 382), France (n = 742), Greece (n = 640) and Belgium (n = 512).

2.4.2 Selected Benchmark Measures

Seven separate benchmark variables were selected (see Table 2.4), which include respondents' physical test result – in particular, grip strength, and respondents' responses to factual questions. Grip strength has been widely used as an objective measure to validate health (Spiegel et al. 1988; Jürges 2007). It was measured by trained interviewers with a dynamometer (Alcser, Benson, et al., 2005). Two measurements were recorded for each hand. In the study, I

used a generated variable indicating the maximum of the four outcomes for each respondent[5]. The other benchmarks were constructed using factual questions in SHARE, such as "In the last month, have you cried at all?", using a "Yes / No" scale. These characteristics differ from the six subjective self-assessment questions as they use a "Yes / No" scale whereas self-assessment uses a five-point Likert rating scale (e.g., from None to Extreme). This indicates that such questions are unlikely to be subject to response styles influencing the subjective self-assessment questions, such as ERS and MRS. Although such scales can be subject to ARS, given that these questions are straightforward factual questions, it is reasonable to believe that response style is less an issue for these items than for the subjective self-rating questions (Sigelman et al. 1981). To illustrate how well this method controls for response styles, I describe the results of two benchmarks (i.e., grip strength and the number of limitations with activities of daily living: ADL) in detail. More information on the other benchmark measures can be found in Table 2.4 and Appendix 2.2 and 2.3.

The validation approach as described in Section 2.3 was applied in this dataset to evaluate whether the use of anchoring vignettes in SHARE can effectively control for response styles.

2.5 Results

As shown in Tables 2.1 and 2.2, I first examined the distributions of the self-assessment question (i.e., percentage of reported moderate to extreme health problems) and the means of the benchmarks by country for each domain. I also ranked each country's health levels, where a smaller number in the rank orders indicates better health. Table 2.2 shows that Germany and

[5] It is known that grip strength differs by gender and age. I checked the gender and age distributions across countries. They are in general comparable across countries.

Sweden have similar means for most of the health benchmarks. We would expect they would report similarly in the self-assessment questions. However, as shown in Table 2.1, their self-reported health differ greatly for the pain, sleep and emotional health domains. Another counterintuitive result is that Sweden ranks better than Spain, Italy and France for all the health benchmarks. Swedish respondents, however, report worse health for breath, mobility and emotional health domains. This inconsistencies between objective health and self-reports could likely be caused by response styles in reporting. This highlights the importance to address response styles in research.

Table 2. 1 Percentage and rank orders of reported moderate to extreme health problems on self-rated questions by health domain and country.

	Pain		**Sleep**		**Mobility**		**Breath**		**Emotional health**	
	%	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**
Germany	37.2	7	30.7	4	26.1	8	13.2	6	20.2	3
Sweden	17.5	2	13.1	1	22.9	6	30.9	8	27.8	8
Netherlands	17.4	1	22.2	3	14.6	2	6.0	1	8.6	1
Spain	36.2	6	30.7	4	26.1	7	9.6	3	25.2	6
Italy	32.1	5	32.1	6	19.0	5	9.6	4	22.7	5
France	38.1	8	40.9	8	17.4	3	16.8	7	21.4	4
Greece	29.6	3	17.6	2	9.4	1	8.3	2	27.2	7
Belgium	31.8	4	37.3	7	17.6	4	11.1	5	17.7	2

Note: a smaller number in rank indicates better health.

Table 2. 2 Mean and rank orders of selected benchmarks by country.

	Max. of grip strength measure		**ADL**		**Charlson index**		**# of symptoms**		**Mobility limitations**		**IADL**		**EURO-D**	
	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank	%	Rank
Germany	36.9	2	0.1	3	1.3	4	1.4	3	1.2	4	0.1	1	1.8	3
Sweden	36.1	3	0.1	4	1.2	3	1.5	5	0.9	2	0.1	4	1.8	2
Netherlands	37.4	1	0.1	2	1.0	1	1.0	1	0.8	1	0.1	3	1.6	1
Spain	29.2	8	0.1	6	1.5	6	1.7	8	1.8	8	0.3	8	2.7	6
Italy	31.6	7	0.1	5	1.4	5	1.6	6	1.3	7	0.2	5	2.7	8
France	33.2	6	0.2	7	1.5	8	1.6	7	1.3	6	0.2	7	2.7	7
Greece	33.5	5	0.1	1	1.2	2	1.1	2	1.1	3	0.1	2	2.2	4
Belgium	35.3	4	0.2	8	1.5	7	1.4	4	1.3	5	0.2	6	2.3	5

Note: a smaller number in rank indicates better health. ADL refers to the number of limitations with activities of daily living. IADL refers to number of limitations with instrumental activities of daily living.

Table 2. 3 Model fit results.

Model	DIC
1. Without Vignettes (Figure 2a)	43,438.3
2. With Vignettes (Figure 2b)	177,697

Table 2.3 shows the model fit results. Based on the deviance information criterion (DIC), unexpectedly, Model 1 is superior to Model 2 and fit the data better.

Table 2. 4 Correlation estimates between θ and B for the models with and without vignettes.

Model	Without Vignettes (Figure 2a)		With Vignettes (Figure 2b)	
Benchmark	Correlation Estimates between θ and B	95% Credible Interval	Correlation Estimates between θ and B	95% Credible Interval
Maximum grip strength	-0.24	(-0.26, -0.23)	-0.22	(-0.23, -0.20)
ADL	0.22	(0.21, 0.24)	0.17	(0.16, 0.19)
Charlson index	0.34	(0.32, 0.35)	0.27	(0.25, 0.29)
# of symptoms	0.46	(0.44, 0.47)	0.33	(0.32, 0.35)
Mobility limitations	0.42	(0.41, 0.43)	0.32	(0.31, 0.34)
IADL	0.23	(0.22, 0.23)	0.18	(0.16, 0.19)
Depression EURO-D	0.46	(0.45, 0.47)	0.35	(0.34, 0.37)

Note. Sample size is 4014. 95% Credible Interval, where the probability that the parameter lies in this particular interval is 95%. ADL, limitations with activities of daily living; IADL, limitations with instrumental activities of daily living.

For the seven benchmark criteria used to evaluate which model can better control for RS, surprisingly, contrary to our hypothesis, for all the criteria used, the correlation between θ (higher number of θ indicates more health problems or worse health) and the benchmark criteria in the model with vignettes (Model 2) are lower than in the model without vignettes (Model 1) (See Table 2.4). For each criterion, the result is significant given that there is no overlap between the two 95% credible intervals. This suggests that adding the vignettes does not necessarily bring

latent health closer to the benchmarks. In other words, the use of anchoring vignettes in this multidimensional IRT model does not seem to effectively control for response styles in the SHARE data.

To better understand why adding the vignettes in this model did not effectively control for RS, I examined the correlations between the latent variables, as shown in Table 2.5. These correlations are based on 10,000 post burn-in iterations for Model 2. Ideally, if the RC and VE are not violated, we would expect to see no correlation between RS tendencies and θ, since vignettes are designed to only reflect RS tendencies but not any of the substantive trait. However, unlike Bolt et al. (2014), we see that the response style tendencies are correlated with the substantive trait – θ: The correlation between θ and s_1 is 0.52, and it shows a general decreasing pattern from the first category (s_1) to the last category (s_5). This shows that respondents with worse perceived health (higher θ) are more likely to select the first two response categories – "none" and "mild". This reflects the way respondents rate for the vignette questions – for a given vignette person, those with poor health tend to rate the vignette as having better health and those with very good health tend to rate the vignette as having worse health. This implies that, when rating the vignette questions, respondents may use themselves as a reference point in evaluating others. It seems that a comparison-based contrast effect (Schwarz & Bless 1992) takes place here, where respondents compare the vignette questions with their own health. This reveals the VE assumption violation in the SHARE data, which assumes that respondents, no matter background, view the vignette persons in the same way.

As expected, the correlation between RS tendencies for adjacent categories (e.g., between s_1 and s_2) is positive and high, suggesting that if respondents have high propensity to select the first category, they also have a high propensity to select the second category. Consequently,

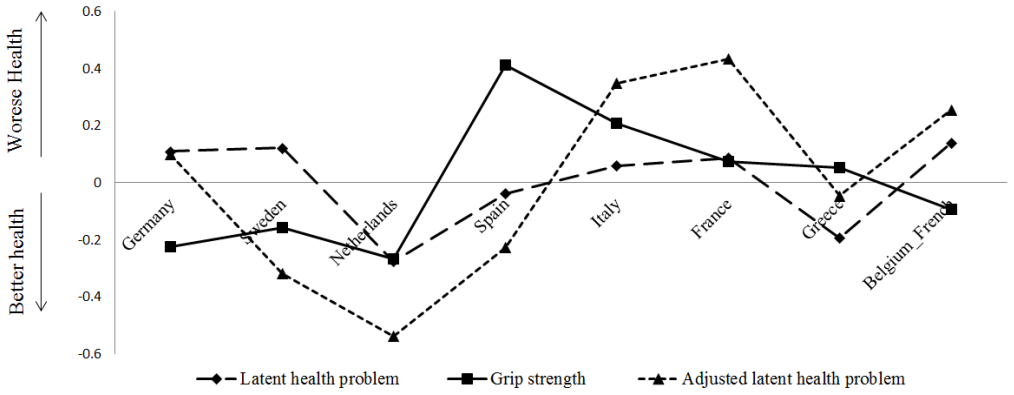
respondents with a high propensity to select the first category are less likely to select the last two categories, with the correlations between s_1 and s_4, s_1 and s_5 being negative. Similar patterns were found for other categories.

Table 2. 5 Respondent-level correlation estimates among health and response styles.

Variable	θ	s_1	s_2	s_3	s_4	s_5
θ						
s_1	0.54					
s_2	0.32	0.73				
s_3	-0.09	0.02	0.45			
s_4	-0.42	-0.58	-0.32	0.43		
s_5	-0.39	-0.64	-0.61	-0.09	0.67	

Note. θ is the latent health problems, where higher θ indicates worse health. s_1 to s_5 indicates the propensity to select the first (None) to the fifth (Extreme difficulty) categories.

2.3A



2.3B

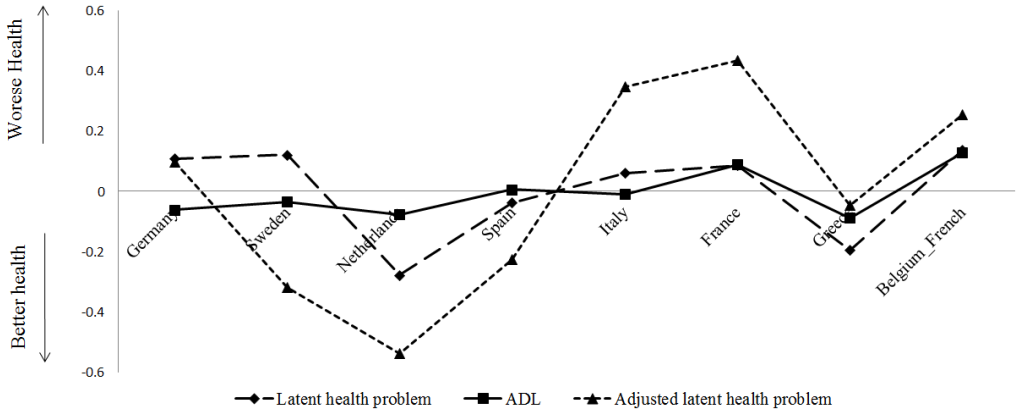


Figure 2. 3 Comparison of latent health across countries before and after adjustments using anchoring vignettes in relation to negative grip strength (A) and ADL (B).

Figures 2.3A and 2.3B show the standardized mean latent health before and after adjustments using anchoring vignette in comparison with the objective measures: the grip strength and ADL. The grip strength and ADL measures are standardized with mean of 0 and a standard deviation of 1. Given that the correlation between latent health and grip strength is negative (where a higher number for latent health indicates worse health and higher grip strength indicates better health), I multiplied grip strength by *-1*in Figure 2.3A so that higher numbers for all three measures indicate worse health. In general, Figure 2.3 shows that the effects of the RS control using anchoring vignettes vary by country. In Figure 2.3A, for Germany, there is no difference between the model with vignettes and the model without. For some countries like Greece, the adjusted health mean is closer to the benchmark. For some other countries like Netherlands and France, however, the adjusted latent health deviates more from the grip strength benchmark. In Figure 2.3B, it shows that the pattern of the latent health before adjustment matches closely with the ADL benchmark, while the adjusted latent health deviates from the benchmark. Similar patterns were found for other benchmarks (see Appendix 2.3).

To further evaluate whether the validity of this method is heterogeneous across countries, I replicated the same validation approach for each country separately. Results are shown in Table 2.6. Similar as we see in Table 2.4, it shows that for each country and all seven benchmarks, the correlation between θ and the benchmark is lower in the model with vignettes, indicating that vignettes are not effectively controlling for RS for individuals in each country.

Table 2. 6 Model comparison results for each country.

Model	Without Vignettes (Figure 2a)		With Vignettes (Figure 2b)	
Benchmark	Correlation Estimates between θ and B	95% Credible Interval	Correlation Estimates between θ and B	95% Credible Interval
Germany (n=472)				
Maximum grip strength	-0.18	(-0.22, -0.13)	-0.17	(-0.22, -0.12)
ADL	0.37	(0.33, 0.40)	0.31	(0.27, 0.36)
Charlson index	0.43	(0.39, 0.47)	0.30	(0.25, 0.35)
# of symptoms	0.38	(0.34, 0.41)	0.34	(0.29, 0.38)
Mobility limitations	0.18	(0.14, 0.22)	0.17	(0.12, 0.21)
IADL	0.21	(0.17, 0.24)	0.19	(0.14, 0.23)
Depression EURO-D	0.45	(0.41, 0.49)	0.33	(0.28, 0.37)
Sweden (n=355)				
Maximum grip strength	-0.27	(-0.32, -0.21)	-0.25	(-0.32, -0.18)
ADL	0.25	(0.20, 0.30)	0.21	(0.14, 0.27)
Charlson index	0.48	(0.43, 0.52)	0.43	(0.37, 0.48)
# of symptoms	0.39	(0.35, 0.44)	0.32	(0.26, 0.38)
Mobility limitations	0.17	(0.12, 0.22)	0.15	(0.09, 0.21)
IADL	0.28	(0.23, 0.32)	0.23	(0.18, 0.29)
Depression EURO-D	0.50	(0.45, 0.54)	0.43	(0.37, 0.49)
Netherlands (n = 483)				
Maximum grip strength	-0.19	(-0.24, -0.14)	-0.21	(-0.27, -0.16)
ADL	0.29	(0.24, 0.34)	0.22	(0.17, 0.27)
Charlson index	0.42	(0.38, 0.46)	0.28	(0.22, 0.33)
# of symptoms	0.40	(0.36, 0.44)	0.31	(0.27, 0.36)
Mobility limitations	0.23	(0.19, 0.26)	0.17	(0.13, 0.21)
IADL	0.22	(0.17, 0.26)	0.19	(0.15, 0.24)

Depression EURO-D	0.42	(0.38, 0.47)	0.21	(0.16, 0.27)
Spain (n = 428)				
Maximum grip strength	-0.32	(-0.36, -0.28)	-0.26	(-0.31, -0.21)
ADL	0.45	(0.42, 0.49)	0.37	(0.32, 0.42)
Charlson index	0.51	(0.47, 0.54)	0.41	(0.36, 0.45)
# of symptoms	0.53	(0.50, 0.56)	0.44	(0.39, 0.49)
Mobility limitations	0.26	(0.22, 0.29)	0.22	(0.18, 0.27)
IADL	0.29	(0.25, 0.32)	0.24	(0.20, 0.29)
Depression EURO-D	0.52	(0.49, 0.55)	0.44	(0.39, 0.48)
Italy (n = 382)				
Maximum grip strength	-0.29	(-0.34, -0.24)	-0.18	(-0.24, -0.12)
ADL	0.34	(0.30, 0.38)	0.23	(0.18, 0.28)
Charlson index	0.48	(0.44, 0.51)	0.31	(0.25, 0.36)
# of symptoms	0.48	(0.45, 0.52)	0.29	(0.24, 0.34)
Mobility limitations	0.18	(0.15, 0.22)	0.08	(0.03, 0.13)
IADL	0.25	(0.21, 0.28)	0.13	(0.07, 0.17)
Depression EURO-D	0.53	(0.49, 0.56)	0.37	(0.31, 0.42)
France (n = 742)				
Maximum grip strength	-0.25	(-0.29, -0.22)	-0.21	(-0.25, -0.17)
ADL	0.34	(0.31, 0.37)	0.27	(0.23, 0.31)
Charlson index	0.45	(0.42, 0.48)	0.31	(0.27, 0.35)
# of symptoms	0.40	(0.37, 0.43)	0.29	(0.25, 0.33)
Mobility limitations	0.27	(0.24, 0.30)	0.19	(0.15, 0.23)
IADL	0.22	(0.19, 0.26)	0.16	(0.12, 0.20)
Depression EURO-D	0.48	(0.45, 0.51)	0.32	(0.28, 0.36)

Greece (n = 640)				
Maximum grip strength	-0.27	(-0.31, -0.22)	-0.24	(-0.29, -0.19)
ADL	0.26	(0.22, 0.30)	0.21	(0.16, 0.26)
Charlson index	0.40	(0.36, 0.43)	0.33	(0.29, 0.38)
# of symptoms	0.37	(0.34, 0.41)	0.32	(0.28, 0.37)
Mobility limitations	0.22	(0.19, 0.25)	0.20	(0.15, 0.24)
IADL	0.22	(0.18, 0.26)	0.20	(0.16, 0.24)
Depression EURO-D	0.39	(0.35, 0.43)	0.33	(0.28, 0.38)
Belgium_French (n = 512)				
Maximum grip strength	-0.21	(-0.25, -0.16)	-0.23	(-0.28, -0.18)
ADL	0.33	(0.29, 0.37)	0.24	(0.19, 0.28)
Charlson index	0.46	(0.42, 0.50)	0.33	(0.28, 0.38)
# of symptoms	0.39	(0.35, 0.43)	0.32	(0.27, 0.36)
Mobility limitations	0.19	(0.14, 0.23)	0.14	(0.09, 0.19)
IADL	0.22	(0.18, 0.27)	0.18	(0.13, 0.23)
Depression EURO-D	0.43	(0.39, 0.47)	0.35	(0.30, 0.40)

Note. ADL, limitations with activities of daily living; IADL, limitations with instrumental activities of daily living.

2.6 Discussion

Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), this study examined whether a newly proposed IRT model using anchoring vignettes can effectively disentangle the response styles and respondents' actual perceptions. Results from this study indicate that this method does not necessarily control for response styles in SHARE, and the adjusted latent health measures do not seem to be closer to the objective benchmarks. Consistent

results were found for all the benchmarks selected and for each of the eight countries. The reason for why vignettes failed in controlling for response styles in SHARE may be due to the violation of VE assumption. It is shown that respondents with different health backgrounds tend to view the vignettes differently – those with poor health tend to rate vignette people as having less health problems (contrast effect). This also highlights the difficulties to construct anchoring vignettes that reveals the same information to each respondent (or meet VE assumption). This research has general implications for the cautious use anchoring vignettes and highlights the importance of the design aspects of vignettes, which may greatly influence the validity of the anchoring vignette approach.

It is important to note that this study is limited in several ways. First, the use of SHARE restricts the proposed study to the evaluations of up to three vignettes. In practical applications, however, some studies may have more than 3 vignette items, such as 5 or 7 vignette items, like World Health Survey (WHS). Future studies can evaluate the effects of using more than 3 vignette items. Second, the objective measures of SHARE data in this proposal contain both physical measures and self-report measures to factual questions. It is noted that the self-reported questions may contain measurement error, such as ARS. However, these are straightforward objective questions (e.g., whether has been diagnosed with certain disease), which have been reported as having good agreement with medical records data in general (Harlow & Linet 1989; Kilbourne et al. 2017). It is thus reasonable to assume that these questions are of better quality than subjective question. In addition to the lower possibility of response styles than self-assessment questions, all seven benchmarks used in this research (including the physical tests of grip strength), lead to consistent results.

There are several areas where additional research is needed. First, as seen in this paper, the success of this model and its validity may largely depend on the assumptions underlying the use of anchoring vignette questions (King et al. 2004; Peracchi & Rossetti 2013), which can be violated in various ways (Bago d'Uva et al. 2011; Kapteyn et al. 2011; Rice et al. 2011). In light of this possibility, simulation studies are essential to evaluate the concurrent validity of this model under the violations of RC and / or VE. Second, future research should further develop and improve this model, and also evaluate alternative latent trait models using anchoring vignette to control for response styles. One potential way to improve this model is to allow the parameters of the response style latent factors to differ across different domains. As for the evaluation of alternative models, for example, one can replace the latent response style factors in the multidimensional IRT model with a latent class variable which classifies respondents into different response style types. Similar models, although without the use of anchoring vignettes, are reported in previous literature to study response styles with a latent class response style variable (Gollwitzer, Eid, & Jurgensen, 2005; Moors, 2003). Third, the traditional method for analyzing anchoring vignette data is the hierarchical ordered probit (HOPIT) model, which uses anchoring vignettes to investigate where respondents' anchoring points (cut-points) lie on the continuum of the true state, in order to control for different usage of response scales. HOPIT and the multidimensional IRT model differ in many ways. For example, HOPIT model depends more on the covariates used to model the true latent health and underlying cut-points for the response scales. Each sub-domain (e.g., pain) can be corrected using the HOPIT model. The multidimensional IRT model, on the other hand, does not depend on covariates, and works better correcting a latent construct with several sub-domains. No studies so far have compared the two models. It is also unclear which of the two models can more effectively control for response

styles. Future studies can compare the validity of HOPIT model and the multidimensional IRT model. Fourth, little is known about the effect of different survey designs (e.g., different question orders) on vignettes. Previous studies of context effects on vignettes have focused only on the U.S. population (e.g., Buckley, 2008; Hopkins & King, 2010), while anchoring vignettes are commonly applied to cross-national comparisons. Future studies can extend this to cross-cultural surveys and explore better design options for vignette questions. For example, researchers can evaluate different designs of vignette introductions (e.g., emphasize respondents to follow the RC assumptions or not), randomize the order of self-reported questions (e.g., before and after the vignette items), experiment on the number and descriptions of vignettes and examine other design options, such as using pictures or videos instead of verbal descriptions for vignettes, in a cross-cultural survey context. In Chapter 3 and 4, I further evaluate the use of pictures as an alternative approach to verbal vignettes. Such research can shed light on improving the future design of anchoring vignettes.

2.7 References

Alcser, K. H., Avendano, M., Börsch-Supan, A., Brunner, J. K., Cornaz, S., Dewey, M., … Winter-Ebmer, R. (2005). *Health, ageing and retirement in Europe: first results from the Survey of Health, Ageing and Retirement in Europe.* Mannheim: Mannheim Research Institute for the Economics of Aging (MEA). Retrieved from http://www.share-project.org/uploads/tx_sharepublications/SHARE_FirstResultsBookWave1.pdf

Alcser, K. H., Benson, G., Börsch-Supan, A., Brugiavini, A., Christelis, D., Croda, E., … Weerman, B. (2005). *The Survey of Health, Aging, and Retirement in Europe–Methodology.* Mannheim Mannheim Research Institute for the Economics of Aging (MEA). Retrieved from http://www.share-project.org/uploads/tx_sharepublications/SHARE_BOOK_METHODOLOGY_Wave1.pdf

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & van Doorslaer, E. (2011). Slipping Anchor?: Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Journal of Human Resources*, *46*(October 2009), 875–906.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*(May), 143–156.

