Improving Anchoring Vignette Methodology in Health Surveys with Image Vignettes

Mengyao Hu[1], Sunghee Lee[1], Hongwei Xu[2],
Roberto Melipillán[3], Jacqui Smith[1] & Arie Kapteyn[4]

[1] *University of Michigan-Ann Arbor*
[2] *Queens College, New York*
[3] *Universidad Del Desarrollo, Chile*
[4] *University of Southern California*

Abstract

The anchoring vignette method is designed to improve comparisons across population groups and adjust for differential item functioning (DIF). Vignette questions are brief descriptions of hypothetical persons for respondents to rate. Although this method has been adopted widely in health surveys, there remain challenges. In particular, vignettes are complex, increasing survey time and respondent burden. Further, the assumptions underlying this method are often violated. To overcome such challenges, this paper introduces an innovative technique, namely image anchoring vignettes, conveying vignette information with varying health levels in images. We conducted a cross-cultural experimental study to examine the performance of image and standard text vignettes in terms of response time, how well they satisfy the assumptions, and their DIF-adjusting quality using a confirmatory factor analysis. The study revealed that respondents can better differentiate the intensity levels of the three vignettes in the image vignette condition, compared to text vignettes. Response consistency assumption appears to be better satisfied for image vignettes than text vignettes. Using well-designed image vignettes greatly reduces survey time without losing the DIF-adjustment quality, indicating the potential of image vignettes to improve overall efficiencies of the anchoring vignette method. Improving vignette equivalence (i.e., minimizing different interpretations of vignettes by different groups), remains a challenge for both text and image vignettes. This study generates new insights into the design and use of image anchoring vignettes.

Keywords: Differential item functioning; Anchoring vignettes; Image vignettes; Cross-cultural comparisons; Self-assessments of health

Self-assessed questions on health are good predictors for mortality and morbidity (Idler & Benyamini, 1997; DeSalvo et al., 2005). Self-assessment health questions often use Likert-type rating scales to measure respondents' attitudes, knowledge, perceptions, and behavior (Krosnick & Abelson, 1992; Lee, Jones, Mineyama, & Zhang, 2002). Ideally, responses obtained from these questions reflect only respondents' true state. This, however, is not always the case. In fact, answers to self-assessment questions reflect both respondents' true state and how they use the scales, a phenomenon known as response-category differential item functioning (DIF) (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007). As described in King and Wand (2007), DIF refers to situations when respondents from different backgrounds map the same state onto the scales in different ways.
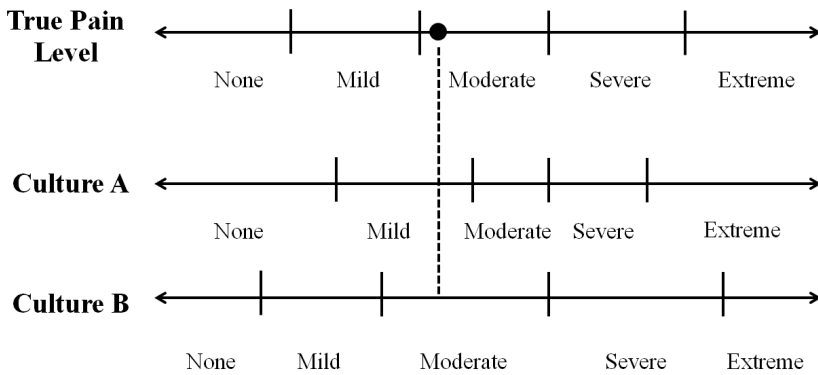
Figure 1 (adapted from Hu, Lee, & Xu, 2018) illustrates a cross-cultural study example of DIF to a self-assessed pain question on an ordinal response scale from "None" to "Extreme". In this example, cultural groups, A and B, use different cut points for a given response category. Assume that two respondents, one from A and one from B, have the same true pain level, both falling on the vertical dashed line. Despite their identical pain levels, the respondent from A will select "Mild," and the respondent from B will choose "Moderate". If this DIF is not accounted for, simple between-culture comparisons will erroneously conclude that the Culture B respondent experiences a higher level of pain (Hu et al., 2018).

An adjustment method for such DIF issues is to use anchoring vignettes (AV), which have been used in multiple national and international health surveys including the Health and Retirement Study (HRS) and the Survey of Health, Ageing and Retirement in Europe (SHARE). The AV approach typically involves two components: a self-assessment question and (typically multiple) anchoring vignette questions. First, respondents are asked to report their own status. For example, in a health survey, a typical self-assessed pain question is: Overall, in the last 30 days, how much pain or bodily aches did you have? The second component consists of vignette questions, each in a few sentences describing a hypothetical person's situation related to the construct measured, and respondents are asked to rate the vignette person. For example, a vignette used in HRS asks, "*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 of a problem did Paul have with bodily aches or pains?*". Usually, more than one vignette question describing varying intensity levels of the measured construct (e.g., low, moderate, and high levels) are asked (see Appendix 1). The vignette ratings can serve

_____

Direct correspondence to
    Mengyao Hu, Survey Research Center, Institute for Social Research,
    University of Michigan, 426 Thompson St., Ann Arbor, MI, USA
    E-mail: maggiehu@umich.edu

Figure 1   DIF for cross-cultural studies. Adapted from Hu, Lee & Xu (2018). The horizontal lines with arrows indicate the continuous scales of the domain (pain level). The short vertical lines indicate the cut points respondents used to answer the self-assessment question. The vertical dashed line indicates respondents responses to self-assessment questions. If a respondent's pain level falls on that line, it indicates that they have the same true pain level.

as benchmarks for the actual unobserved self-assessed pain level that researchers intend to measure.

The successful use of anchoring vignettes depends on two key assumptions: response consistency (RC) and vignette equivalence (VE). RC requires respondents to rate vignette persons in the same way as they would rate themselves (King et al., 2004). VE assumes that vignette descriptions are perceived similarly across respondents (King et al., 2004), essentially requiring vignettes to provide the same stimuli across respondents.

Promises and Pitfalls of the Current Anchoring Vignette Approach

Anchoring