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678.

Bollen, K. A. Structural Equations with Latent Variables. 1989 New York. *NY Wiley*.

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using

anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., … Zuber, S. (2013). Data Resource Profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*(4), 992–1001.

Buckley, J. (2008). Survey context effects in anchoring vignettes. URL http://polmeth. wustl. edu/media/Paper/surveyartifacts. pdf.

de Jong, M. G., Steenkamp, J.-B. E. ., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research*, *45*(1), 104–115.

Dolnicar, S., & Grun, B. (2009). Response Style Contamination of Student Evaluation Data. *Journal of Marketing Education*, *31*, 160–172.

Gollwitzer, M., Eid, M., & Jurgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological assessment*, 17(1), 56-68.

Greenleaf, E. (1992). Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles. *Journal of Marketing Research*, *29*(2), 176–188.

Harlow, S. D., & Linet, M. S. (1989). Agreement between questionnaire data and medical records: the evidence for accuracy of recall. *American Journal of Epidemiology*, *129*(2), 233–248.

Harzing, A. W., Brown, M., Köster, K., & Zhao, S. (2012). Response Style Differences in Cross-National Research: Dispositional and Situational Determinants. *Management International Review*, *52*, 341–363.

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response Styles and Personality Traits: A Multilevel Analysis. *Journal of Cross-Cultural Psychology*, *45*, 1028–

1045.

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, *55*(7), 794–800.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*(681 Icc), 201–222.

Javaras, K. N., & Ripley, B. D. (2007). An "Unfolding" Latent Variable Model for Likert Attitude Data. *Journal of the American Statistical Association*, *102*(May 2014), 454–463.

Johnson, T. R., & Bolt, D. M. (2010). On the Use of Factor-Analytic Multinomial Logit Item Response Models to Account for Individual Differences in Response Style. *Journal of Educational and Behavioral Statistics*, *35*(1), 92–114.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics*, *16*(2), 163–78.

Kam, C. C. S., & Zhou, M. (2014). Does Acquiescence Affect Individual Items Consistently? *Educational and Psychological Measurement*, 1–21.

Kaplan, D. (2012). Bayesian Structural Equation Modeling. In *Handbook of Structural Equation Modeling* (pp. 650–673).

Kapteyn, A., Smith, J. P., Van Soest, A., & Vonková, H. (2011). Anchoring Vignettes and Response Consistency Consistency. *Working Paper*. Retrieved from http://www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR840.pdf

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International journal of public opinion research*, 22(3), 320-342.

Kilbourne, A. M., Schumacher, K., Frayne, S. M., Cypel, Y., Barbaresso, M. M., Nord, K. M., ...

& Gleason, T. C. (2017). Physical Health Conditions Among a Population-Based Cohort of Vietnam-Era Women Veterans: Agreement Between Self-Report and Medical Records. *Journal of Women's Health*, 26(11), 1244-1251.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enchancing the Validty and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review, 98*, 191–207.

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly*, 79(4), 952-975.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity*, 37(3), 277-302.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, 42, 779–794.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with extreme response style in cross- cultural research: a restricted latent class factor analysis approach. *Sociological Methodology*, *41*(1), 13–47.

Mottus, R., Allik, J., Realo, a., Rossier, J., Zecca, G., Ah-Kion, J., … Johnson, W. (2012). The Effect of Response Style on Self-Reported Conscientiousness Across 20 Countries. *Personality and Social Psychology Bulletin*, *38*, 1423–1436.

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

Plieninger, H., & Meiser, T. (2014). Validity of Multiprocess IRT Models for Separating Content

and Response Styles. *Educational and Psychological Measurement*, 0013164413514998-.

Rice, N., Robone, S., & Smith, P. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *European Journal of Health Economics*, *12*, 141–162.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming Scale Usage Heterogeneity. *Journal of the American Statistical Association*, *96*(January 2015), 20–31.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. *The Construction of Social Judgments*, 217–245:

Sigelman, C. K., Budd, E. C., Spanhel, C. L., & Schoenrock, C. J. (1981). When in doubt, say yes: Acquiescence in interviews with mentally retarded persons. *Mental Retardation*, *19*(2), 53–58.

Spiegel, J. S., Leake, B., Spiegel, T. M., Paulus, H. E., Kane, R. L., Ward, N. B., & Ware, J. E. (1988). What are we measuring? an examination of self-reported functional status measures. *Arthritis & Rheumatism*, *31*(6), 721–728.

Thomas, T. D., Abts, K., & Vander Weyden, P. (2014). Measurement Invariance, Response Styles, and Rural-Urban Measurement Comparability. *Journal of Cross-Cultural Psychology*, *45*(7), 1011–1027.

Vaerenbergh, Y. Van, & Thomas, T. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied*

Psychological Measurement, *34*, 105–121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*(1), 96–110.

Yang, Y., Harkness, J. a., Chin, T.-Y., & Villar, A. (2010). Response Styles and Culture. Survey Methods in Multinational, Multiregional, and Multicultural Contexts, (1984), 203–223.

Appendix 2. 1 Self-Assessment and Vignette Questions in SHARE Data

Self-assessment questions and vignette questions on six health domains in SHARE.

Pain	
Self-assessment question	
Overall, in the last 30 days, how much pain or bodily aches did you have?	
Corresponding Vignette items	
[Paul] has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs. Overall, in the last 30 days, how much pain or bodily aches did [Paul] have?	Low Pain
[Henry] has pain that radiates down his right arm and wrist during his day at work. This is slightly relieved in the evenings when he is no longer working on his computer. Overall, in the last 30 days, how much pain or bodily aches did [Henry] have?	Medium Pain
[Charles] has pain in his knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, he feels uncomfortable when moving around, holding and lifting things. Overall, in the last 30 days, how much pain or bodily aches did [Charles] have?	High Pain
Response Scale for both the self-assessment questions and vignette items:	
None Mild Moderate Severe Extreme	
Sleep	
Self-assessment question	
In the last 30 days, how much difficulty did you have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?	
Corresponding Vignette items	
[Alice] falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night. In the last 30 days, how much difficulty did [Alice] have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?	Low
[Maria] takes about two hours every night to fall asleep. She wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again. In the last 30 days, how much difficulty did [Maria] have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?	Medium
[Karen] wakes up almost once every hour during the night. When he wakes up in the night, it takes around 15 minutes for her to go back to sleep. In the morning she does not feel well-rested. In the last 30 days, how much difficulty did [Karen] have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?	High
Response Scale for both the self-assessment questions and vignette items:	
None Mild Moderate Severe Extreme	
Mobility	
Self-assessment question	

Overall, in the last 30 days, how much of a problem did you have with moving around?

Corresponding Vignette items

[Rob] 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. Overall, in the last 30 days, how much of a problem did [Rob] have with moving around?	Low
[Kevin] 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. Overall, in the last 30 days, how much of a problem did [Kevin] have with moving around?	Medium
[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. Overall, in the last 30 days, how much of a problem did [Tom] have with moving around?	High

Response Scale for both the self-assessment questions and vignette items:

None Mild Moderate Severe Extreme

Breath

Self-assessment question

In the last 30 days, how much of a problem did you have because of shortness of breath?

Corresponding Vignette items

[Mark] has no problems with walking slowly. He gets out of breath easily when climbing uphill for 20 meters or a flight of stairs. In the last 30 days, how much of a problem did [Mark] have because of shortness of breath?	Low
[Paul] suffers from respiratory infections about once every year. He is short of breath 3 or 4 times a week and had to be admitted in hospital twice in the past month with a bad cough that required treatment with antibiotics. In the last 30 days, how much of a problem did [Paul] have because of shortness of breath?	Medium
[Henry] has been a heavy smoker for 30 years and wakes up with a cough every morning. He gets short of breath even while resting anddoes not leave the house anymore. He often needs to be put on oxygen. In the last 30 days, how much of a problem did [Henry] have because of shortness of breath?	High

Response Scale for both the self-assessment questions and vignette items:

None Mild Moderate Severe Extreme

Affect

Self-assessment question

Overall, in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?

Corresponding Vignette items

[Karen] enjoys her work and social activities and is generally satisfied	Low

with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to-day activities. Overall, in the last 30 days, how much of a problem did [Karen] have with feeling sad, low or depressed?	
[Maria] feels nervous and anxious. She worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests her. When she is alone she tends to feel useless and empty. Overall, in the last 30 days, how much of a problem did [Maria] have with feeling sad, low or depressed?	Medium
[Anna] feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead. Overall, in the last 30 days, how much of a problem did [Anna] have with feeling sad, low or depressed?	High
Response Scale for both the self-assessment questions and vignette items:	
None Mild Moderate Severe Extreme	

Note: Vignette questions have the same response categories as the self-assessment question. They both have the same question wording – for example, it asks "Overall, in the last 30 days, how much of a problem did [FILL] have with feeling sad, low or depressed?" for all questions within that domain. The order of the vignette questions are randomized in the questionnaire.

Appendix 2. 2 Description of the benchmarks.

Grip strength

Respondent's handgrip strength is measured using dynamometer, where respondents are asked to press as hard as they can. The. Multiple measurements were taken, and the maximum one is recorded.

The Charlson index

The Charlson index is a weighted health index that considers the number and the seriousness of comorbid diseases developed by individuals (Charlson et al. 1987). The assigned weights for diseases using SHARE data are presented in the table below, which follows Charlson et al. (1987). The constructed Charlson index is highly correlated with the number of chronic condition index, as provided by SHARE.

Table. Charlson index construction.

Assigned weight for disease	Conditions
1	Heart attack / Heart trouble or angina, chest pain during exercise
	High blood pressure or hypertension
	High blood cholesterol
	A stroke or cerebral vascular disease
	Diabetes or high blood sugar
	Chronic lung disease
	Asthma
	Arthritis, including osteoarthritis, or rheumatism
	Osteoporosis
	Stomach or duodenal ulcer, peptic ulcer
	Parkinson disease
	Peripheral vascular disease
2	Hemiplegia
	Kidney cancer
	Any other tumor
3	Liver cancer

Other benchmarks

All other benchmarks are existing indexes provided by SHARE (See http://www.share-project.org/fileadmin/pdf_documentation/SHARE_release_guide_6-0-0.pdf for more information).

Detailed descriptions for each benchmark is presented below:

ADL and IADL

Both the Activities of Daily Living (ADL) the Instrumental Activities of Daily Living (IADL) are constructed based on question PH049_ in SHARE questionnaire, which includes a total of 6 ADL and 7 IADL activities. Responses were coded as 0 or 1 for each activity and summed to give a total ADL / IADL score. Question PH049_ is copied below. The first six activities are ADL activities and activities seven to thirteen are IADL activities.

PH049_

Please look at card 10. Here are a few more everyday activities. Please tell me if you have any difficulty with these because of a physical, mental, emotional or memory problem. Again exclude any difficulties you expect to last less than three months. (Because of a health or memory problem, do you have difficulty doing any of the activities on card 10?)

1. Dressing, including putting on shoes and socks
2. Walking across a room
3. Bathing or showering
4. Eating, such as cutting up your food
5. Getting in or out of bed
6. Using the toilet, including getting up or down
7. Using a map to figure out how to get around in a strange place
8. Preparing a hot meal
9. Shopping for groceries
10. Making telephone calls 11. Taking medications
12. Doing work around the house or garden
13. Managing money, such as paying bills and keeping track of expenses

# of symptoms

Number of symptoms is constructed based on question PH010_ in SHARE questionnaire. It includes a total of 11 symptoms. Responses were coded as 0 or 1 for each symptom and summed to give a total score. Question PH010_ is copied below.

PH010_

Please look at card 7. For the past six months at least, have you been bothered by any of the health conditions on this card? Please tell me the number or numbers.

1. Pain in your back, knees, hips or any other joint
2. Heart trouble or angina, chest pain during exercise
3. Breathlessness, difficulty breathing
4. Persistent cough
5. Swollen legs
6. Sleeping problems
7. Falling down
8. Fear of falling down
9. Dizziness, faints or blackouts
10. Stomach or intestine problems, including constipation, air, diarrhoea
11. Incontinence or involuntary loss of urine

Mobility limitations

Variable indicating mobility limitations is based on question PH048_ in SHARE questionnaire. It includes a total of 10 activities, each asking whether respondents have any difficulty doing the particular activity. Responses were coded as 0 or 1 for each activity and summed to give a total score. Question PH048_ is copied below.

PH048_

Please look at card 9. We need to understand difficulties people may have with various activities because of a health or physical problem. Please tell me whether you have any difficulty doing each of the everyday activities on card 9. Exclude any difficulties that you expect to last less than three months. (Because of a health problem, do you have difficulty doing any of the activities on this card?)

1. Walking 100 metres
2. Sitting for about two hours
3. Getting up from a chair after sitting for long periods
4. Climbing several flights of stairs without resting
5. Climbing one flight of stairs without resting
6. Stooping, kneeling, or crouching
7. Reaching or extending your arms above shoulder level
8. Pulling or pushing large objects like a living room chair
9. Lifting or carrying weights over 10 pounds/5 kilos, like a heavy bag of groceries
10. Picking up a small coin from a table

EURO-D

EURO-D scale is constructed based on the following 12 questions in SHARE. Responses were coded as 0 or 1 for each depressive symptom and summed to give a total score.

Question 1: SAD OR DEPRESSED LAST MONTH

'In the last month, have you been sad or depressed?'

0 No

1 Yes

Question 2: HOPES FOR THE FUTURE

'What are your hopes for the future?'

0 Any hopes mentioned

1 No hopes mentioned

Question 3: FELT WOULD RATHER BE DEAD

'In the last month, have you felt that you would rather be dead?'

0 No such feelings

1 Any mention of suicidal feelings or wishing to be dead

Question 4: FEELS GUILTY

'Do you tend to blame yourself or feel guilty about anything?'

0 No such feelings

1 Obvious excessive guilt or self-blame, mentions guilt or self-blame, but it is unclear if these constitute obvious, or excessive guilt or self-blame

Question 5: TROUBLE SLEEPING

'Have you had trouble sleeping recently?'

0 No trouble sleeping

1 Trouble with sleep or recent change in pattern

Question 6: LESS OR SAME INTEREST IN THINGS

'In the last month, what is your interest in things?'

0 No mention of loss of interest, non-specific or uncodeable response

1 Less interest than usual mentioned

Question 7: IRRITABILITY

'Have you been irritable recently?'

0 No

1 Yes

Question 8: APPETITE

'What has your appetite been like?'

0 No diminution in desire for food, non-specific or uncodeable response

1 Diminution in desire for food

Question 9: FATIGUE

'In the last month, have you had too little energy to do the things you wanted to do?'

0 No

1 Yes

Question 10: CONCENTRATION

'How is your concentration?' (Difficulty in concentrating on entertainment or reading)

1 Difficulty in concentrating on entertainment

2 No such difficulty mentioned

Question 11: ENJOYMENT

'What have you enjoyed doing recently?'

0 Mentions any enjoyment from activity

1 Fails to mention any enjoyable activity

Question 12: TEARFULNESS

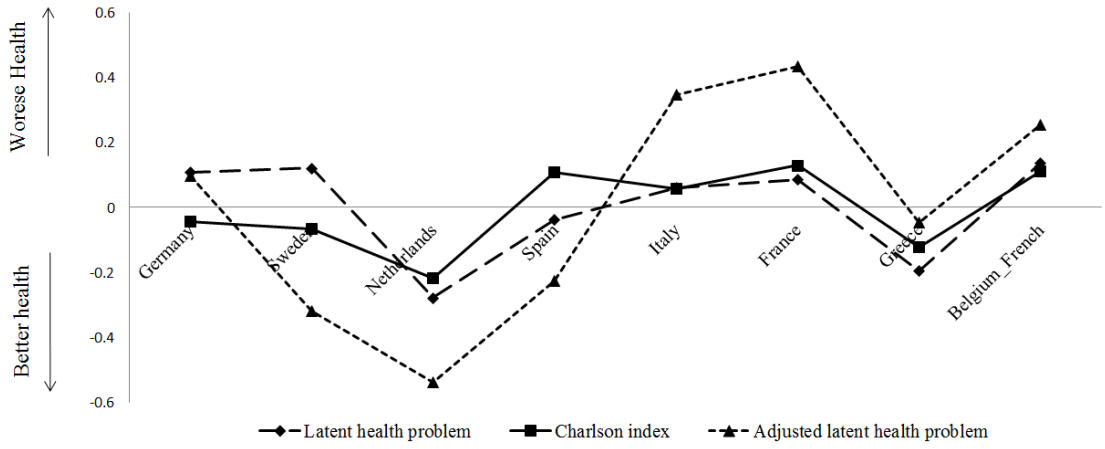
'In the last month, have you cried at all?'

0 No

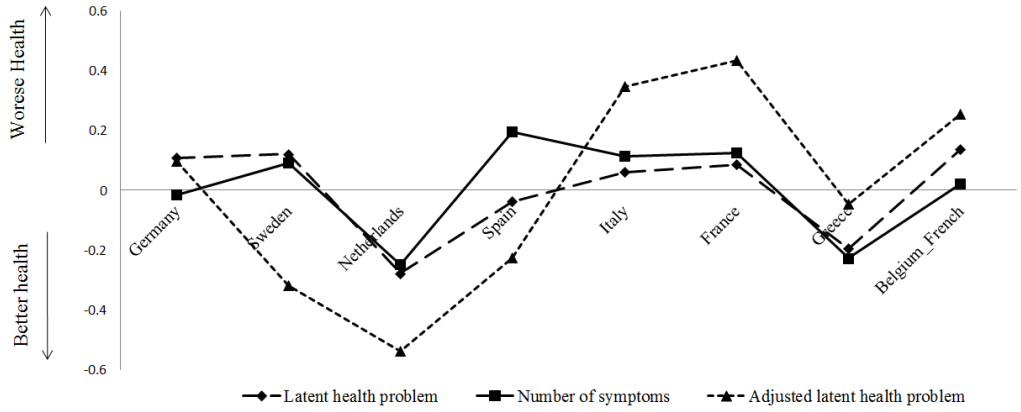
1 Yes

Appendix 2. 3 Latent health across countries before and after adjustments

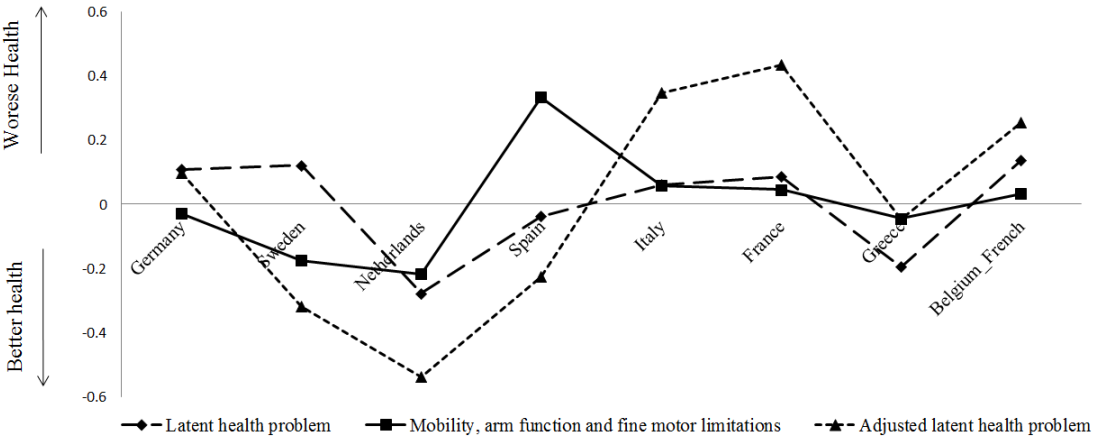
A



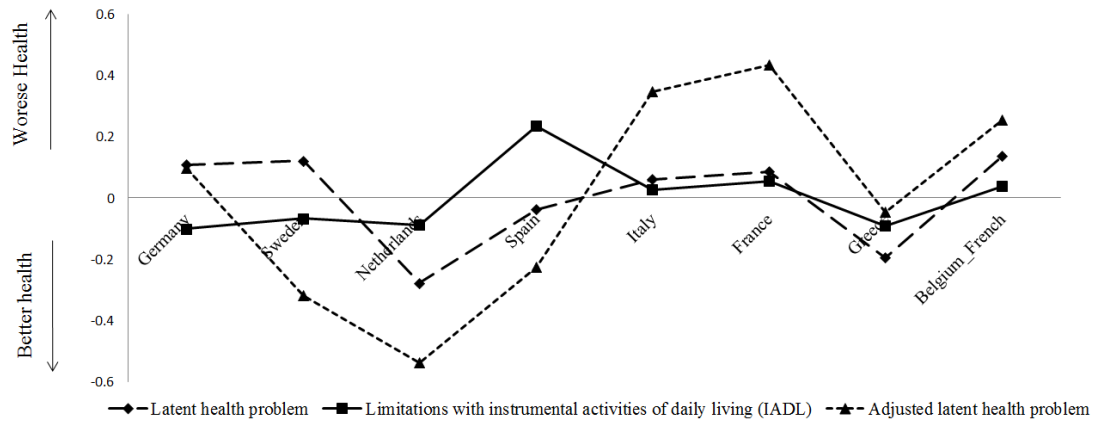
B



C



D



E

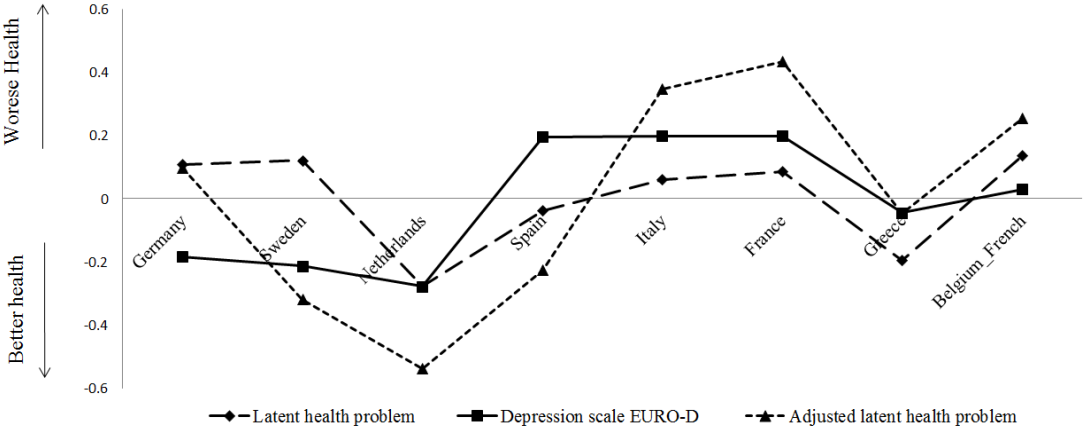


Figure 2. 4 Comparison of latent health across countries before and after adjustments using anchoring vignettes in relation to Charlson index (A), number of symptoms (B), mobility limitations (C), limitations with instrumental activities of daily living (D) and depression scale EURO-D (E).

CHAPTER III

The Use of Visual Anchoring Vignettes as an Alternative to Verbal Vignettes

3.1 Introduction

One major source of measurement error in self-assessment questions is response-category differential item functioning (DIF) (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007) that arises due to differences in respondents' usage of response scales. Anchoring vignette is a method developed to correct for DIF in interpersonal and cross-cultural comparisons. This method has been used to on various topics (e.g., health, political efficacy and life satisfaction). It also has been included in a wide range of surveys, such as the World Health Surveys (WHS), the Study on Global Ageing and Adult Health (SAGE), the Health Retirement Survey (HRS) and the Survey of Health, Ageing and Retirement in Europe (SHARE).

The anchoring vignette method starts from the premises that 1) for a given domain, there is a true and unobservable state for each person that lies on a continuum of the domain; 2) in data collection, respondents use unobservable cutpoints that enable them to report on an ordinal response scale; and 3) where these unobserved cutpoints are located interacts with respondents' cultural backgrounds systematically, as seen in Figure 1.1. The goal of the anchoring vignette method is to investigate the locations of the cutpoints by asking respondents to rate their own states and then the vignette persons' state. These evaluations allow researchers to compare where respondent's self-assessment stands relative to their assessments of the vignette persons. The

comparison results reveal where respondents' response anchoring points lie on the true state continuum, which, in turn, enables correction for DIF.

The successful use of anchoring vignettes depends on their ability to meet two key measurement assumptions. One is response consistency (RC), which means that respondents rate vignette persons in the same way as they rate themselves. The other one is vignette equivalence (VE), which means that the situations posed in vignettes are perceived similarly across respondents. As indicated in previous literature (e.g., King et al., 2004), anchoring vignettes should be designed in a way to achieve both RC and VE. To produce such successful vignettes, King et al. (2004) suggest that 1) the introduction explicitly state that, respondents are to rate the vignette questions as they would for themselves in the self-assessment questions; and that 2) the anchoring vignette questions be designed to provide the same stimuli across respondents. Other important design characteristics include the number and descriptions of anchoring vignette questions. To be more specific, usually more than one vignette is used to correct for one self-assessment question. The vignette questions describe different intensity levels of the measured construct – low, medium and high for a three-vignette example (e.g., for pain, see Appendix 3.1 for an example).

3.1.1 Promises and Pitfalls of Current Anchoring Vignette Approach

Anchoring vignettes have been reported in many studies as a promising tool to correct for DIF that is due to respondents' differential response scale usage (e.g., Mojtabai, 2015; Murray, Tandon, Salomon, Mathers, & Sadana, 2002). Despite this promise, studies of the effectiveness of the standard anchoring vignettes (which rely on the verbal descriptions of vignette persons) have yielded mixed results. While some studies have found that verbal vignettes can effectively correct for DIF (e.g., Van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011), other studies, have

reported that verbal vignettes do not necessarily provide comparable results among population groups (e.g., Grol-Prokopczyk, Verdes-Tennant, McEniry, & Ispány, 2015). In addition to the equivocal results on the validity of this approach, previous studies have also shown that, although verbal vignettes are designed to achieve RC and VE, these two assumptions can be violated in a real population (Bolt et al. 2014; Ferrer-i-Carbonell et al. 2011; Kapteyn et al. 2011; Rice et al. 2012).

The assumption violations are likely due to several critical practical challenges related to the anchoring vignette design. The first and most obvious challenge concerns the question difficulty (Hopkins & King 2010). Unlike typical questions on surveys that focus on respondents, anchoring vignette questions require respondents to imagine a set of hypothetical persons based on verbal descriptions and shift their focus from themselves to these imagined hypothetical persons. The second challenge is that vignettes describe hypothetical situations, and hence result in questions that contain much more text than self-assessment questions and other typical survey questions. This leads to a substantive increase in survey time (Hu and Lee, 2016). Given that usually more than one vignette is asked per domain (e.g., pain), vignettes in combination require a non-trivial amount of response time and cognitive ability (Hirve et al. 2013; Hopkins & King 2010; King et al. 2004).

The third challenge centers on what to include in the vignette descriptions. As acknowledged by Kapteyn et al., (2011), it is difficult to make the vignette descriptions as comprehensive as respondents' knowledge about their own state, which means that respondents may rate themselves using criteria different from those they use for vignettes (RC violation) (Kapteyn et al., 2011). Moreover, if respondents include information other than what is described in the vignettes in their understanding of the vignette scenario, VE is likely to be violated. These

issues have been identified in cognitive interviews by Hu and Lee (2013), which revealed that respondents do use their own experience and imagination about the hypothetical person beyond the given description, causing their perceptions of the vignette person to vary across respondents. The potential for these problems is even greater in cross-cultural research, for which it is even more challenging to design equivalent and comparable vignettes.

Fourth, the use of anchoring vignettes in cross-cultural research raises yet another issue with verbal vignettes, namely questionnaire translation. For example, the description of a pain vignette used in HRS states that "[Vignette person] has pain that …. In the last 30 days, how much pain or bodily aches did [vignette person] have?" However, when this was translated into Chinese in the China Health and Retirement Longitudinal Study (CHARLS), a 30-day frame was added to the description rather than the question as "In last 30 days, [vignette person] has pain that… How much pain or bodily aches did [vignette person] have". This translation directly influenced the description of the vignette person – i.e., the one without a 30-day qualifier was viewed more severe, since it is considered chronic, while the one with the qualifier was not. As Hu and Lee (2013) found, this contributes to the VE violation.

Although previous literature has greatly emphasized the importance of the design and pretesting of verbal anchoring vignettes, no clear design guidelines have been established to address the above limitations and practical challenges.

3.1.2 Visual Anchoring Vignettes

To remedy the aforementioned limitations of verbal vignettes, it is proposed in this study to use visual vignettes with well-designed and carefully-selected images. This strategy has several potential advantages. First, visual images may require less cognitive effort to process

their meaning than verbal descriptions do. According to previous literature, compared to texts, images are processed in a quicker and more automatic way, allowing respondents to form more "direct" connections between images and their meaning (Luna & Peracchio 2003; Paivio 2013; Townsend & Kahn 2014). In the case of anchoring vignettes (which requires imagining hypothetical persons), the use of visual images is advantageous for both low-literacy respondents and those who are unable to create mental images based on verbal vignettes. For these respondents, the saying "A picture is worth a thousand words" is particularly relevant considering the challenge of maneuvering through the lengthy verbal descriptions of understanding the vignette scenario (Hibbing & Rankin-Erickson 2003).

In addition to the ease of understanding, since respondents can process image information all at once for visual vignettes, we can reasonably expect that the use of visual vignettes can not only reduce respondents' cognitive burden but also survey time. Both of these can also further help with survey data quality, including reduction of break-offs in surveys and respondents' satisficing behavior.

A second advantage of visual vignettes is that they could potentially help with assumption fulfillments. As mentioned earlier, one challenge of designing verbal vignettes is that respondents may interpret the vignette questions in different ways (violation of VE). For example, it has been found that first names used in verbal vignettes (e.g., "Alice falls asleep easily at night…") can lead to respondents' inferences about that person's characteristics, such as age, gender and racial/ethnic information (e.g., Jürges & Winter, 2013). If respondents from different groups perceive the vignette person as having different characteristics, VE is likely to be violated. This is less a concern in visual vignette designs where the physical characteristics of

the vignette person are clearly presented, limiting the possibility of different interpretations of the vignette person.

However, one may argue that the use of an image in a survey question could influence or even bias survey responses. For example, Couper and his colleagues (2007) examined the effect that pictures of a healthy woman exercising versus a sick woman in a hospital bed have on self-rated health. They found that respondents consistently rate their own health lower when exposed to a picture of a fit woman, than when exposed to a picture of a sick woman (Couper et al. 2007). As explained by the authors, this is due to contrast effects, where respondents use the image as a standard to which they compare their own situations. Another reported context effect with image presentations is assimilation effects, which occur when the judgment of self-assessment question becomes more like the judgment of the image stimulus (Couper et al. 2007; Manis et al. 1991). The nature of visual anchoring vignettes, however, is different from those experiments with regard to contrast and assimilation effects, in that respondents provide a rating for the visual vignette persons, instead of rating for themselves. Rather than contrast or assimilation effects on self-assessment questions, anchoring vignettes are concerned more about whether respondents are able to differentiate the intensity level of each vignette (e.g., from least to most pain). Of course, we cannot completely rule out the possibility that respondents may use their own situation as a standard for rating vignette persons. However, we anticipate that respondents would take less time to answer visual vignettes than verbal vignettes, which gives them less opportunity to link their own situation with the rating for visual vignettes. Accordingly, we can reasonably assume that such context effects related to visual vignettes would be minimal and most likely would be less than those related with verbal vignettes. Thus, the issue of context effects should not apply directly to visual anchoring vignettes.

Since there are no prior studies on the use of visual anchoring vignettes, it remains an open question whether this approach can remedy some of the limitations of the current verbal vignettes and effectively correct for DIF. To fill this gap, this research aims to 1) evaluate the use of visual anchoring vignettes as an alternative to verbal vignettes and 2) compare the performance of visual and standard verbal vignettes in terms of response time, respondents' cognitive burden and how well they meet the RC and VE. Given that vignette persons' characteristics can potentially affect the use of anchoring vignette method, to better understand potential sources of variations in the visual vignette performances, a secondary objective of this study is to examine the effect of different vignette persons' characteristics (e.g., male vs. female randomization) on the performance of visual vignettes in terms of how well they correct for DIF that is due to respondents' differential response scale usage.

3.2 Method

The method section contains four parts. The first describes the design of visual vignettes. The second introduces a pretest to evaluate selected visual vignette candidates and to select three well-designed visual vignettes per domain. The third involves a web survey experiment which collects data from various racial / ethnic groups (i.e., non-Hispanic (NH) white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic). The fourth describes the analysis strategies. This research focused on four health domains, sleep, affect, mobility and pain, which are known to be subject to DIF (e.g., d'Uva, O'Donnell, & van Doorslaer, 2008).

3.2.1 Design of visual vignettes

Prior to the design of visual vignettes, I evaluated what to include in the visual image, and determined the criteria for image selection or creation. A three-step approach was used to

develop these criteria: specifically, I 1) thoroughly examined critical elements of the four health domains, 2) identified common elements applicable across groups (e.g., arm pain) based on literature reviews, and 3) based on the elements identified, I selected or designed images with these elements at different intensity level for each domain (e.g., from no pain to extreme pain for pain domain). The first two steps of finding common critical elements are important to meet the VE assumption because respondents with different cultural backgrounds may use different elements to evaluate their health, and they may weigh the same elements differently (Hu & Lee, 2013).

Based on the developed criteria, images that matched these criteria were then selected from commercial websites of images and photos (e.g., www.istockphoto.com/). In situations where no criteria-meeting images for a health domain were found on those websites, I 1) recruited eligible volunteers from different platforms (e.g., own friends or family networks) to serve as models in the photos, 2) obtained their consent to take photo and to use it in our study, and 3) took the photo and edited for use in the visual vignettes. To remove potential confounding effects of various image elements, such as the background, size, resolution and color balance, the selected images or photos were further edited by students with expertise in image-editing at the University of Michigan.

To evaluate the relevance of domain-specific vignette protagonists' characteristics on the performance of visual vignettes, a set of two visual vignette conditions with different characteristics design (e.g., male vs. female; young vs. old randomization) were developed and further tested in the analysis. The ultimate goal of visual vignette design for the current study was to have three well-designed visual vignettes per design condition per domain (e.g., three male pain vignettes and three female pain vignettes). For the purpose of selecting the most

comparable images across cultures, I first designed about six images (two images per intensity level – e.g., two no/low pain, two middle pain and two extreme pain vignettes) per characteristic design condition per domain (resulting in about 50 images in total), and eventually selected three out of six for each condition in the pretest. The selected images (see Appendix 3.1) were then used in the web survey experiment as described below. The vignette designs for each domain and reasons for the design choices are described below.

Sleep. For sleep, I used male vs. female visual vignettes designs. Other vignette characteristics, including race, age and fitness level were controlled to be the same across the two designs. I chose male and female vignette designs for sleep because research has demonstrated that women and men differ in their sleep patterns, and risks and reasons for sleep disorders (Jean-Louis et al. 2000). Male's and female's self-assessed sleep levels were found to correlate differently with objective measures (Hoch et al. 1987), suggesting that reporting bias may differ between the two groups. It is also reported in previous literature that although women have better sleep quality than men, they have more complaints about their sleep than men (Vitiello et al. 2004; Krishnan & Collop 2006). Despite the gender differences in sleep quality perceptions and actual sleep patterns, it remains an open question whether male and female respondents would rate for male vs. female visual vignettes differently, and how these potential differences can affect the performance of visual vignette methodology.

Mobility. For mobility, I varied the vignette persons' body size (i.e., obese vs. optimal weight) in the two visual vignettes design conditions. Similar to other domains, other vignette characteristics, including race, age and gender were controlled to be the same across the two designs. The choice of body size as a design factor for mobility is made for the following two reasons. First, research has demonstrated that obesity is associated with mobility disabilities

(Forhan & Gill 2013; Stenholm et al. 2009). Second, similar to other visible characteristics such as gender, race and age, body size is another characteristic which can impact the perceptions of vignettes(Trautner et al. 2013; Penny & Haddock 2007). Given these two reasons, it is worth considering the impact of body size on respondents' ratings of vignette persons' mobility, and evaluating whether this influences the performance of visual vignette method in terms of how well it corrects for DIF.

Affect. For affect, in one visual vignette design, I matched respondents' race with the vignette persons' race. In the other design condition, respondents were randomly assigned to vignettes of other races. Similar to the other domains, other vignette characteristics, including age, gender and body size were controlled to be the same across the two designs. Racial/ethnic expectations may allow prejudicial judgment to emerge (Martinez 2012), which may influence respondents' ratings for vignette persons of the same or different races. Studies on emotion and affect recogniation across cultural groups suggest that a cultural advantages may take place in the process of recognizing other people's emotions (Anderson & Keltner 2002). This has been described by the cultural advantage model, indicating that people process facial emotion experssions of same-race individuals more accurately and efficiently than those of other race groups (Soto & Levenson 2009; O'Toole et al. 1996). Given the possibility of cultrual advantage, experimenting with the effect of race-matching for this domain can reveal whether the cultural advantage model applies to visual anchoring vingettes, and how it influences the performances of visual vignette methods.

Pain. For pain, I varied the vignette persons' age (i.e., older adult vs. young adult) in the two visual vignettes design conditions. Similar to the other domains, other vignette characteristics, including gender, race and body size were controlled to be the same across the

two designs. Age was chosen as a design factor for pain for the following two reasons. First, physical pain is more prevalent and significant in older adults (Herr 2002). Despite the prevalence, older adults are more likely to under-report their pain. Second, pain assessment of vignette persons are likely to differ by the vignette persons' age. One study has evaluated the vignette person's age influences on pain assessment using virtual human (VH) images, and found that older VH were viewed as having worse pain, worse coping and a greater need for medical treatment than younger VH (Hirsh et al. 2008). Given the likely difference in respondents' ratings between older and younger visual vignette persons, it is worthwhile to investigate how these potential differences may affect the use of visual vignettes.

Ideally, I could have experimented and randomized all four characteristic (i.e., gender, race, age and body size) for each domain. However, this is not practical given the large sample size needed and the limited budget constrain. As a starting point, for each domain, I decided to select one potentially most-influential characteristic and keep the other three characteristics constant. Future studies can expand the experiment design and evaluate other untested characteristics for each domain.

Also we note that visual vignettes do not necessarily need to be designed in such a way to reveal the same meanings as existing verbal vignettes. Instead, similar to the intention of verbal vignette designs, they need to be designed so that they can better meet RC and VE, in order to adjust for DIF.

3.2.2 Pretesting

To conduct the pretest, I used Amazon Mechanical Turk (MTurk). On MTurk, I posted the survey announcement, also known as Amazon's human intelligence tasks (HITs). Eligible respondents can browse the HITs and decide if they would like to take the survey or not. The

announcement contains a link to the pretest survey, which was programmed with Qualtrics. The pretest was open to U.S. workers who are 18 or older. A $0.45 incentive was offered to each complete. To recruit respondents of all age groups, toward the end of the data collection, I posted a HIT open only to older respondents with the same incentive. In total, 201 respondents completed the pretest survey, with about half being 50 years or older. The main criteria applied to evaluate and select proper images was based on whether respondents could correctly rank order vignettes as expected. This method was first used by WHO in their pretesting of anchoring vignettes (Murray et al. 2003:376). For the two sets of image options, the image with the higher correct ranking rate (i.e., the percentage of respondents who can correctly rank the vignette series) was selected. The final correct ranking rates ranged from about 80% to 97% for all health domains.

3.2.3 Web survey

The main data collection was based on a web survey using non-probability online panel. Respondents from the four different racial/ethnic groups – NH white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic – were recruited through Qualtrics online survey panel. Respondents from each racial/ethnic group were randomized in to three conditions – the standard verbal vignette condition and two visual vignette conditions different in vignette person's characteristics as described in *design of visual vignettes* section.

3.2.3.1 Measures

Besides self-assessment and vignette questions, this study also included measures on objective questions regarding these health domains, health behavior questions, and demographic and socio-economic information.

Vignette question designs. For the verbal condition, I adopted the verbal vignette descriptions from those widely used in many major surveys (e.g., HRS). Each domain has a series of three vignettes, describing different intensity levels of the measured construct – low, medium and high (e.g., from least to most pain). For the visual condition, I used the visual vignettes designed and selected in the pretest with three vignettes per condition, depicting three levels of difficulty/intensity of symptoms in each domain. Introduction to the vignette questions also followed the standard approach used in earlier surveys like HRS. I experimentally randomized the order of the domains and the order of three vignettes per domain presented to respondents, so as to isolate question order effects.

Objective questions. Objective questions on each of the four health domains were collected. These questions were needed to evaluate the validity of both verbal and visual vignette methods. For example, objective questions for sleep included how long respondents sleep every night, how easily respondents can wake up or get back to sleep at night and whether respondents are taking medications for sleep issues or not. See Appendix 3.2 for question details.

Health behavior questions. Health behavior questions such as those targeting eating and drinking habits, physical activities were included in the web survey questionnaire.

Respondents' characteristics. Variables including respondents' age, gender, education, marital status, income, cultural identification, language preference, height and weight were collected.

Paradata. To evaluate whether the use of visual vignettes can help to reduce survey time, paradata such as time stamps were also collected.

3.2.3.2 Translation of survey instruments

In translating the instrument into Spanish for Spanish-speaking Hispanics, this study followed the set of best practices developed by the United States Census Bureau (Pan & De La Puente, 2005) and Cross-Cultural Survey Guidelines developed by the survey research center at the University of Michigan (Mohler, Dorer, de Jong & Hu, 2016). Translation was conducted by the HRS translation team at the University of Michigan. The translated questionnaire was then reviewed and tested by 20 bilingual speakers who are fluent in English and are native Spanish speakers.

3.2.3.3 Data collection

The online survey questionnaire was programmed in Qualtrics. The Qualtrics online panel team sampled respondents through their panel. Except for Hispanics speaking Spanish, about 750 respondents were sampled for each of the three other race /ethnic groups. Each of the three sampled subgroup had about equal proportions of 1) male and female, 2) below or equal to high school education and higher than high school education, and 3) 18 - 49 and over 49 year-old respondents. For Spanish-speaking Hispanics[6], 889 respondents were sampled with about 43% male respondents. Detailed information of the sample profile is presented in Table 3.1.

Email invitations were sent to selected respondents, with the link to the survey included in the email. Respondents from each racial / ethnic group were randomly assigned to one of the three vignette type conditions - one verbal condition and two visual conditions.

[6] Due to the difficulties to recruit Spanish-speaking Hispnics, Qualtrics ended up collecting more respondents for this group in order to meet the targeted number for male Spanish-speaking Hispnics who are 50 and above and have high school or below education.

Table 3. 1 Respondents' characteristics.

	White %	Black %	Hispanic – English %	Hispanic - Spanish %
n	760	750	750	889
Male	50.39	50	50	42.52
Age				
Age 18 - 29	14.34	22.8	22.13	21.37
Age 30 - 49	33.68	25.73	26.53	35.77
Age 50 - 64	30.13	36.27	34.53	33.52
Age 65 and above	21.84	15.2	16.8	9.34
Higher than high school education	49.47	50	50	57.82
Married	53.42	36.67	50.93	54.78
Employed	50.92	52	56.13	57.14
Income				
Income below $40,000	35	35.87	33.07	34.76
Income between $40,000 - $69,999	33.95	42.93	41.33	45.67
Income $70,000 or more	31.05	21.2	25.6	19.57

3.2.4 Analysis Strategy

I first examined the distributions of the self-assessment question and vignette questions by vignette type and designs for each domain. I then analyzed survey time using time stamp data. The mean response time for verbal vignette questions and visual vignette questions are compared. For each of the time variables, I trimmed the high response-time observations downwards – for the observations which exceed the 99^{th} percentile of the time variable, their values were replaced as the 99^{th} percentile time of this variable.

I then checked whether respondents could correctly rank order vignettes based on their intensity level. Several previous studies refer to this analysis as a weak test for VE, stating that correct rank-ordering is a "necessary but not sufficient" condition for vignette equivalence (e.g., Grol-Prokopczyk et al. 2015; Kristensen & Johansson 2008). It is true that if VE is held, with

good vignette designs, respondents should be able to correctly rank the vignettes. However, this is not always the case. For example, if two vignettes cannot be easily distinguished from each other, even if VE is held, it may be possible that most respondents have mistakenly ranked the order. Thus, the correct rank ordering is more a test for the quality of the vignette design, which should be performed before testing VE and RC.

It is possible that respondents may rate two or three vignettes identically. For example, if a respondent has a very high threshold for what is "mild" pain, he / she may rate the first two vignettes (low and middle pain) or all vignettes as no pain. This is referred to as "ties" in vignette-ratings. Although it is possible that a respondent may have *true* ties for all three vignettes (i.e., view the three vignettes with similar intensity level and rate them identically), this is actually quite unlikely given the differences of the intensity levels in the vignette design. Identical ratings for all three vignettes are more likely due to satisficing behavior in reporting, where respondents may choose the same categories to reduce their cognitive burden. Thus, here I only consider two situations of the ties – 1) ties between the first two vignettes (low and middle intensity) and 2) ties between the last two vignettes (middle and high intensity).

3.2.4.1 Test of VE

The test of VE was conducted following Grol-Prokopczyk & Carr (2017). This method was first developed by Bago d'Uva et al. (2011), and applied in several other studies (Grol-Prokopczyk et al. 2015; Grol-Prokopczyk & Carr 2017; Molina 2016). The rationale behind this test is that if respondents view each vignette in the same way (VE), the distance between any two vignettes on the latent spectrum should be the same for all respondents (Bago d'Uva et al. 2011). The test is based on likelihood-ratio test (LRT) of two models. Both models, which are variations of the hierarchical ordered probit (HOPIT) model, include only the vignette component of the

HOPIT model. More information on the HOPIT model can be found in Appendix 3.3. Below I list the key differences between the two models. The first model, Model A[7], is as below:

$$V_{ij}{}^{*} = \alpha_j + \varepsilon_{ij} \qquad (A)$$

where $V_{ij}{}^{*}$ is respondent i's perceived location of vignette j, α_j is a constant term and ε_{ij} is the random error term. ε_{ij} is assumed to be normally distributed with a mean of zero and a variance of 1. α for the reference vignette is set to be 0 for model identification. Note that Model A does not include covariates to predict the perceived vignette locations. This is consistent with VE, namely that respondents' perceptions of vignettes do not depend on their background and are constant across different population groups.

In Model B, a vector of covariates, X_i, is added to the model to predict the perceived vignette locations. In this particular study, X_i include marital status, employment status, age, gender, education, income level and racial / ethnic groups. λ_j is the estimated coefficients for each non-referent vignette. Similar to Model A, the α for the reference vignette is set to 0.

$$V_{ij}{}^{*} = \alpha_j + \lambda_j X_i + \varepsilon_{ij} \qquad (B)$$

If VE is fulfilled, λ_j will be 0 for each j, and the LR test will indicate no differences between the two models. If, however, a significant LR test statistic appears, it indicates that respondents from different groups perceive the severity of the vignettes differently. In other words, VE is violated. The estimated coefficient λ_j and the significance of each covariate will inform us which covariates are driving the violations.

3.2.4.2 Test of RC

The test of RC was conducted following Grol-Prokopczyk et al. (2015) and Bago d'Uva et al. (2011). This less stringent test of RC was based on visual comparisons of two sets of

[7] In describing the models, I used the same notation as Grol-Prokopczyk & Carr (2017).

cutpoints. One set was generated from vignettes only, based on Model A as in the tests of VE. The other set was generated from self-assessments, where I included objective health measures to predict the self-assments (Model C). The cutpoints from the two models were then graphed in a figure for visual comparisons. The RC test basically compared the "shape" (Grol-Prokopczyk et al. 2015) of the two sets of cutpoints. A similar cutpoint "shape" would indicate that respondents had similar standards when rating vignettes and rating themselves (RC). As mentioned in Grol-Prokopczyk and Carr (2017), this test can be viewed only as suggestive – since each set of the cutpoints was calculated on a different scale, the abosolute and relative positions of the cutpoints on the latent spectrum were unknown. Actual violations of RC may be larger than what appears in the "shape" comparisons. The objective measures used in this present study include: whether respondents have seen a doctor about their difficulties with sleep, whether respondents on average sleep less than 7 hours or over 9 hours each day, a sleep quality score[8], total pain index[9], number of mobility activities that respondents have difficulty with, number of chronic health conditions and the Kessler Psychological Distress Scale (K6) (Kessler et al. 2002).

3.3 Results

3.3.1 Descriptive analysis

As shown in Figures 3.1 to 3.4, I first examined the distributions of the self-assessment question and vignette questions by vignette type and designs for each domain. As expected for a randomized study, for each domain, the distributions for the self-assessment questions do not

[8] This sleep quality score is constructed based respondents' responses to three sleep questions, asking respectively whether and how often respondents 1) have trouble falling asleep, 2) wake up several times at night and 3) wake up earlier than planned at night and were unable to fall asleep again.
[9] This total pain index is constructed following Ray et al. (2009).

differ by vignette type or visual vignette designs. Comparing vignette distributions by vignette type, in general, the intensity levels of the visual vignettes can be better differentiated than those of the verbal vignettes. For the sleep (Figure 3.2) and mobility (Figure 3.3) verbal vignettes, it seems that most respondents fail to distinguish the intensity levels for the second (medium severity) and the third (most severity) vignettes.

For the sleep, pain and affect domains, the distributions for the two visual vignette designs are very similar. I thus only present *verbal* vs. *visual vignette regardless of design* for these three domains in Figure 3.1, 3.2 and 3.4. For the mobility domain (Figure 3.3), respondents assigned to the fit vignette design rate the vignettes with less difficulty than those assigned to the obese vignette design.

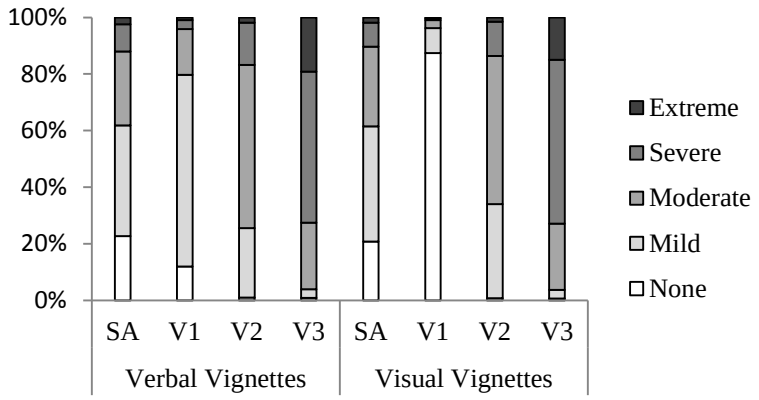


Figure 3. 1 Responses to pain self-assessment (SA) and the three vignettes difficulty/intensity questions (V1=none/mild; V2 = moderate; V3= severe/extreme) by vignette types.

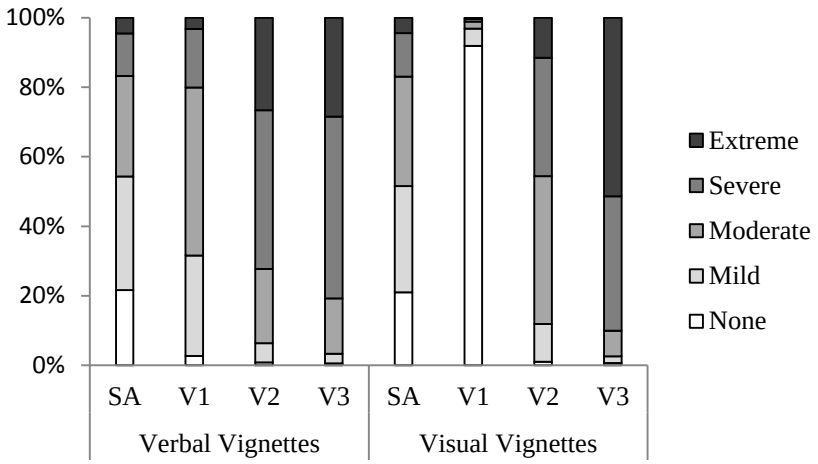


Figure 3. 2 Responses to sleep self-assessment (SA) and the three vignettes difficulty/intensity questions (V1=none/mild; V2 = moderate; V3= severe/extreme) by vignette types.

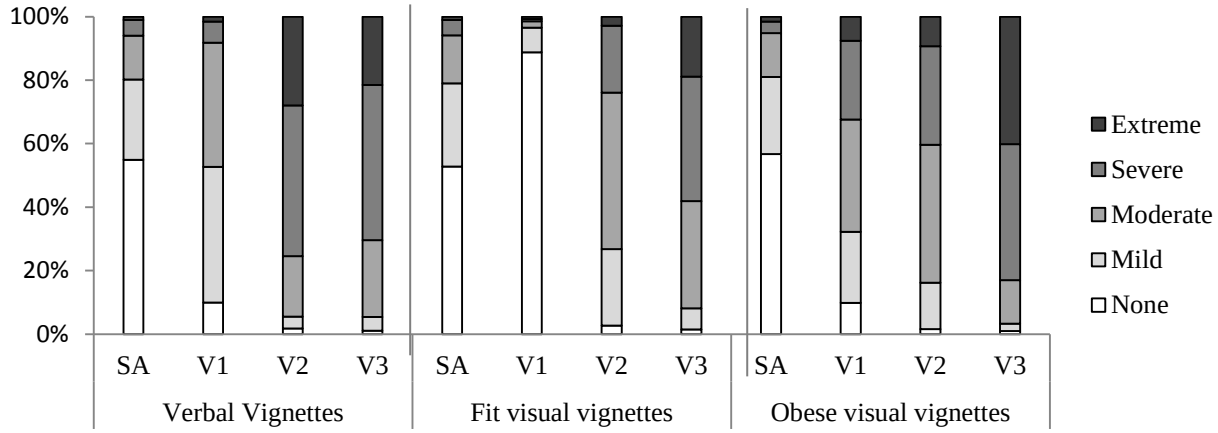


Figure 3. 3 Responses to mobility self-assessment (SA) and the three vignettes difficulty/intensity questions (V1=none/mild; V2 = moderate; V3= severe/extreme) by vignette types.

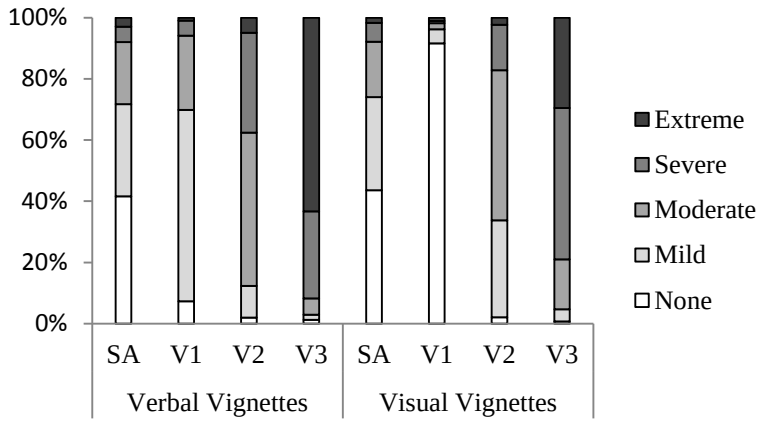


Figure 3. 4 Responses to affect self-assessment (SA) and the three vignettes difficulty/intensity questions (V1=none/mild; V2 = moderate; V3= severe/extreme) by vignette types.

3.3.2 Time analysis

As shown in Table 3.2, regardless of domain, the average time respondents spent on a verbal vignette question is much longer than (about two times) that of a visual vignette question. This is consistent with our expectation.

Table 3. 2 Average time (in seconds) spent on a verbal vignette and visual vignette question by health domain.

	Pain		Sleep		Mobility		Affect	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Verbal vignette	15.93	8.81	15.73	8.25	17.85	10.31	18.05	10.59
Visual vignette	7.95	3.58	8.38	3.61	8.33	3.79	7.42	3.17

3.3.3 Rank-ordering

Table 3.3 shows the percentage of respondents whose ratings for the vignettes are consistent with the expected order (i.e., low intensity to high intensity). The percentages ranged from 17% to about 90%, depending on the domain. It is noted that for each of the four domains, the percentage is significantly higher for the visual vignette conditions than for the verbal condition. In other words, respondents assigned to the visual conditions can better rank the order

of the vignettes than those assigned to the verbal condition. Consistent with what we see in Figures 3.2 and 3.3, respondents seem to have difficulty differentiating the rank-orders of sleep and mobility *verbal* vignettes, with only less than 20% able to correctly rank for these domains. Although significantly better than the verbal condition, only 25% respondents can correctly rank the vignettes for the mobility domain with obese visual vignette designs.

Table 3. 3 Percentage of respondents ordering vignettes consistently with expected ordering.

	Pain		Sleep		Mobility		Affect	
	n	%	n	%	n	%	n	%
Verbal vignette	1051	47.6	1051	17.7	1051	19.8	1051	67.1
Visual vignette design 1	1040	74.0	1068	69.1	1061	61.3	1043	87.9
Visual vignette design 2	1058	85.4	1030	79.1	1037	25.2	1055	77.5
Visual vignette regardless of design	2098	79.7	2098	74.0	2098	43.4	2098	81.8

Note: design 1 vs. design 2 of visual vignettes conditions for each domain are: pain – old (design 1) vs. young (design 2); sleep – female (design 1) vs. male (design 2); mobility – fit (design 1) vs. obese (design 2); affect – same race group (design 1) vs. different race groups (design 2).

Table 3.4 below shows the results when allowing the two tie situations – 1) ties between the first two vignettes (low and middle intensity) and 2) ties between the last two vignettes (middle and high intensity).

Table 3. 4 Percentage of respondents ordering vignettes consistently with the expected ordering (including ties between the first two and the last two vignettes).

	Pain		Sleep		Mobility		Affect	
	n	%	n	%	n	%	n	%
Verbal vignette	1051	85.8	1051	58.4	1051	58.5	1051	87.0
Visual vignette design 1	1040	95.2	1068	90.6	1061	84.1	1043	92.1
Visual vignette design 2	1058	96.8	1030	94.2	1037	69.1	1055	93.0
Visual vignette regardless of design	2098	96.0	2098	92.4	2098	76.6	2098	92.6

Note: design 1 vs. design 2 of visual vignettes conditions for each domain are: pain – old (design 1) vs. young (design 2); sleep – female (design 1) vs. male (design 2); mobility – fit (design 1) vs. obese (design 2); affect – same race group (design 1) vs. different race groups (design 2).

Not surprisingly, when allowing ties, the percentage of respondents whose ratings for the vignettes are consistent with the expected order (Table 3.4) are higher than the percentages for the no-tie situations (Table 3.3). Consistently with Table 3.3, respondents assigned to the visual conditions did a better job in recognizing the rank orders of the vignettes than those assigned to the verbal condition.

3.3.4 VE Testing Results

Table 3.5 presents the test results of VE, which shows that the VE assumption is violated in almost all conditions, except for the sleep verbal vignettes (detailed results see Appendix 3.4, Figure 3.7). Due to the space restrictions, this paper presents only the model results for predicting vignette locations (i.e., where it lies on the latent health spectrum) for the pain verbal vignette condition in Table 3.6.

Table 3. 5 Likelihood ratio tests of vignette equivalence.

	Pain		Sleep		Mobility		Affect	
	df	LR Test	df	LR Test	df	LR Test	df	LR Test
Verbal vignettes	24	70.4***	24	24.4	24	55.1***	24	110.9***
Visual vignette design 1	24	75.0***	24	99.5***	24	127.9***	24	95.6***
Visual vignette design 2	24	98.0***	24	97.7***	24	47.0**	24	90.8***
Visual vignette regardless of design	24	137.4***	24	158.8***	24	67.1***	24	154.3***

Note: design 1 vs. design 2 of visual vignettes conditions for each domain are: pain – old (design 1) vs. young (design 2); sleep – female (design 1) vs. male (design 2); mobility – fit (design 1) vs. obese (design 2); affect – same race group (design 1) vs. different race groups (design 2).
*, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$.

As shown in Table 3.6, the reference vignette is Vignette 3 – the vignette describing the highest pain level. Gender, marital status and racial / ethnic groups are the main predictors that drive the violations of VE for pain verbal vignettes. Those who are married view the first

vignette (the vignette with the least pain) as farther away from the reference vignette on the latent spectrum, with a positive coefficient of 0.31 (p-value = 0.02). In other words, those who are married view the first vignette as depicting better health (or less pain) than those who are not married. Males, compared to females, view Vignette 1 as depicting worse health (or more pain). Hispanics, in general, view the first vignette as depicting more pain compared to the White respondents.

Table 3. 6 Predictors for the perceived vignette locations on the latent health spectrum for pain verbal vignettes.

	Coefficient	SE	p-value
Vignette 1 (no/mild difficulty/intensity)			
Constant	3.20***	0.25	0.00
Married	0.31*	0.13	0.02
Male	-0.49***	0.13	0.00
Employed	-0.10	0.14	0.47
Education above high school	-0.09	0.13	0.49
Age 18 - 29	0.25	0.23	0.29
Age 30 - 49	-0.17	0.21	0.42
Age 50 - 64	0.09	0.20	0.66
Middle income	0.04	0.14	0.80
High income	-0.15	0.18	0.38
Black	-0.12	0.19	0.52
Hispanics (English)	-0.47**	0.19	0.01
Hispanics (Spanish)	-0.82***	0.18	0.00
Vignette 2 (moderate difficulty/intensity)			
Constant	1.38***	0.21	0.00
Married	0.09	0.11	0.42
Male	-0.11	0.11	0.34
Employed	-0.03	0.12	0.80
Education above high school	-0.12	0.11	0.29
Age 18 - 29	0.13	0.20	0.50
Age 30 - 49	-0.06	0.18	0.71
Age 50 - 64	0.01	0.17	0.96
Middle income	0.06	0.12	0.62
High income	-0.12	0.15	0.43
Black	0.27	0.16	0.09
Hispanics (English)	-0.13	0.15	0.39
Hispanics (Spanish)	0.02	0.15	0.89

Notes: Vignette 3 (highest difficulty / intensity) is the reference vignette. *, P≤0.05; **, P≤0.01; ***, P≤0.001.

The estimated vignette locations on the latent health spectrum by racial / ethnic groups are presented in Figure 3.5, including 3.5A for pain verbal vignettes and 3.5B for pain visual vignettes. Since the patterns for the two visual vignette conditions are similar, I present only 3.5B which includes data for both visual conditions. If VE is met, we would expect the estimated

pain vignette locations to be exactly the same for each racial / ethnic group. This is not the case, as can be seen from Table 3.6 and also shown in Figure 3.5. For both verbal and visual vignettes, Hispanics who completed the Spanish survey view the first vignette person as having more pain compared to the White respondents, while Hispanics who completed the English survey view the first vignette person also as having more pain than White under the verbal condition but not the visual vignette condition. Results for other domains are presented in Appendix 3.4.

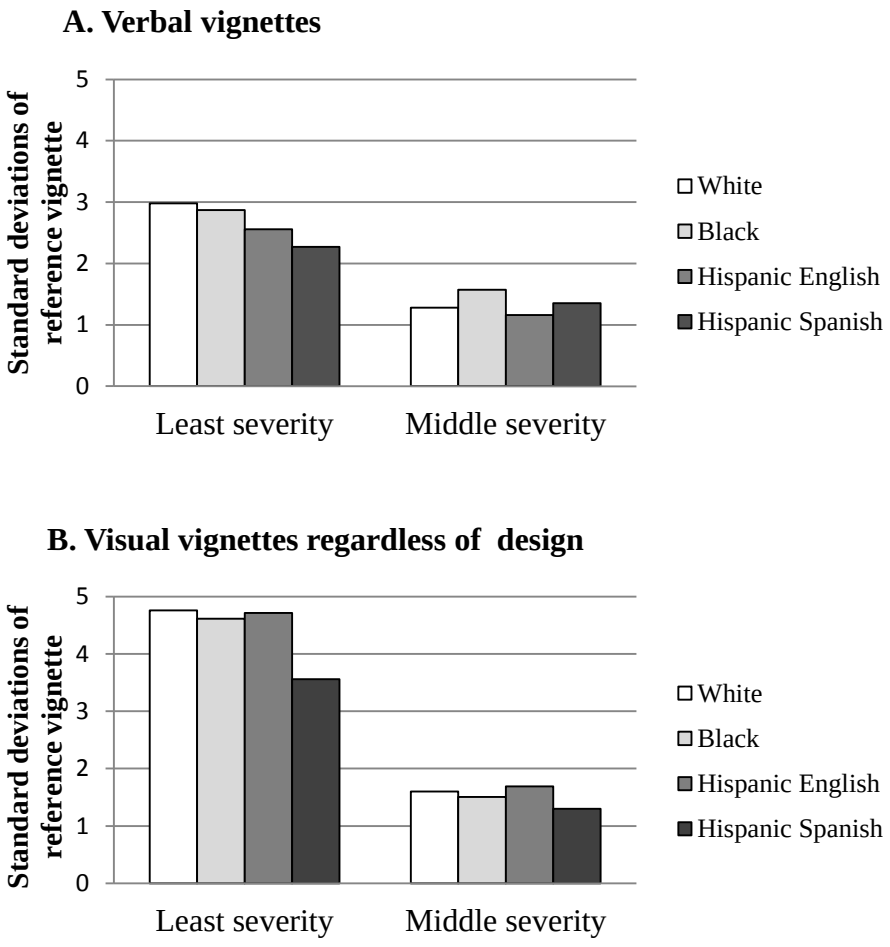
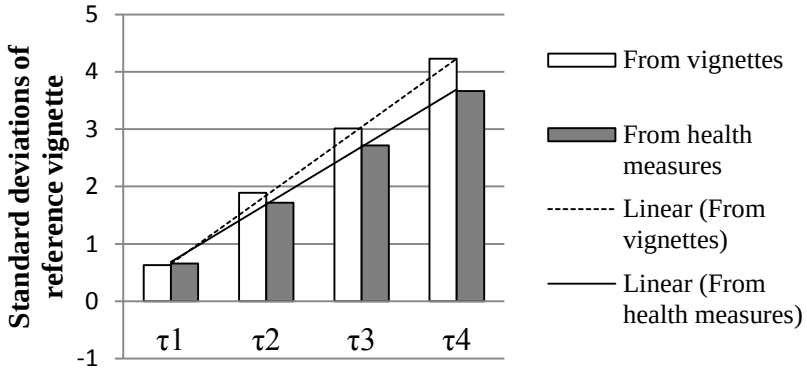


Figure 3. 5 Estimated pain vignette locations (on latent health spectrum; relative to Severity 3). Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

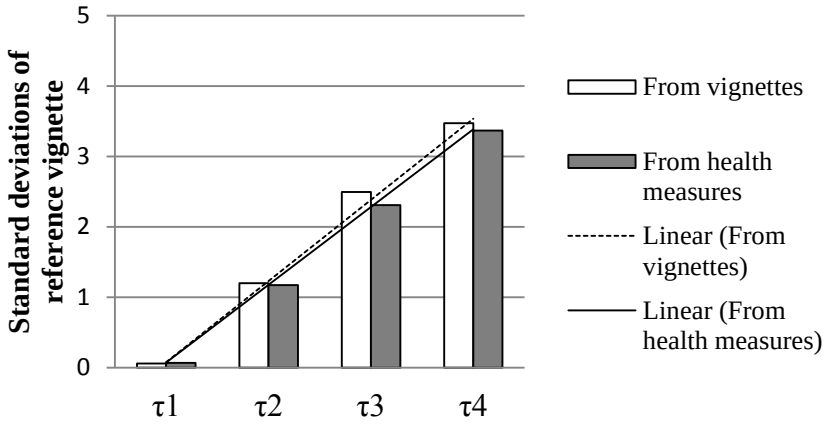
3.3.5 RC Testing Results

As described in the *Analysis* section, the RC assumption test is based on visual comparisons of two sets of cutpoints – one from Model A which has only vignettes, another from Model C which includes self-assessments and objective measures. Figure 3.6 shows the estimated cutpoints for mobility. As shown in Figure 3.6, for both verbal and visual vignette conditions, the vignette-derived cutoff point patterns are quite similar to the health measures-derived cutpoints, indicating minor violations of RC. Among the three vignette designs (Figure 3.6A – 3.6C), visual vignettes with fit vignette figures appear to adhere RC the most (the vignette-derived cut points closely align with SA- derived thresholds). Test results of RC for other domains can be found in Appendix 3.5.

A. Verbal vignettes



B. Visual vignettes with fit vignette figures



C. Visual vignettes with obese vignette figures

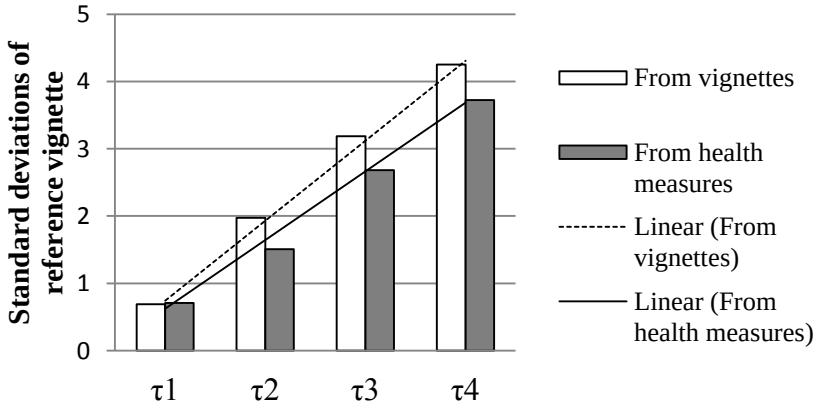


Figure 3. 6 Estimated cutpoints for mobility based on vignettes and health measures. τ1 – τ4 are cutpoints for the five-point response scale from "None" to "Extreme" (e.g., τ1 is the cutpoint between "None" and "Mild").

3.4 Discussion

Anchoring vignettes have become widely used in comparative studies. However, several practical challenges are associated with this method. In particular, vignette questions, given their complexity of descriptions, increase survey time and respondents' burden. Further, the fulfillment of the VE and RC assumptions behind this method is not always assured. To remedy the limitations of verbal vignettes, this paper proposes the use of visual vignettes, consisting of carefully designed and pre-tested images. This is the first study that examined the use of visual vignettes. In this study, the performances of both verbal and visual anchoring vignettes are compared using various approaches, including time and tests of assumptions. This research has important implications for future design and improvements of anchoring vignette method.

The results indicate that the use of visual vignettes can potentially reduce respondents' cognitive burden and greatly reduce survey time. Visual vignettes also outperform verbal vignettes in that respondents can better distinguish the different intensity levels in visual vignettes (e.g., from no pain to extreme pain) than verbal vignettes. For both verbal and visual vignettes, respondents' perceptions of the vignettes can differ by groups (VE violations). This implies that designing "universal" anchoring vignettes (Grol-Prokopczyk & Carr 2017) that reveal the same information to every respondent is still a challenge for both verbal and visual vignettes.

This study also revealed several important findings to deepen our understanding of vignette methodology. First, when rating for the two visual vignettes (i.e., fit vs. obese) for mobility domain, respondents tend to rate the obese vignette person (of the same difficulty level with the fit vignette person) as of more mobility difficulties. This is not surprising given that

those who are obese have higher risks of having mobility difficulties than those who are not obese (Koster et al. 2007). Although beyond the scope of this study, future research could look into how the perceptions of fit vs. obese vignettes may vary among respondents with different body sizes.

Second, the rank-ordering results for affect domain suggest that when respondents rate for vignettes of the same racial / ethnicity, they can better distinguish the intensity levels of the three vignettes. Similar results are observed for the mobility domain: in this sample, most respondents are not obese (the sample contains about 20% obese respondents). The fit vignette images, which match the body size of the majority respondents, lead to a higher correct rate for rank-ordering. Although these findings clearly need to be replicated in other studies, the results suggest that respondents may better perceive the visual vignettes when the vignette figures match more closely with the respondents.

Third, this study revealed many interesting findings on how different respondents view and rate vignettes. It is found that male respondents view the first pain vignette as describing more pain than female. It may be because female experience more pain than male (Cepeda & Carr 2003). They may use themselves as a comparison standard when rating the vignette person, and thus view the first vignette person as depicting minimal pain. Due to the space restrictions, this study cannot present and discuss detailed results for all covariates. Future studies can look more into this. Finally, I also examined the DIF-adjusting results using HOPIT models[10]. Results are similar for both verbal and visual vignettes. Despite the similar assumption-testing and the DIF-adjusting results among verbal and visual vignettes, given the clear advantage of visual vignettes in reducing survey time, lowering respondents' cognitive burden and better

[10] Due to space restrains, results are not shown in this paper.

differentiating the intensity levels, we believe there is a great potential in the use of visual anchoring vignettes in adjusting DIF.

Despite our encouraging results, this study is limited in several ways. First, as mentioned in the Methods section (Section 3.2.4.2), the current RC test is not stringent, and should be viewed only as "suggestive but not definitive" (Grol-Prokopczyk & Carr 2017). The RC violations may actually be larger than what is shown in Figure 3.6. To explore this, future research could use more stringent RC test to compare the two vignette types. Second, the objective health measures used in the RC tests may not fully capture actual health. One may also argue that these objective health questions are based on self-reports and may be subject to reporting errors. Note that these questions designed to capture objective health are straightforward factual questions, for which reporting errors may be less an issue. Also, many of the objective measures used in this study are based on widely-used existing scales with many successful applications described in previous literature (Ray et al. 2009; Kessler et al. 2002). It is true that better ways to capture objective health exist, such as collecting biomarker data, obtaining doctor assessment records and performing body measurements. These measurements, however, are not practical given the survey budget constraints and the mode of the data collection (web survey). Third, this study examined the most commonly used verbal vignettes, included in HRS, SHARE and many other large-scale surveys. It is possible that the verbal vignettes with better worded descriptions may perform better regarding the tests of assumptions than the current verbal vignettes. The same may be true for visual vignettes that better-designed pictures may perform better for assumption-testing. Future research could compare the verbal and visual vignettes with different descriptions or designs. Although more tests are definitely needed to further compare verbal and visual vignettes, it is reasonable to conclude that well-

designed visual vignettes will reduce respondents' cognitive burden and reduce survey response time than verbal vignettes.

Our research identifies several important directions for future research. First, VE is found violated in both verbal and visual vignettes. Future research can further evaluate how to better improve vignettes' VE adherence. Second, all respondents in this study are from the U.S. Future research could evaluate the use of vignettes in a less homogeneous group, such as extending this study to cross-nation surveys and / or to a wide variety of other racial / ethnic groups. Third, the current use of static images for visual vignettes may not well present measures related to change over time and location, such as a slow or quick walking speed. Future research can evaluate other visual vignette designs such as using short videos. Forth, the ways vignettes are presented and their applications can vary by survey modes, which may influence their performances. For example, respondents can hear verbal vignettes in telephone and face-to-face interviews, but they need to read the descriptions in mail and web surveys. Although visual vignettes can be viewed only on a screen or on paper and cannot be delivered orally, the presentation of visual vignettes may also differ by modes. For example, in face to face surveys, interviewers can show the pictures on a show card to respondents and ask them questions, while in mail and web surveys, no interviewers are involved. Given these different ways of asking and displaying vignettes in different modes, future research could evaluate the mode effects for both verbal and visual vignettes. Fifth, this paper focuses on the parametric analysis of anchoring vignettes. Future research could compare verbal and visual vignette performances using nonparametric methods as discussed in King et al. (2004). Finally, this study evaluates only the vignette method with three anchoring vignettes. Future research could evaluate the use of different number of verbal and visual vignettes. In addition, given the budget constraints, it is not practical to test all possible

combinations of vignette characteristics in visual vignettes (such as age, gender, racial / ethnic groups, fitness level or even clothing). Future research could further evaluate the effects of various vignette characteristics on visual vignette performances.

In conclusion, this study indicates that the use of well-designed visual vignettes (with images) can greatly reduce survey time and respondents' cognitive burden than verbal vignettes. Improving VE adherence, or in other words, minimizing different interpretations of vignettes by groups, is critical for both verbal and visual vignettes and requires further investigations. Future implementations of anchoring vignettes can use the findings of this study to introduce efficiencies in the survey designs.

3.5 References

Anderson, C., & Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behavioral and Brain Sciences*, *25*(1), 21-22.

Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Journal of Human Resources*, *46*(October 2009), 875–906.

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541.

Cepeda, M. S., & Carr, D. B. (2003). Women experience more pain and require more morphine than men to achieve a similar degree of analgesia. *Anesthesia and Analgesia*, 1464–1468.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, *71*(4), 623–634.

d'Uva, T. B., O'Donnell, O., & van Doorslaer, E. (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), 1375–1383.

d'Uva, T. B., Lindeboom, M., O'Donnell, O., & Van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. Journal of Human Resources, 46(4), 875-906.

Ferrer-i-Carbonell, A., Van Praag, B. M., & Theodossiou, I. (2011). Vignette equivalence and response consistency: The case of job satisfaction.

Forhan, M., & Gill, S. V. (2013). Obesity, functional mobility and quality of life. *Best Practice & Research Clinical Endocrinology & Metabolism*, *27*(2), 129–137.

Greene, W. H., & Hensher, D. A. (2009). Modelling ordered choices. *Department of Economics, Stern School of Business, New York University, New York, NY, 10012*.

Grol-Prokopczyk, H., & Carr, D. (2017). In Pursuit of Anchoring Vignettes That Work: Evaluating Generality Versus Specificity in Vignette Texts, *0*(0), 1–10.

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and Pitfalls of Anchoring Vignettes in Health Survey Research. *Demography*.

Herr, K. (2002). Chronic pain: Challenges and assessment strategies. *Journal of Gerontological Nursing*, *28*(1), 20–27.

Hibbing, A. N., & Rankin-Erickson, J. L. (2003). A Picture Is Worth a Thousand Words: Using Visual Images to Improve Comprehension for Middle School Struggling Readers, *56*(8), 758–770.

Hirsh, A. T., Alqudah, A. F., Stutts, L. A., & Robinson, M. E. (2008). Virtual human technology: Capturing sex, race, and age influences in individual pain decision policies. *Pain*, *140*(1), 231–238.

Hirve, S., Gómez-Olivé, X., Oti, S., Debpuur, C., Juvekar, S., Tollman, S., … Ng, N. (2013). Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia--testing assumptions. *Global Health Action*, *6*, 21064.

Hoch, C. C., Reynolds, C. F., Kupfer, D. J., Berman, S. R., Houck, P. R., & Stack, J. A. (1987). Empirical note: self-report versus recorded sleep in healthy seniors. *Psychophysiology*, *24*(3), 293–299.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*(681 Icc), 201–222.

Hu, M., and Lee, S., (2016) Context Effects in Anchoring Vignette Questions. *The 71st Annual Conference of the American Association for Public Opinion Research*, Austin, Texas.

Hu, M., Lee, S., and Liu, M. (2013). Using Vignette Questions to Correct for Response Scale

Usage. *Paper presented at the 38th Annual Conference of Midwest Association for Public Opinion Research*, Chicago, Illinois.

Jean-Louis, G., Kripke, D. F., Ancoli-Israel, S., Klauber, M. R., & Sepulveda, R. S. (2000). Sleep duration, illumination, and activity patterns in a population sample: effects of gender and ethnicity. *Biological Psychiatry*, *47*(10), 921–927.

Jürges, H., & Winter, J. (2013). ARE ANCHORING VIGNETTES RATINGS SENSITIVE TO VIGNETTE AGE AND SEX? *Health Economics*, *19*(22), 1–13.

Kapteyn, A., Smith, J. P., Van Soest, A., & Vonková, H. (2011). Anchoring Vignettes and Response Consistency Consistency. *Working Paper*. Retrieved from http://www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR840.pdf

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, *32*(6), 959-976.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enchancing the Validty and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, *98*, 191–207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*, 46–66.

Koster, A., Penninx, B. W. J. H., Newman, A. B., Visser, M., Van Gool, C. H., Harris, T. B., … Kritchevsky, S. B. (2007). Lifestyle factors and incident mobility limitation in obese and non-obese older adults. *Obesity*, *15*(12), 3122–3132.

Krishnan, V., & Collop, N. A. (2006). Gender differences in sleep disorders. *Current Opinion in Pulmonary Medicine*, *12*(6), 383–389.

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

Luna, D., & Peracchio, L. A. (2003). Visual and linguistic processing of ads by bilingual consumers. *Persuasive Imagery: A Consumer Response Perspective*, 153–175.

Manis, M., Biernat, M., & Nelson, T. F. (1991). Comparison and expectancy processes in human judgment. *Journal of Personality and Social Psychology*, *61*(2), 203–211.

Martinez, U. (2012). Cultur (ally) jammed: Culture jams as a form of culturally responsive teaching, *65*(5), 12–18.

Mohler, P., Dorer, B., de Jong, J. & Hu, M. (2016). Translation: Overview. *Guidelines for Best Practice in Cross-Cultural Surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from http://www.ccsg.isr.umich.edu/.

Mojtabai, R. (2015). Depressed Mood in Middle-Aged and Older Adults in Europe and the United States: A Comparative Study Using Anchoring Vignettes. *Journal of Aging and Health*, 1–23.

Molina, T. (2016). Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries. *Demography*, *53*(2), 295–323.

Murray, C. J. L., Özaltin, E., Tandon, A., Salomon, J. A., Sadana, R., & Chatterji, S. (2003). Empirical evaluation of the anchoring vignette approach in health surveys. In C. J. L.Murray & D. B. Evans (Eds.), *Health systems performance assessment: Debates, methods and empiricism* (pp. 369–399). Geneva, Switzerland: World Health Organization.

Murray, C. J. L., Tandon, A., Salomon, J. A., Mathers, C. D., & Sadana, R. (2002). New approaches to enhance cross-population comparability of survey results. *Summary Measures of Population Health: Concepts, Ethics, Measurement, and Applications*, 421–

432.

O'Toole, A. J., Peterson, J., & Deffenbacher, K. A. (1996). An "other-race effect"for categorizing faces by sex. *Perception*, *25*(6), 669–676.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Pan, Y., & De La Puente, M. (2005). Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed. *Survey Methodology*, *6*.

Penny, H., & Haddock, G. (2007). Children's stereotypes of overweight children. *British Journal of Developmental Psychology*, *25*(3), 409–418.

Ray, L., Lipton, R. B., Zimmerman, M. E., Katz, M. J., & Derby, C. A. (2009). Mechanisms of association between obesity and chronic pain in the elderly. *Pain*.

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

Soto, J. A., & Levenson, R. W. (2009). Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. *Emotion*, *9*(6), 874–84.

Stenholm, S., Alley, D., Bandinelli, S., Griswold, M. E., Koskinen, S., Rantanen, T., … Ferrucci, L. (2009). The effect of obesity combined with low muscle strength on decline in mobility in older persons: results from the InCHIANTI study. *International Journal of Obesity*, *33*(6), 635–644.

Townsend, C., & Kahn, B. E. (2014). The "Visual Preference Heuristic": The Influence of Visual versus Verbal Depiction on Assortment Processing, Perceived Variety, and Choice Overload. *Journal of Consumer Research*, *40*(5), 993–1015.

Trautner, M. N., Kwan, S., & Savage, S. V. (2013). Masculinity, Competence, and Health: The Influence of Weight and Race on Social Perceptions of Men. *Men and Masculinities, 16*(4), 432-451.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (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: Statistics in Society, 174*, 575–595.

Vitiello, M. V, Larsen, L. H., & Moe, K. E. (2004). Age-related sleep change: gender and estrogen effects on the subjective--objective sleep quality relationships of healthy, noncomplaining older men and women. *Journal of Psychosomatic Research, 56*(5), 503–510.

Appendix 3. 1 Verbal and visual vignettes used for the web survey for each domain.

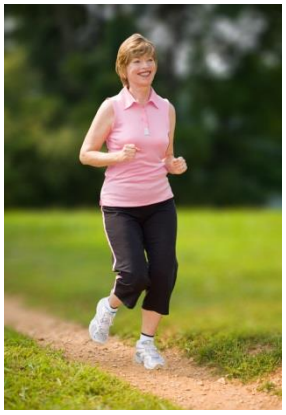
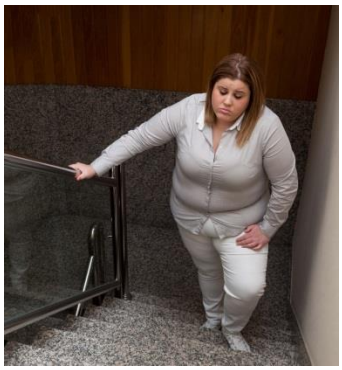
Pain

Pain Intensity Level	Verbal Vignette	Visual Design One (young adults)	Visual Design Two (seniors)
No / low pain	Karen has a headache once a month that is relieved after taking a pill. During the headache she can carry on with her day-to-day affairs.		
Middle Pain	Jennifer has pain that radiates down her right arm and wrist during her day at work. This is slightly relieved in the evenings when she is no longer working on her computer.		
High Pain	Mary 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.		

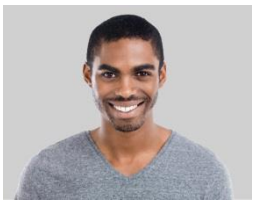
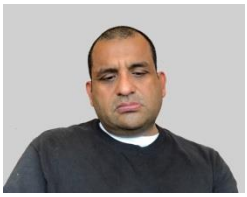
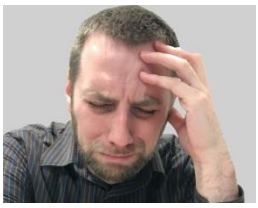
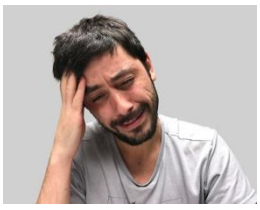
Sleep

Sleep Difficulty Level	Verbal Vignette	Visual Design One (female)	Visual Design Two (male)
No / low Difficulty	Sara / Sam falls asleep easily at night, but two nights a week she / he wakes up in the middle of the night and cannot go back to sleep for the rest of the night.		
Middle Difficulty	Susan / Scott wakes up almost once every hour during the night. When she / he wakes up in the night, it takes around 15 minutes for her / him to go back to sleep. In the morning she does not feel well-rested.		
High Difficulty	Patty / Paul takes about two hours every night to fall asleep. She / he wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again.		

Mobility

Mobility Difficulty Level	Verbal Vignette	Visual Design One (optimal weight / fit)	Visual Design Two (obese)
No / low Difficulty	Laura 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. She has no problems with day-to-day activities, such as carrying food from the market.		
Middle Difficulty	Sandy does not exercise. She cannot climb stairs or do other physical activities because she is obese. She is able to carry the groceries and do some light household work.		
High Difficulty	Lisa 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.		

Affect

Depression Level	Verbal Vignette	White	Black	Hispanic
No / low Depression	Matt enjoys his work and social activities and is generally satisfied with hisr life. He gets depressed every 3 weeks for a day or two and loses interest in what he usually enjoys but is able to carry on with his day-to-day activities.			
Middle Depression Level	David feels nervous and anxious. He worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests her. When he is alone he tends to feel useless and empty.			
High Depression	Leo feels depressed most of the time. He weeps frequently and feels hopeless about the future. He feels that he has become a burden on others and that he would be better dead.			

Appendix 3. 2 Key objective questions included the questionnaire.

Mobility

The following items are about activities you might do during a typical day. Because of a health problem, did you have difficulty doing any of the activities in the **last 30 days**?

	Yes	No
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	1	2
b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2
c. Lifting or carrying groceries	1	2
d. Climbing **several** flights of stairs	1	2
e. Climbing **one** flight of stairs	1	2
f. Bending, kneeling, or stooping	1	2
g. Walking **more than a mile**	1	2
h. Walking **several blocks**	1	2
i. Walking **one block**	1	2
j. Bathing or dressing yourself	1	2

Pain

The following questions are about pain you may have experienced in the **last 30 days**.

During the **last 30 days**, how often did you have pain in the …	None of the time	A little of the time	Some of the time	Most of the time	All of the time
a. head (e.g., persistent headaches or migraine)	1	2	3	4	5
b. teeth					
c. face (e.g., facial ache or pain in the jaw muscles or the joint in front of the ear)	1	2	3	4	5
d. neck and shoulder	1	2	3	4	5
e. back	1	2	3	4	5
f. Arms, elbows, hands, wrist or fingers	1	2	3	4	5
g. Legs, knees, feet, ankles,	1	2	3	4	5
h. chest	1	2	3	4	5
i. abdomen	1	2	3	4	5
j. hips or pelvis	1	2	3	4	5

Affect

The following questions ask about how you have been feeling during **the last 30 days**. For each question, please select the number that best describes how often you had this feeling.

In **the last 30 days**, about how often did you feel …	None of the time	A little of the time	Some of the time	Most of the time	All of the time
a. …nervous?	1	2	3	4	5
b. …hopeless?	1	2	3	4	5
c. …restless or fidgety?	1	2	3	4	5
d. …worthless?	1	2	3	4	5
e. …that everything was an effort?	1	2	3	4	5
f. …so depressed that nothing could cheer you up?	1	2	3	4	5

Sleep

1. These questions are about your sleep habits **in the last 30 days**. Please select one option for each of the following questions. Pick the answer that best describes how often you experienced the situation in the last 30 days.

In the last 30 days...	No, not in the last 30 days	Yes, less than once a week	Yes, 1 or 2 times a week	Yes, 3 or 4 times a week	Yes, 5 or more times a week
a. Did you have trouble falling asleep?	1	2	3	4	5
b. Did you wake up several times a night?	1	2	3	4	5
c. Did you wake up earlier than you had planned to, and were unable to fall asleep again?	1	2	3	4	5

2. In the last 30 days, did you have trouble falling asleep?

No, not in the last 30 days
Yes, less than once a week

Yes, 1 or 2 times a week
Yes, 3 or 4 times a week
Yes, 5 or more times a week

3. In the last 30 days, did you wake up several times a night?

No, not in the last 30 days
Yes, less than once a week
Yes, 1 or 2 times a week
Yes, 3 or 4 times a week
Yes, 5 or more times a week

4. In the last 30 days, did you wake up earlier than you had planned to, and were unable to fall asleep again?

No, not in the last 30 days
Yes, less than once a week
Yes, 1 or 2 times a week
Yes, 3 or 4 times a week
Yes, 5 or more times a week

Appendix 3. 3 HOPIT model specifications.

The method most often used for anchoring vignette data analysis involves the Hierarchical Ordered Probit (HOPIT) model, which is a generalized ordered probit model (Greene & Hensher, 2009; King et al., 2004). As mentioned earlier, HOPIT model uses vignettes to identify the group-specific cutpoints and compares where respondent's self-assessments stand relative to their cutpoints for the categories. Thus, it involves two components: the model for self-assessment and the model for vignettes.

Model for Self-Assessment:

The self-assessment component starts with the distribution of true state for person i, Y_i^*, where $i = 1, \ldots, N$

$$Y_i^* \sim N(\mu_i, \sigma^2).$$

Note that Y_i^* reflects an unobserved continuum (e.g., pain level). The mean of Y_i^*, μ_i, is a latent variable further modeled using linear regression on covariates, $\boldsymbol{X}_i$, such as cultural groups, age and gender, as $\mu_i = \boldsymbol{X}_i\beta + \eta_i$, where η_i is an independent random error, which follows a normal distribution as:

$$\eta_i \sim N(0, \omega^2).$$

The actual response in the self-assessment question is denoted as Y_i, where the question is asked with a response scale with K response categories, $k = 1, \ldots, K$. Y_i reflects Y_i^* as follows:

$$Y_i = k, \text{ if } \tau_i^{k-1} < Y_i^* \leq \tau_i^{k}.$$

Here, $\tau_i = (\tau_i{}^0, \tau_i{}^1, \ldots, \tau_i{}^k, \ldots, \tau_i{}^K)$ is a vector of thresholds (i.e., cutpoints) respondent *i* uses to answer the question. For instance, $\tau_i{}^1$ is where response categories of "None" and "Mild" in Figure 1.1 are differentiated. Naturally, $\tau_i{}^0 = 0$ and $\tau_i{}^K = \infty$. These person-specific thresholds are modeled as:

$$\tau_i{}^1 = \gamma^1 \boldsymbol{Z}_i \text{ for } k = 1 \text{ and } \tau_i{}^k = \tau_i{}^{k-1} + \mathrm{e}^{\gamma^k \boldsymbol{Z}_i} \text{ for } k = 2, \ldots, K-1,$$

where $\boldsymbol{Z}_i$ is a vector of covariates relevant to the thresholds, which may or may not be the same as $\boldsymbol{X}_i$.

Model for Vignette Responses:

Vignette responses are modeled in a similar way as the self-assessment. Let $V_{ij}{}^*$ be the unobserved state of the person in the j^{th} vignette item perceived by respondent i, which follows a normal distribution with a random error denoted as:

$$V_{ij}{}^* \sim N(\alpha_j, \sigma_j^2),$$

where $i = 1, \ldots, N$ (the respondents can be different from those who answered for self-assessments, if vignettes are only asked to a subset of the whole sample) and $j = 1, \ldots, J$. Note that α_j is assumed to be the same across respondents, which implies the VE assumption.

Respondents' response to the j^{th} vignette is denoted as v_{ij} , which is obtained by mapping the latent $V_{ij}{}^*$ onto the response scale and modeled as:

$$v_{ij} = k, \text{ if } \tau_l{}^{k-1} < V_{ij}{}^* \leq \tau_l{}^k.$$

The thresholds of vignettes are modeled exactly the same as the self-assessment model.

$$\tau_i{}^1 = \gamma^1 \boldsymbol{Z}_i \text{ for } k = 1 \text{ and } \tau_i{}^k = \tau_i{}^{k-1} + \mathrm{e}^{\gamma^k \boldsymbol{Z}_i} \text{ for } k = 2, \ldots, K-1.$$

Imposing the thresholds to be modelled the same for the vignettes and the self-assessment implies the RC assumption.

Appendix 3. 4 VE assumption tests for sleep, mobility and affect.

Figure 3.7 shows the estimated vignette locations for sleep domain – 3.7A for verbal vignettes and 3.7B for visual vignettes. As can be seen from Figure 3.7A, the estimated vignette locations across racial / ethnic groups are very similar, indicating that respondents regardless of racial / ethnic background view the vignettes in similar ways. However, it is worthwhile to note that the perceived vignette location for the second vignette is not significantly different from the reference vignette, suggesting that the sleep verbal vignettes, at the first place, failed to provide good distinction of the second and third vignettes. As shown in Figure 3.7B, despite the VE violation (e.g., Hispanics who took the Spanish survey view the first vignette person as with more sleep difficulties comparing with the White respondents), visual vignettes did a much better job to differentiate the intensity levels of the three vignettes.

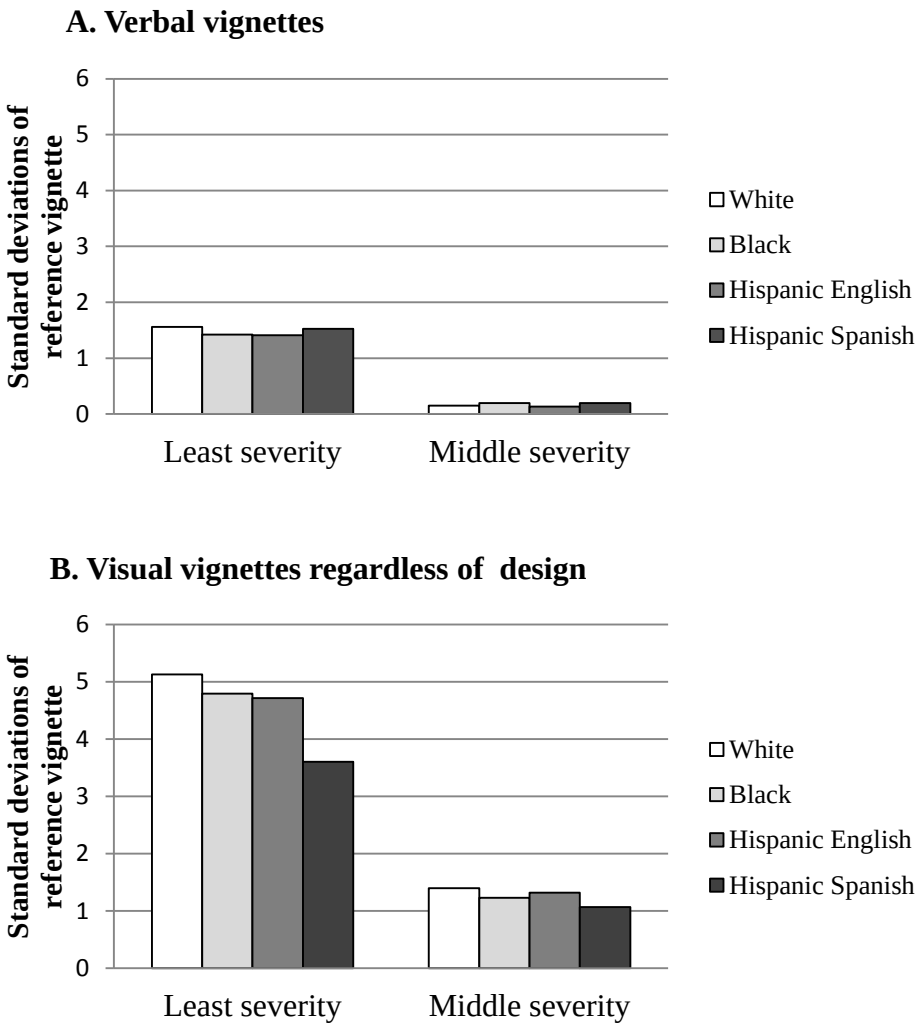
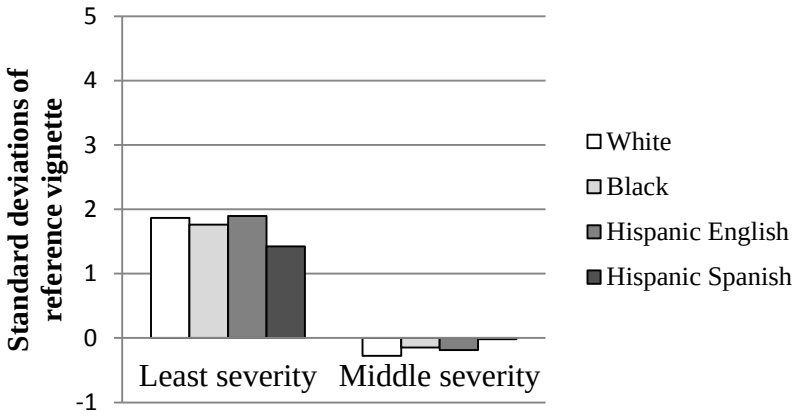


Figure 3. 7 Estimated sleep vignette locations (on latent health spectrum; relative to Severity 3). Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

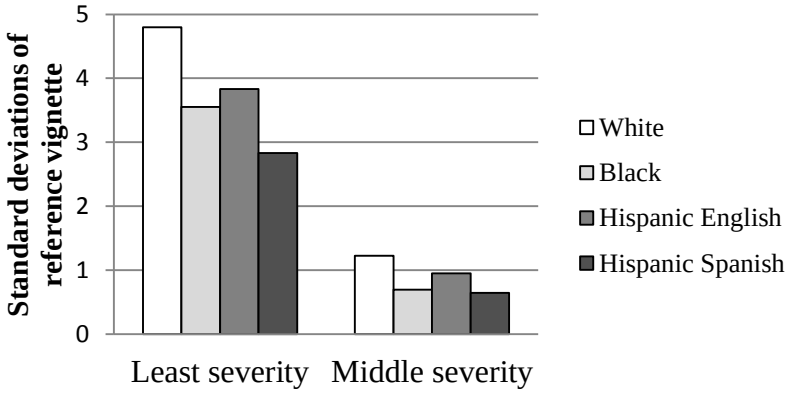
Figure 3.8 shows the estimated vignette locations for mobility domain – 3.8A for verbal vignettes, 3.8B for visual vignettes using fit vignette figure and 3.8C for visual vignettes with obese vignette figure. As can be seen from Figure 3.8A – 3.8C, the estimated vignette locations vary by racial / ethnic groups for all three conditions – the violation of VE. For Figure 3.8A (verbal vignettes), similar as what we see for the sleep domain (Figure 3.8A), the perceived

vignette location for the second vignette is not significantly different from the reference vignette, indicating the difficulties to distinguish the two vignettes. Visual vignettes, both Figure 3.8B and 3.8C, did a much better job to differentiate the middle severity vignette and the highest severity vignette. Comparing Figure 3.8B to Figure 3.8C, respondents can better distinguish the least and middle severity vignettes when presented with the fit vignette figures (Figure 3.8B), while the difference for the two vignettes in Figure 3.8C is much smaller.

A. Verbal vignettes



B. Visual vignettes with fit vignette figures



C. Visual vignettes with obese vignette figures

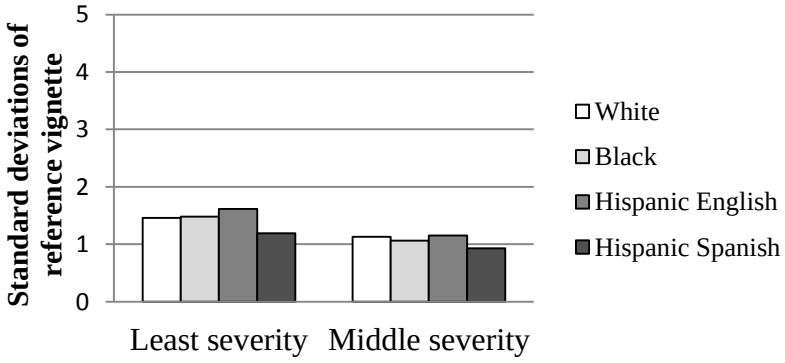


Figure 3. 8 Estimated mobility vignette locations (on latent health spectrum; relative to Severity 3). Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

Figure 3.9 shows the estimated vignette locations for affect domain – 3.9A for verbal vignettes, 3.9B for visual vignettes. As shown in Figure 3.9, the estimated vignette locations vary by racial / ethnic groups for both verbal and visual vignettes – Hispanics who completed the Spanish questionnaires seem to view the same vignettes as with more pain than the other racial / ethnic groups.

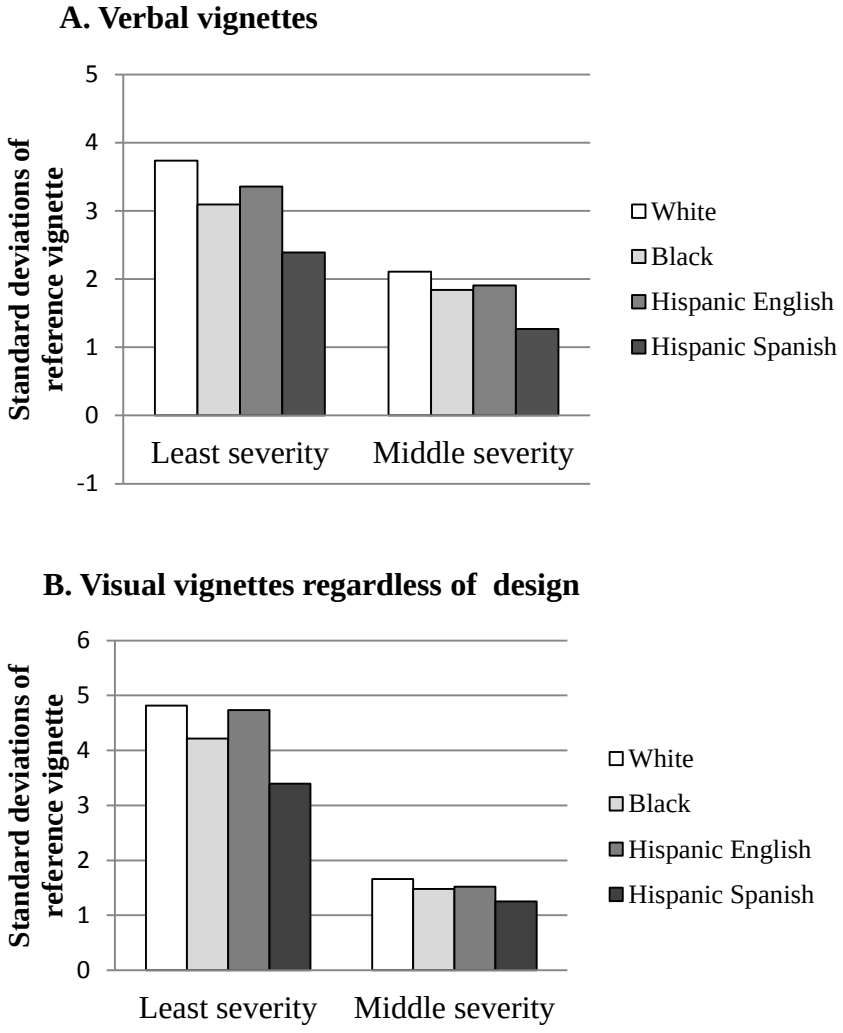


Figure 3. 9 Estimated affect vignette locations (on latent health spectrum; relative to Severity 3). Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

Appendix 3.5. RC assumption test results for the pain, sleep and affect domain.

Pain

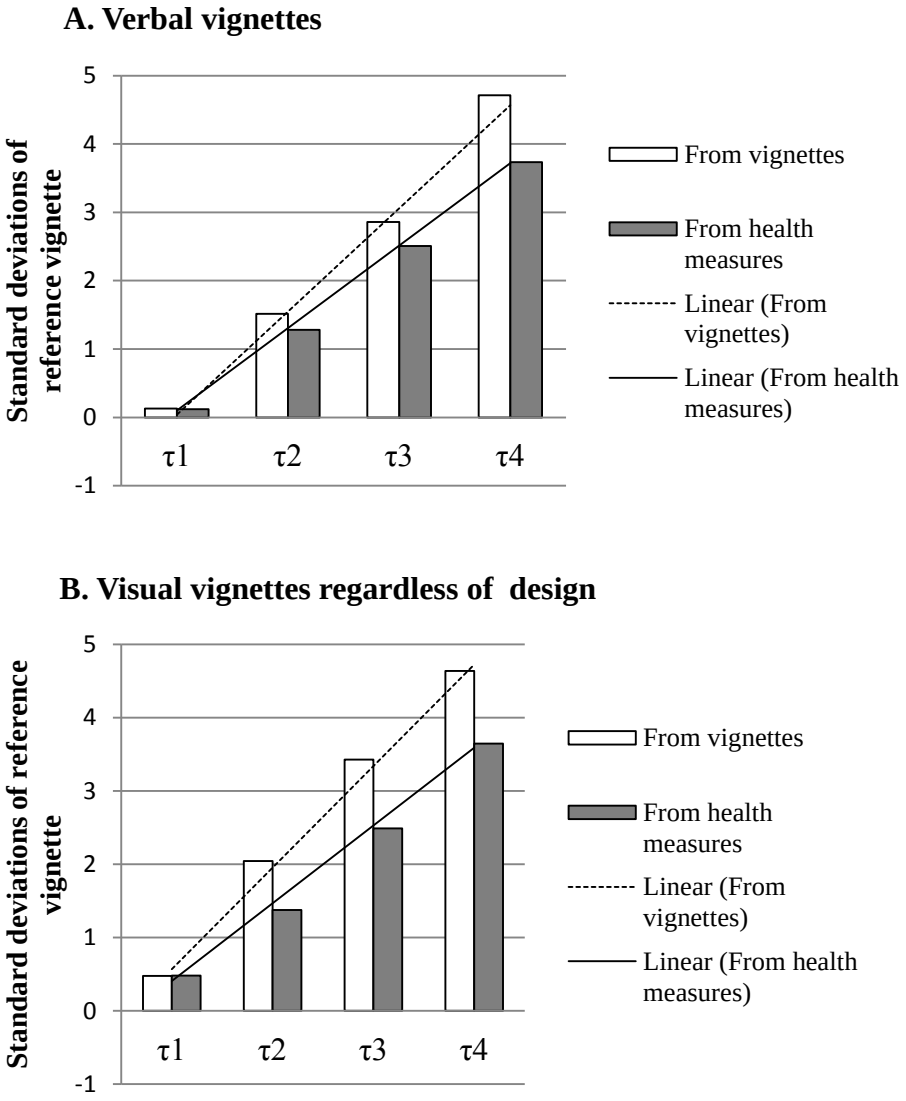


Figure 3. 10 Estimated cutpoints for pain based on vignettes and health measures. τ1 – τ4 are cutpoints for the five-point response scale from "None" to "Extreme" (e.g., τ1 is the cutpoint between "None" and "Mild").

Sleep

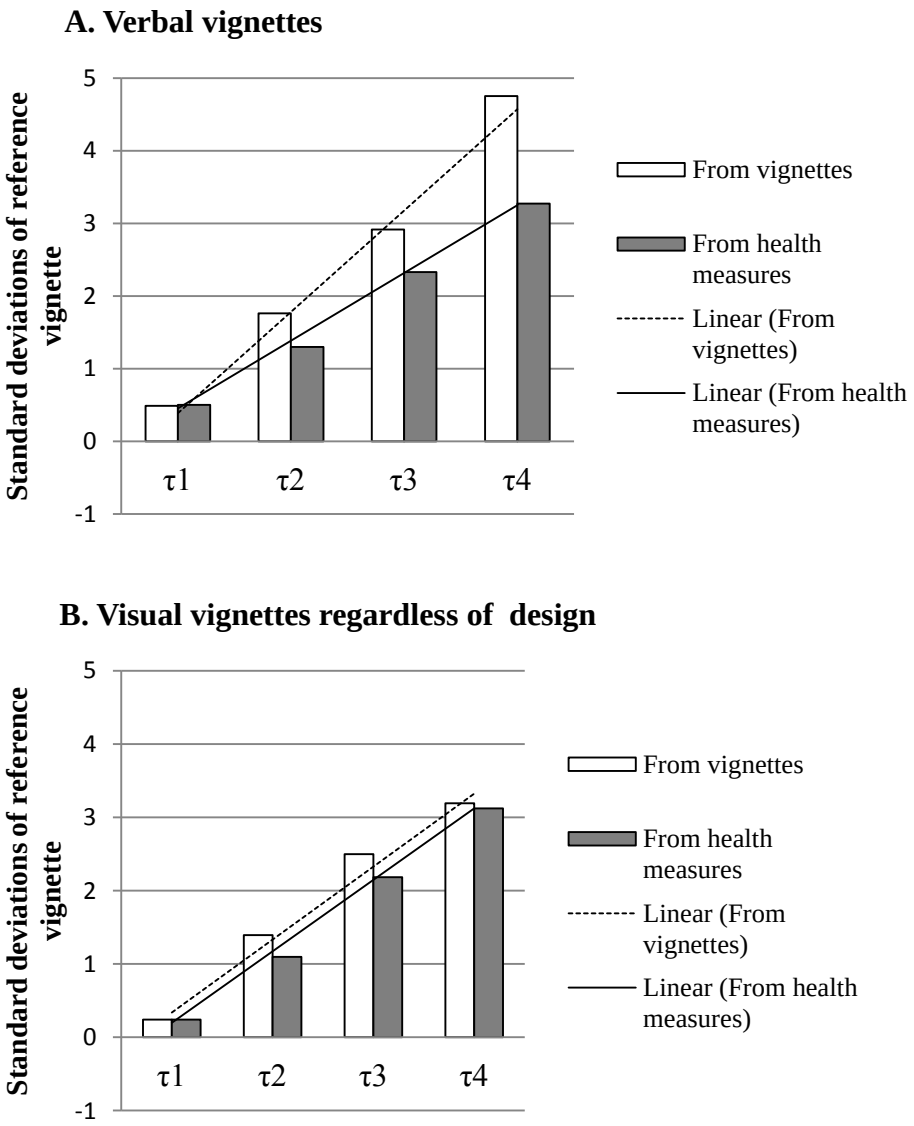


Figure 3. 11 Estimated cutpoints for sleep based on vignettes and health measures. τ1 – τ4 are cutpoints for the five-point response scale from "None" to "Extreme" (e.g., τ1 is the cutpoint between "None" and "Mild").

Affect

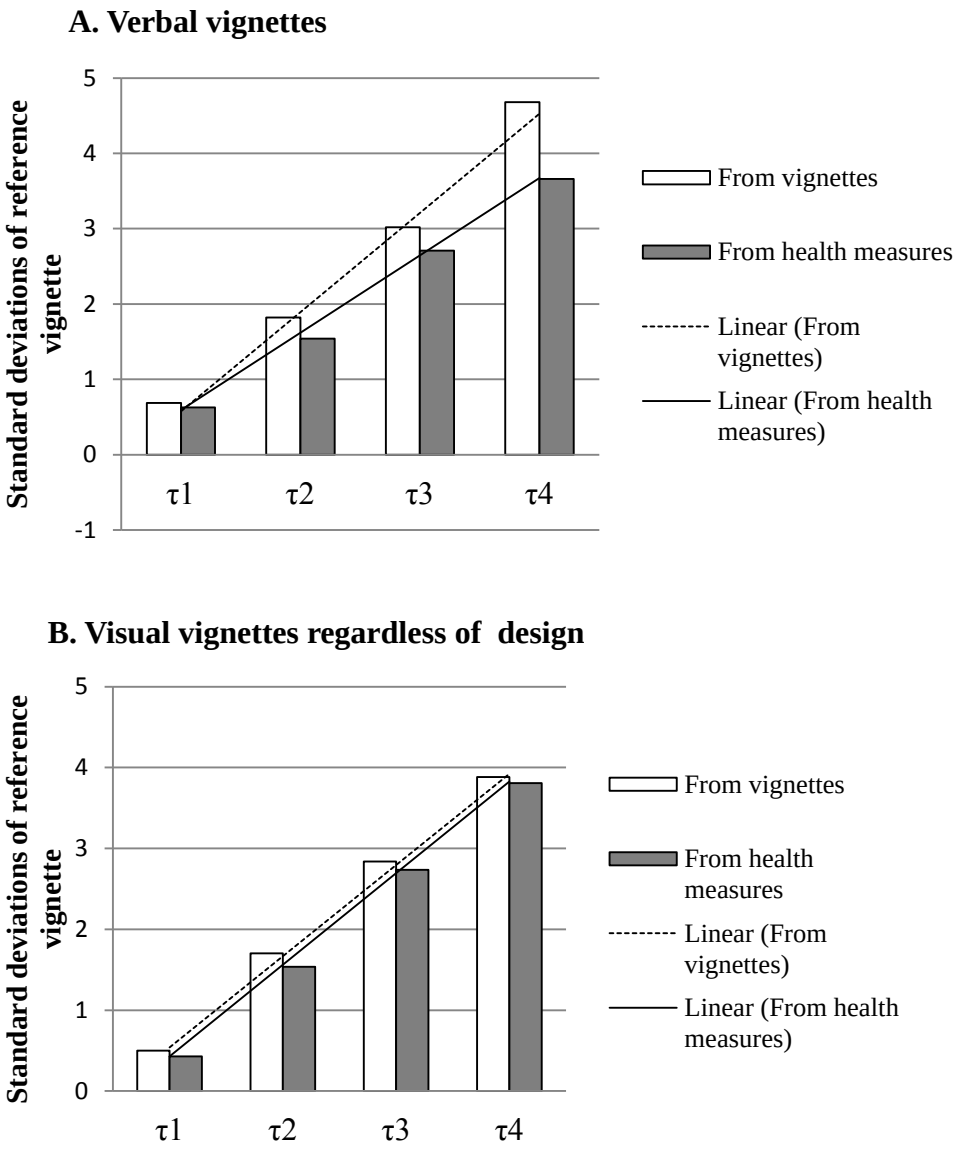


Figure 3. 12 Estimated cutpoints for affect based on vignettes and health measures. τ1 – τ4 are cutpoints for the five-point response scale from "None" to "Extreme" (e.g., τ1 is the cutpoint between "None" and "Mild").

CHAPTER IV

Survey Context Effects in Anchoring Vignettes

4.1 Introduction

Despite the use of anchoring vignettes in many surveys worldwide, little work has been done to address the issues of their administration and design considerations. Specifically, an overlooked but critical question surrounding anchoring vignette method is the placement of self-assessments relative to the vignette questions. This is important for two reasons. First, given that vignette questions are asked together with subjective self-assessment questions for the same construct, the similarity in question topics imply potential question order effects (Schuman & Presser 1996). Second, the subjective nature of the self-assessment questions suggest that they are prone to question order effects (Schuman & Presser 1996).

Question order effects refer to the phenomenon in which preceding questions in a questionnaire may influence respondents' responses to a later question (Schuman & Presser 1996). The presence of question order effect can lead to several potential issues that affect survey data quality. First, it is likely to be a source of systematic measurement error, which may bias survey estimates and prevent generalization of the results. Second, in longitudinal data analysis, comparisons across time or trend analysis may be invalid if different orders are used at different waves. Third, it prevents valid comparisons of the same construct measured in surveys with different question contexts. These issues become even more critical for the use of anchoring vignette method, given that they are commonly used in cross-cultural research, which often requires comparisons of data from different surveys.

Given all these aforementioned issues and the likelihood of question order effects on self-assessment questions, research evaluating potential question order effects and appropriate placement of self-assessments relative to vignettes are of critical importance. It is, therefore, surprising that, to date, only two studies have evaluated the question order effects related with traditional anchoring vignette method. Buckley (2008) first evaluated how the placement of self-assessment questions relative to vignettes influences survey response to parent-teacher relationship, and found evidence of question order effects. As discusses in their paper, placing the self-assessment item before the vignette series leads to a systematic positive inflation of responses for the self-assessment (Buckley 2008). In other words, a priming effect may occur when vignettes are asked before self-assessments, and respondents may rate the relationship more negatively. In light of this finding, Buckley recommends that survey researchers should randomize the placement of self-assessment questions relative to vignette questions in practice.

Another study of question order effects on political efficacy-related vignettes also found that vignettes, when presented first, can prime respondents to understand the self-assessment question and response scales as intended (Hopkins & King 2010). The authors thus recommended placing self-assessment questions after the vignettes. Although this recommendation seems reasonable, in fact, there are two important challenges that cannot be ignored. One key challenge concerns the implementation of anchoring vignettes in data collection. Given that multiple vignettes are designed to correct for differential item functioning (DIF) for each self-assessment question, the anchoring vignettes can lead to increased survey time and respondent burden as mentioned in Chapter 3. To address this issue, vignettes are often asked to only a small random subsample (e.g., in HRS, CHARLS and SHARE) and the results from this subsample can then be used to correct DIF for all respondents in the survey, including

those who were not asked the vignette questions as suggested in King et al. (2004). If such an approach is followed, the placement of self-assessment question after vignettes for only a subsample can impose question order effects only to the selected subsample, leading to incomparability of self-assessments between this subsample and the rest respondents. Another challenge centers on the use of these questions in cross-cultural studies. As mentioned in Chapter 3, anchoring vignettes are commonly used in multi-cultural and multi-lingual studies. It is found in previous literature that the question order effects on self-assessed health (e.g., before or after specific chronic health conditions) can vary by interview languages (Lee & Grant 2009). If the priming effects of vignettes on self-assessments differ by cultural groups or interview language, it may introduce additional source of measurement error to cross-cultural comparability.

Besides the lack of consensus on the placement of self-assessment relative to vignettes, the existing literatures have several noteworthy weaknesses. The first lies in their lack of attention to the evaluation of question order effects on health-domain vignettes, which are the most widely applied vignette questions worldwide (e.g., as in HRS, SHARE, ELSA, CHARLS and many other cross-national surveys). It is likely that question order effects differ by topics / domains (e.g., more subjective questions may be more prone to question order effects), which means that results of previous studies on placement of vignettes (e.g., political efficacy and education related vignettes) cannot be applied directly to health domains. Second, no prior studies have evaluated the question order effects related to other types of vignette designs, such as visual vignettes. It remains unknown whether the same findings found for verbal vignettes can be applied to visual vignettes. Third, prior studies have ignored the influence of potential question order effects for different racial / ethnic groups. This is of particular importance to the

application of anchoring vignette method, since it is commonly used in cross-cultural / national studies.

In an attempt to fill the research gaps, this chapter evaluates the question order effects related to the placement of self-assessment for both verbal and visual vignettes on four health domains using a sample that includes multiple racial / ethnic groups. Specifically, this study aims to investigate the following research questions:

1) Are there any question order effects on the self-assessed pain, sleep, mobility and affect, and are some domains more susceptible to question order effects?

2) For each domain, do the question order effects differ by vignette types (i.e., verbal vs. visual vignettes)?

3) For each domain, do the question order effects differ by racial / ethnic groups?

4.2 Method

The data was collected in the same web survey as described in Chapter 3. I randomized two experimental factors in the web survey – the placement of self-assessments (before vs. after vignettes) and vignette types (one verbal vignette condition and two visual vignette conditions, see Chapter 3 Methods). The data included 760 non-Hispanic White, 750 non-Hispanic Black, 750 Hispanics interviewed in English and 889 Hispanics interviewed in Spanish. For each racial / ethnic group, about 20% of the respondents were assigned to the *vignettes first, self-assessment last* (V → SA) order – 160 non-Hispanic White, 150 non-Hispanic Black, 158 Hispanics interviewed in English and 191 Hispanics interviewed in Spanish. The rest 78% - 80% were

assigned to the *self-assessment first, vignettes last* (SA → V) order condition[11]. For each of the four health domains (i.e., pain, sleep, mobility and affect), the orders of the three vignette questions were randomized. The orders of the four health domains were also randomized. As respondents were randomized, there were no significant differences in demographic characteristics (e.g., age, gender, education) between the two question order conditions.

For each domain, the self-assessment estimates were first compared by question order and vignette type. I then compared the estimates by question order and racial / ethnic groups. For mobility domain, which was found subject to question order effects, I further examined the effect of question order on self-assessed mobility by racial / ethnic groups in each vignette type condition. Chi-square tests were performed to determine significance. SAS 9.3 was used for analysis.

4.3 Results

Table 4.1 shows the response distribution of self-assessed difficulty on the four health domains by vignette types and question order. There were no question order effects for pain, sleep and affect domains, regardless of the vignette type. The only domain which showed differences between the two order conditions was mobility. Overall, compared with the "self-assessment (SA) first" condition, the percentage of respondents who reported "none" for the mobility problem questions decreased from about 57% to 47%, while estimates for the other four categories increased. The same pattern was found when we broke down to each vignette type. For verbal vignette condition, the estimate for "none" decreased from about 58% to 44% when

[11] Ideally, I could have equal numbers of the two order conditions. However, this is not practical given the limited budget constrain. In addition, the power analysis indicates that the sample size is sufficient to detect question order effects.

the self-assessment was placed after the vignettes. The Chi-square tests indicated question order effects for mobility in the verbal vignettes (p-value = 0.003), and in the visual vignettes with fit person in the images (p = 0.04). Although not statistically significant, the estimate changes were in the same direction for the visual vignettes with obese persons.

To evaluate whether the question order effects differ by racial / ethnic groups, I further evaluated the four health domains by racial / ethnic groups and question order. Consistent with the results presented in Table 4.1, there were no question order effects for pain, sleep and affect domains by racial / ethnic groups. Thus, here I only present the response distribution of mobility by racial / ethnic groups and question order in Table 4.2. The Chi-square test for the mobility domain indicated significant question order effects for non-Hispanic Black and Hispanics interviewed in English, where the estimates for "none" decreased from about 60% to 44% and from about 57% to 42%, respectively. Although the Chi-square tests were not significant for the non-Hispanic White and Hispanics interviewed in Spanish, the estimate changes were in the same direction with the other groups. Substantively, those 16 percent point decrease for non-Hispanic Black and the 15 percent point decrease for Hispanics interviewed in English are much larger than the 4 percent point decrease for Non-Hispanic White.

Table 4. 1 Response distribution of the self-assessments for four health domains by vignette types and question order.

	Pain		Sleep		Mobility		Affect	
	SR → V	V → SR	SR → V	V → SR	SR → V	V → SR	SR → V	V → SR
Verbal vignettes								
n	840	211	840	211	840	211	840	211
None (%)	23.3	20.4	22.7	17.5	57.6	44.1	43.0	36.0
Mild (%)	39.3	38.4	30.5	41.2	24.4	28.9	29.2	34.1
Moderate (%)	25.7	27.5	29.6	26.1	13.0	17.1	19.4	23.7
Severe (%)	9.5	10.4	12.4	11.4	4.2	8.1	5.4	3.8
Extreme (%)	2.1	3.3	4.8	3.8	0.8	1.9	3.1	2.4
χ^2	2.0		9.3		16.0**		6.0	
Visual vignettes Design 1	Older adult		Female		Fit		Same	
n	826	214	841	227	841	220	825	218
None (%)	20.6	23.4	20.1	19.8	54.5	46.4	42.4	40.8
Mild (%)	39.6	39.3	32.0	35.2	24.7	31.8	29.6	32.6
Moderate (%)	30.5	26.6	30.8	28.2	15.7	13.2	20.0	19.3
Severe (%)	7.5	8.9	12.1	14.5	4.2	7.3	6.1	7.3
Extreme (%)	1.8	1.9	5.0	2.2	1.0	1.4	1.9	0.0
χ^2	1.9		5.0		10.0*		5.3	
Visual vignettes Design 2	Young adult		Male		Obese		Different	
n	824	234	809	221	809	228	825	230
None (%)	20.2	21.4	21.8	22.6	58.5	50.4	43.4	51.3
Mild (%)	41.8	42.7	29.4	24.4	23.4	27.6	31.5	27.4
Moderate (%)	26.8	25.6	31.8	36.2	13.2	15.8	16.6	15.7
Severe (%)	9.3	9.0	12.5	12.7	3.5	4.4	6.4	4.4
Extreme (%)	1.9	1.3	4.6	4.1	1.5	1.8	2.1	1.3
χ^2	0.7		2.7		4.7		5.4	
All vignettes								
n	2490	659	2490	659	2490	659	2490	659
None (%)	21.4	21.7	21.5	20.0	56.8	47.0	42.9	42.9
Mild (%)	40.2	40.2	30.6	33.5	24.2	29.4	30.1	31.3
Moderate (%)	27.7	26.6	30.7	30.2	14.0	15.3	18.7	19.4
Severe (%)	8.8	9.4	12.3	12.9	3.9	6.5	5.9	5.2
Extreme (%)	2.0	2.1	4.8	3.3	1.1	1.7	2.4	1.2
χ^2	0.5		4.5		24.8***		4.2	

Notes: design 1 vs. design 2 of visual vignettes conditions for each domain are: pain – old (design 1) vs. young (design 2); sleep – female vs. male; mobility – fit vs. obese; affect – same race group vs. different race groups. *, P≤0.05; **, P≤0.01; ***, P≤0.001.

Table 4. 2 Response distribution of the health domains by racial / ethnicity groups and question order for mobility domain.

	Non-Hispanic White		**Non-Hispanic Black**		**Hispanics (English)**		**Hispanics (Spanish)**	
	SA → V	V → SA	SA → V	V → SA	SA → V	V → SA	SA → V	V → SA
n	600	160	600	150	592	158	698	191
None	54.0	50.0	60.2	44.0	56.6	42.4	56.6	50.8
Mild	27.3	30.0	22.8	29.3	26.5	30.4	20.6	28.3
Moderate	14.7	13.1	12.5	14.7	11.5	18.4	16.8	15.2
Severe	3.0	5.0	4.0	10.7	4.4	6.3	4.3	4.7
Extreme	1.0	1.9	0.5	1.3	1.0	2.5	1.7	1.1
Rao–Scott χ2	3.2		19.3***		13.0*		5.6	

Notes: *, P≤0.05; **, P≤0.01; ***, P≤0.001.

When the self-rated mobility was dichotomized by combining ‘‘moderate’’, “severe” and ‘‘extreme’’ into one category, the question order effect was even more apparent. As shown in Figure 4.1, significant differences by question order were shown for the Non-Hispanic (NH) White and the Hispanics interviewed in English, with the estimate for the “moderate to extreme” increased from about 17% to about 27%. The order effects were not significant for the other two racial / ethnic groups. This pattern, in general, holds for other domains – although not statistically significant for domains other than mobility, Hispanics interviewed in English tend to have the highest increase rate for most of the domains, comparing to other racial / ethnic groups.

I also examined the interaction effect between question order and racial / ethnicity by adding the interaction term in a logistic regression predicting “moderate to extreme” reports. The model result confirms the significant interaction effects ($p<0.05$).

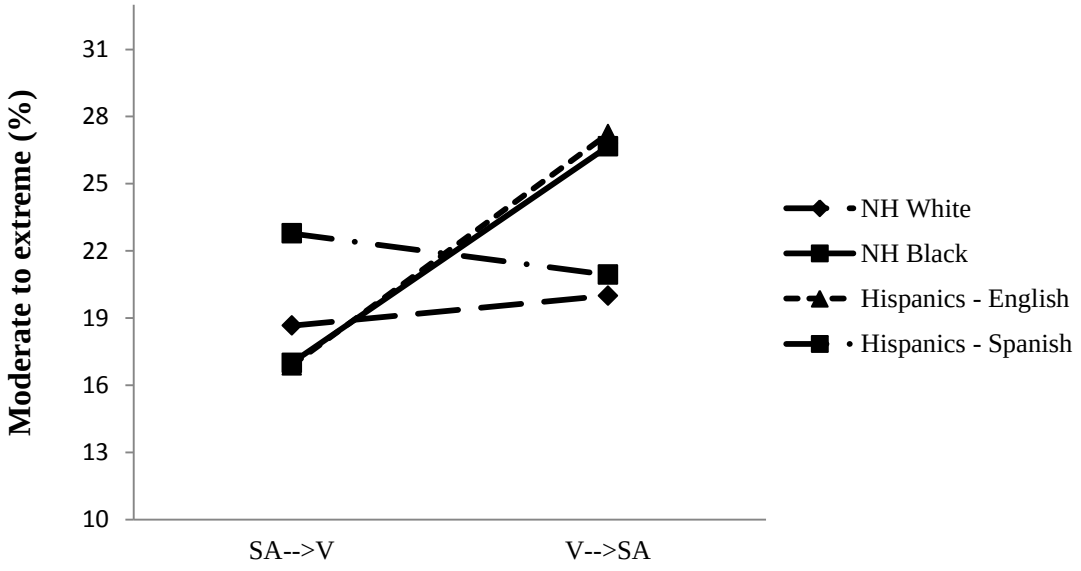


Figure 4. 1 Percentage of reported moderate to extreme mobility problems by racial / ethnic groups and question order. SA, self-assessment question; V, vignette questions.

I further examined the effect of question order on self-rated mobility by racial / ethnic groups in each vignette type condition (Figure 4.2). Consistent with Figure 4.1, significant differences by question order were shown for NH Black (the estimate for "moderate to extreme" increased from 14.0% to about 28.0%) and Hispanics interviewed in English (the estimate increased from 17.0% to about 32%) for the verbal vignette condition (Figure 4.2A). For NH Whites assigned to the verbal condition, there is marginally significant question order effects (Chi-square test p-value=0.08), with the estimate for the "moderate to extreme" increasing from 17% to 28%. For the fit vignette condition, the question order effects are marginally significant for NH White (Chi-square test p-value =0.08) and Hispanics interviewed in English (Chi-square test p-value =0.06) (Figure 4.2B). For NH White, the estimate for the "moderate to extreme" decreased from 22% to 12%, while for Hispanics interviewed in English, the estimate increased from 17% to 28%. As for the obese vignette condition (Figure 4.2C), significant differences by

question order were shown for NH Black, with the estimate for the “moderate to extreme” increasing from about 16% to 31%.

Similar as for Figure 4.1, I examined the interaction effect between question order and racial / ethnicity by adding the interaction term in a logistic regression predicting “moderate to extreme” reports for each vignette type condition. The Wald chi-square test results indicate that the interaction terms are not significant for all conditions (likely related to the smaller sample size). Results for other domains are shown in Appendix 4.1.

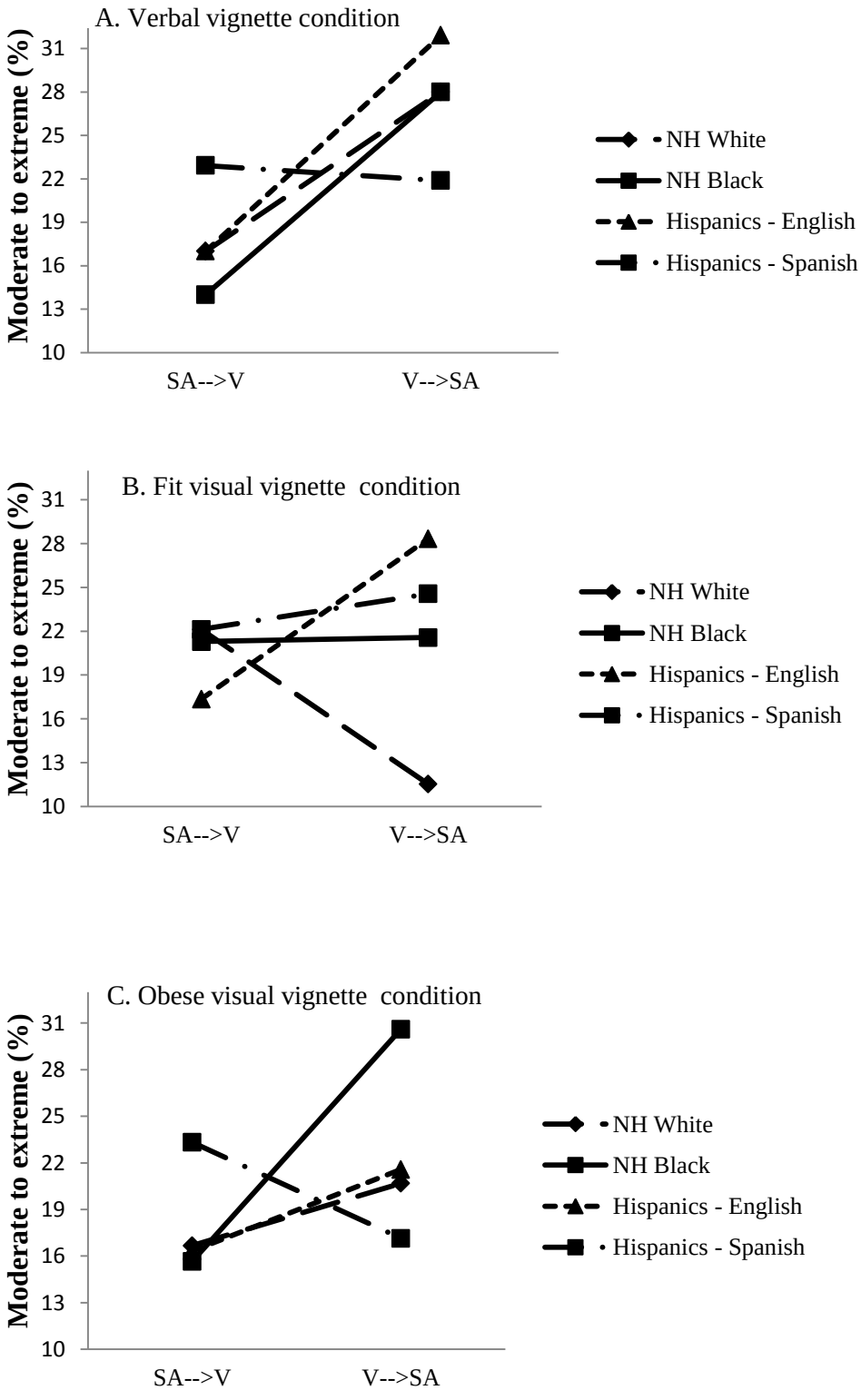


Figure 4. 2 Percentage of reported moderate to extreme mobility problems by racial / ethnic groups and question order among those who were assigned into A) the verbal condition, B) fit vignette condition and C) obese vignette condition. SA, self-assessment question; V, vignette questions.

4.4 Discussion

This study examined question order effects on self-rated pain, sleep, mobility and affect, in the context of anchoring vignette questionnaire design in a cross-cultural survey setting. For the first research question, this study showed that for the four health domains examined, only the mobility domain showed significant question order effects. This may be due to the fact that compared with the other three domains (pain, sleep and affect), mobility seemed to be more visible, through signs like loss of balance and slow walking speed, which can be directly observed. Pain, sleep and affect, on the other hand, were thought to be the "unmeasurable parameters", which are, by their nature, more abstract constructs and difficult to quantify (Aitken, 1969; Huskisson, 1974). For example, pain is an internal experience, but can only be observed through external signs such as facial expressions, body positions and movements (Snow et al. 2004). It may be the case that the mobility vignettes, with more visible signs, could provide more easily-accessible cues to respondents and trigger their relevant memory, and thus influence their responses. This finding suggests that question order effects likely differ by domains, and that the priming effect does not always occur.

Results for the question order effects by vignette types (the second research question) indicate that, for the mobility domain, self-assessments in both verbal and visual vignette conditions may be susceptible to question order effects, while verbal vignettes seem to have more priming effects on self-reported mobility. This is not surprising given that respondents in general spent more time on verbal vignettes (see Chapter 2), which provides them more time to pick up on the cues in vignette descriptions, retrieve relevant information and adjust their responses. For the two mobility vignette designs – fit vs. obese vignettes, only the fit vignette

condition showed question order effects. This may be because that majority of the respondents were not obese (the sample contains about 20% obese respondents), and the fit vignette figures are easier for respondents to relate to, leading to more priming effects.

Results for the question order effects by racial / ethnic groups (the third research question) indicate that question order effects could differ among racial / ethnic groups (i.e., there is question order effects for non-Hispanic Black and Hispanics, but not for non-Hispanic White for mobility domain). It should also be noted that the choice of question order may greatly change the inference. As shown in Figure 4.1 and Figure 4.2, when vignettes were asked first, the differences between the cultural groups increased dramatically and significantly, comparing to when vignettes were asked last (e.g., NH White and NH Black in Figure 4.1). These findings have general implications for anchoring vignette design and survey practice. Since anchoring vignettes are often used in cross-national and cross-cultural studies, it is important to ensure that the results are not confounded by context effects. When the priming effects on self-assessments differ by cultural groups or by language, this introduces an additional source of measurement error to cross-cultural comparability. The priming effects may inflate or deflate the observed differences across cultural groups and make the data even less comparable. It may also likely bias the vignette adjustment effects, adding to the difficulties of disentangling context effects as well as reporting heterogeneity and true attitudes / perceptions. Future studies can further evaluate this. Results for the question order effects by racial / ethnic groups suggest that researchers and practitioners need to be cautious of the context effects on self-assessments when using the *vignette first, self-assessment last* order, especially in cross-cultural studies.

Given the importance of the four health-domain measures and the growing need to improve the anchoring vignette designs, this study identifies three important directions for future

research. First, as question order effects could differ by health domains, future research can expand the current work and evaluate other health domains, such as cognition and vision. Second, this study focuses on question order effects on self-assessments. In future research, in order to determine which vignette order works better in terms of DIF-controlling and to provide better recommendations on vignettes question orders, I will evaluate the question order effects on vignettes, and the visual and verbal vignette performances in different survey contexts. Third, it is unknown what context effects may exist for other racial / ethnic groups other than the three groups (e.g., Asians). Future research could further evaluate whether the order effects may appear in other racial / ethnicity groups.

In conclusion, this study highlights question order effects on health self-assessment questions, and that the effects could differ by health domains, vignette designs and by racial / ethnic groups. This study has important implications for future methodology research and survey practice. It suggests that a survey researcher using anchoring vignettes should be aware of the question order effects, especially in a cross-cultural study. Decisions on the question order designs need to be based on careful consideration and pretesting results.

4.5 References

Aitken, R. C. (1969). Measurement of feelings using visual analogue scales. Proceedings of the Royal Society of Medicine, 62(10), 989–993.

Buckley, J., 2008. Survey Context Effects in Anchoring Vignettes. , (2007), pp.1–14.

Hopkins, D.J. & King, G., 2010. Improving anchoring vignettes designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, 74(681 Icc), pp.201–222.

Huskisson, E. C. (1974). Measurement of pain. *The lancet*, *304*(7889), 1127-1131.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enchancing the Validty and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, *98*, 191–207.

Lee, S. & Grant, D., 2009. The effect of question order on self-rated general health status in a multilingual survey context. *American journal of epidemiology*, 169(12), pp.1525–30.

Schuman, H. & Presser, S., 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*, Sage.

Snow, A. L., O'Malley, K. J., Cody, M., Kunik, M. E., Ashton, C. M., Beck, C., … Novy, D. (2004). A conceptual model of pain assessment for noncommunicative persons with dementia. *The Gerontologist*, *44*(6), 807–817.

Appendix 4. 1 Question order effects for pain, sleep and affect.

Pain

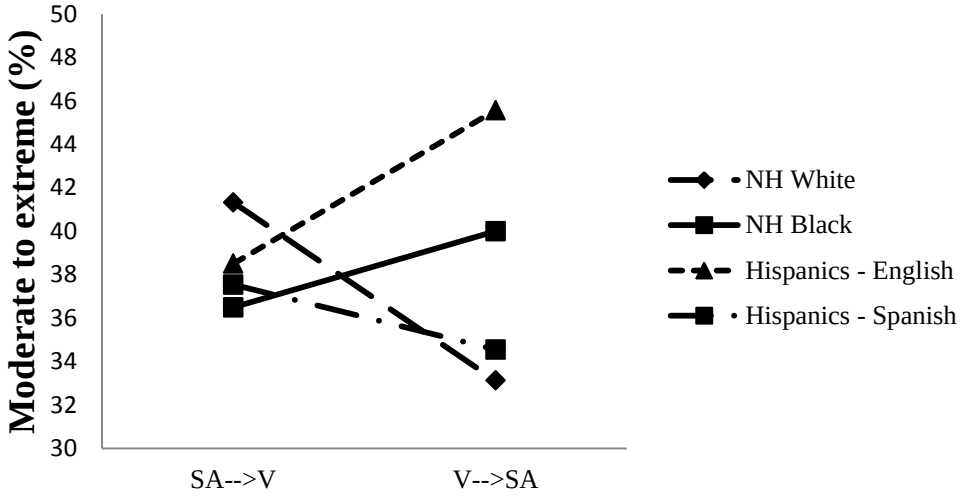


Figure 4. 3 Percentage of reported moderate to extreme pain by racial / ethnic groups and question order. SA, self-assessment question; V, vignette questions.

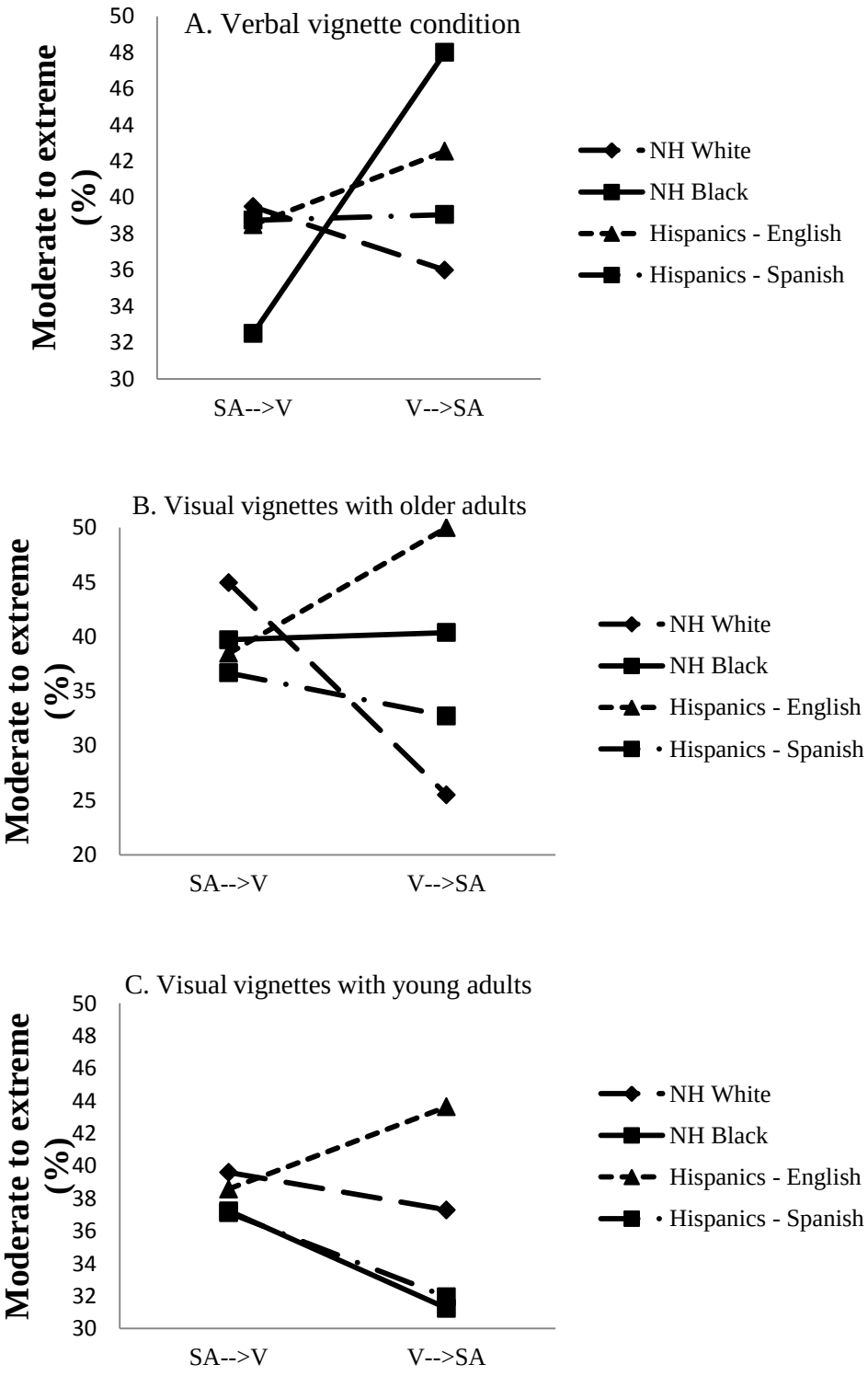


Figure 4. 4 Percentage of reported moderate to extreme pain by racial / ethnic groups and question order among those who were assigned into A) the verbal condition, B) older adults vignette condition and C) young adults vignette condition. SA, self-assessment question; V, vignette questions.

Sleep

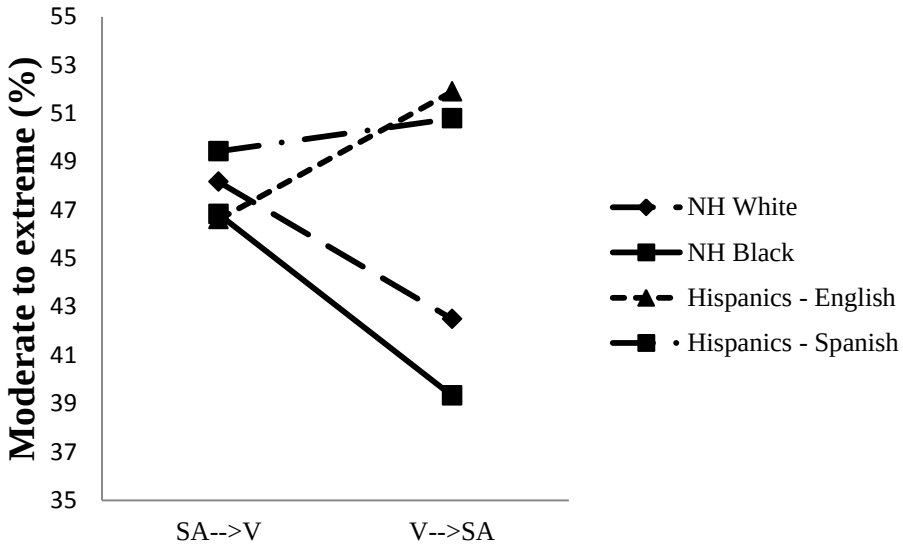


Figure 4. 5 Percentage of reported moderate to extreme sleep difficulties by racial / ethnic groups and question order. SA, self-assessment question; V, vignette questions.

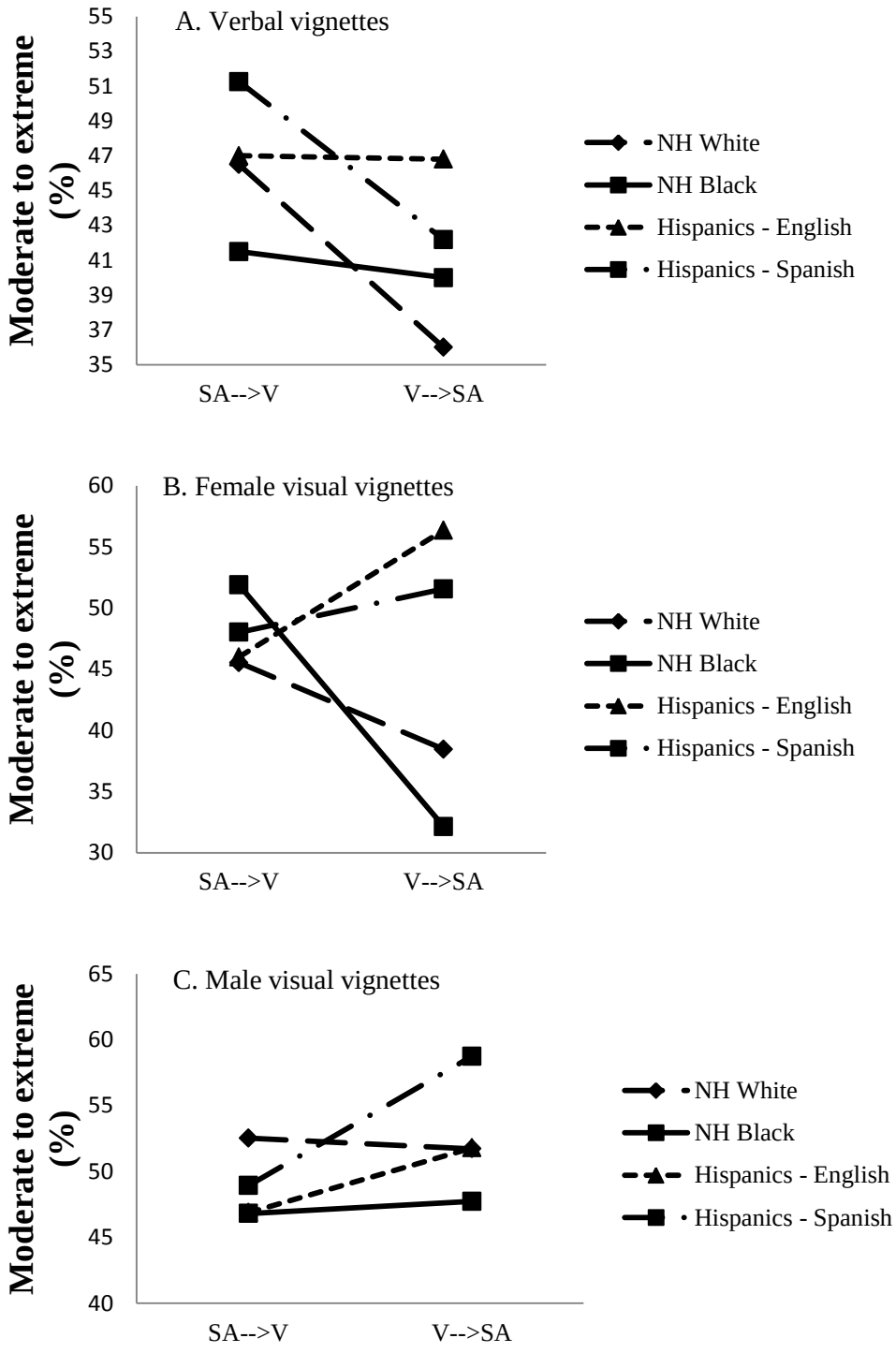


Figure 4. 6 Percentage of reported moderate to extreme sleep difficulties by racial / ethnic groups and question order among those who were assigned into A) the verbal condition, B) older adults vignette condition and C) young adults vignette condition. SA, self-assessment question; V, vignette questions.

Affect

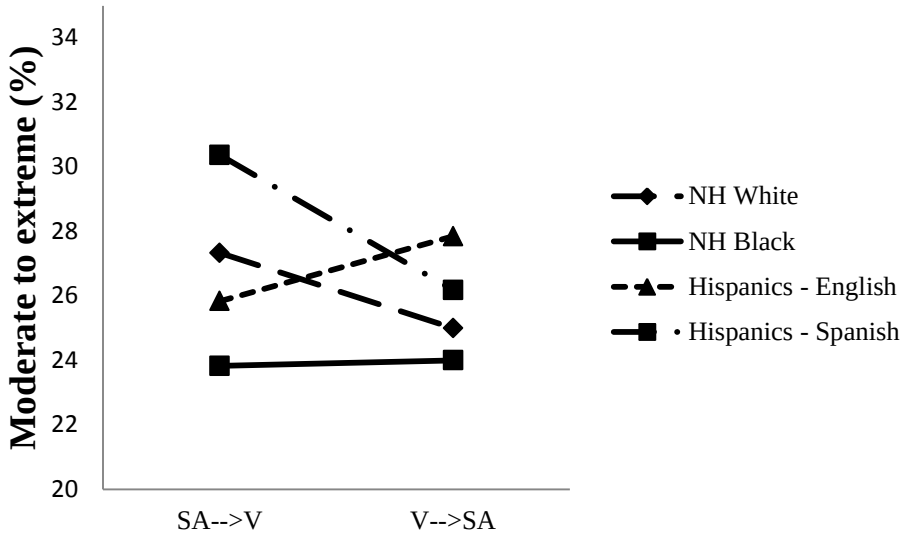


Figure 4. 7 Percentage of reported moderate to extreme affect problems by racial / ethnic groups and question order. SA, self-assessment question; V, vignette questions.

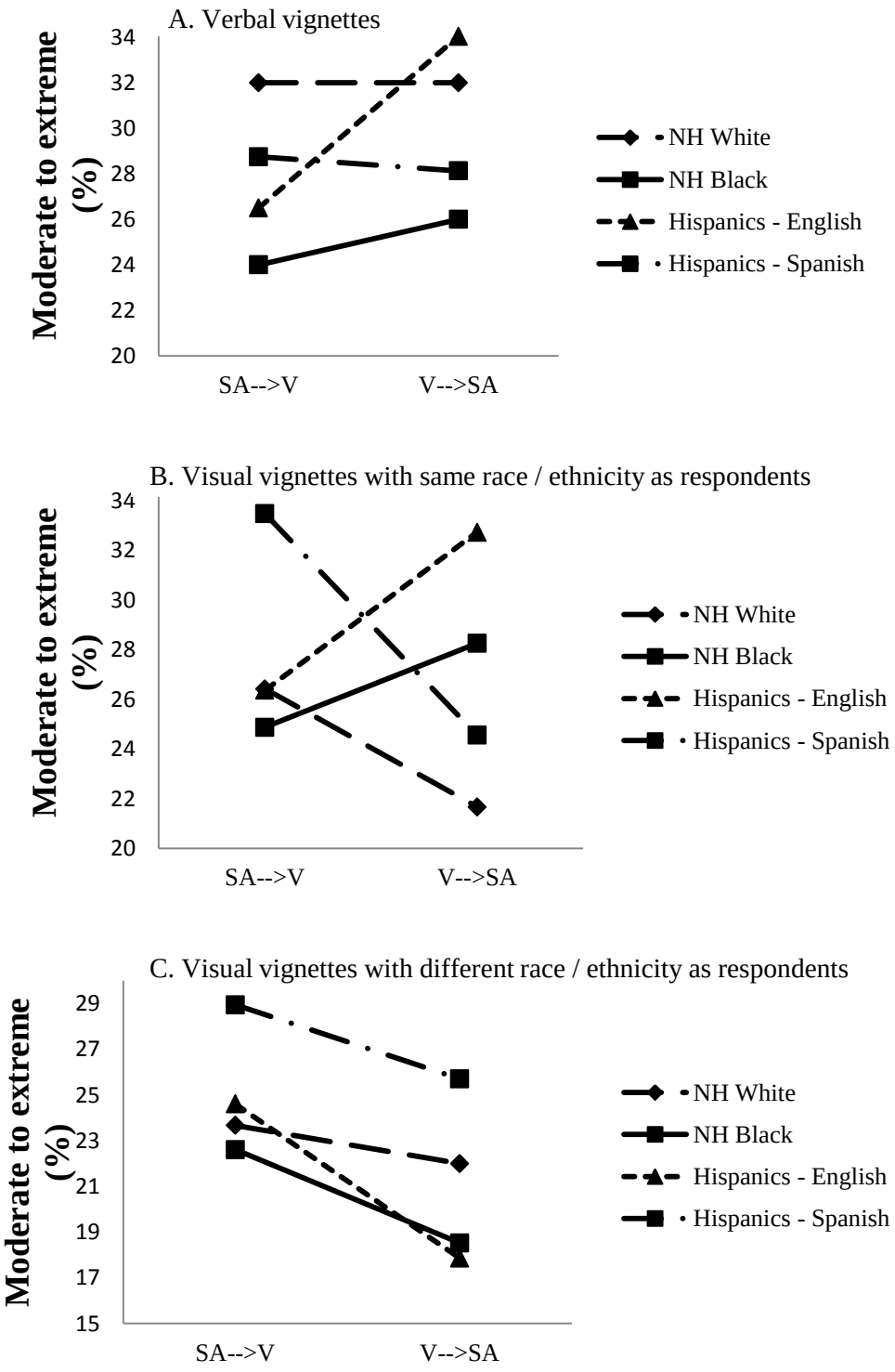


Figure 4. 8 Percentage of reported moderate to extreme affect problems by racial / ethnic groups and question order among those who were assigned into A) the verbal condition, B) older adults vignette condition and C) young adults vignette condition. SA, self-assessment question; V, vignette questions.

CHAPTER V

Conclusion

With the increasing popularity of comparative studies (e.g., cross-cultural and cross-national surveys), researchers are currently facing a difficult problem: respondents with different backgrounds often use different standards when answering survey questions, resulting in incomparability of responses across groups (or *Differential Item Functioning*, *DIF*). Among the many techniques to ameliorate the problem, *anchoring vignettes* is a method designed to calibrate responses to a common scale of measurement across groups. Despite its widespread use, there are several critical practical challenges, and how well it improves comparability across groups is not assured. The purpose of this dissertation was to examine one contemporary approach of using verbal anchoring vignettes to address DIF and to evaluate alternative methods and designs of anchoring vignettes. This dissertation had three primary objectives. The first objective was to evaluate how well a recently-proposed multidimensional IRT model approach using anchoring vignettes can control for response styles in interpersonal and cross-cultural comparisons (Chapter 2). The second objective was to evaluate the use of visual vignettes as an alternative design to current verbal vignette designs (Chapter 3). To date, no studies have examined the use of images for anchoring vignette questions. The third objective was to examine the survey context effects (specifically question order of self-assessment question relative to vignettes) on self-assessed health (Chapter 4). The ultimate goal was to provide some practical evidence as well as scientific contributions to improving future designs of anchoring vignettes.

Chapter 2 tested the validity of the multidimensional IRT model approach proposed by Bolt, Lu, and Kim (2014). More specifically, it uses several objective benchmarks to evaluate the validity of this approach by comparing the model with vignette questions (to adjust reporting errors) and the model without vignette questions. The first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) is selected for analysis, including data from 8 countries. The findings from this study indicate that adding the vignettes does not necessarily bring latent health closer to the objective benchmarks. In other words, the use of anchoring vignettes in this multidimensional IRT model does not effectively control for reporting errors in the SHARE data. This may be due to the violations of vignette measurement assumptions – it was found that for a given vignette person depicting health domains, those with better health tend to rate the vignette as having worse health than those with poor health. It seems that respondents use themselves as a reference point in evaluating others. These findings highlight the importance of constructing good vignettes that fulfill the assumptions. Future studies can further evaluate how well the model controls for response styles under the violations of RC and / or VE using simulation studies.

Chapter 3 examines the use of visual vignettes as an alternative design to current verbal vignette designs for four health domains – pain, mobility, sleep and affect. The visual vignettes were developed mainly out of the concern about the practical challenges in the use of verbal vignettes, specifically the increased respondents' cognitive burden and survey time due to the lengthy descriptions. To compare the performances of verbal and visual anchoring vignettes, this study conducted a web survey experiment, collecting data from various racial / ethnic groups (i.e., non-Hispanic (NH) white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic). The result shows that well-designed visual vignettes can greatly reduce survey time

and respondents' burden without the loss of the DIF-adjusting quality. The findings indicate that the vignette equivalence (VE) assumption is violated for both verbal and visual vignettes. This implies that designing "universal" anchoring vignettes that reveal the same information to every respondent remains a challenge for both verbal and visual vignettes. This is the first study that examined the use of visual vignettes. It shows the great potential of using visual vignettes to improve efficiencies in the anchoring vignette designs. Future research can expend this study to other cultural groups and evaluate other visual vignette designs such as using short videos.

Chapter 4 studies the question order effects (specifically the order of self-assessment question relative to vignettes) on self-assessed health measures. Researchers believe that vignettes, when asked first, can prime respondents to understand the self-assessment question and response scales as intended and improve the validity of the anchoring vignette method (Hopkins & King 2010). This may not hold in the following circumstances. First, when vignettes are asked to only a small random subsample (e.g., in HRS, CHARLS and SHARE) but used to correct DIF for all respondents in the survey, question order effects can be imposed to only to the selected subsample, leading to incomparability of self-assessments between this subsample and the rest respondents. Second, if the priming effects of vignettes on self-assessments differ by population groups, such as cultural groups or interview language used, it may introduce additional source of measurement error to the comparisons groups and affect the DIF-adjusting. This study aims to evaluate 1) whether there are any question order effects on the self-assessed pain, sleep, mobility and affect, and are some domains more susceptible to question order effects, 2) for each domain, whether the question order effects differ by vignette types (i.e., verbal vs. visual vignettes) and 3) whether the question order effects differ by racial / ethnic groups. The results suggest that the priming effects do not always occur for every health domain. It appears

that more easily observable domains like mobility are more susceptible to question order effects. Both verbal and visual vignette conditions may be susceptible to question order effects (e.g., for mobility domain), while verbal vignettes seem to have more priming effects on self-reported mobility. This may be because in verbal condition, with the increased response time, respondents had more time to process the cues in vignette descriptions, retrieve relevant information and adjust their responses. Results for the question order effects by racial / ethnic groups indicate that question order effects could differ among racial / ethnic groups, suggesting that researchers and practitioners need to be cautious of the context effects on self-assessments when using the *vignette first, self-assessment last* order, especially in cross-cultural studies. In future research, I will further evaluate how well the vignette methods work for different order conditions. Future research can expand the current work and evaluate other health domains (such as cognition and vision) and other racial / ethnic groups (e.g., Asians).

This dissertation presented a first attempt to examine several previously unexplored issues related to anchoring vignette designs and DIF-adjusting. From a methodological perspective, the results of this dissertation create basic theoretical knowledge regarding the impact of different anchoring vignette designs on correcting DIF. This dissertation also provides directions for future designs of visual anchoring vignettes. From a practical perspective, the results will help practitioners from various fields (e.g., epidemiology, public health and clinical research) to make better survey design decisions. Given the wide use of anchoring vignettes in cross-cultural surveys, the results of this dissertation also shed light on improving the comparability in future cross-cultural survey designs.

References

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*(681 Icc), 201–222.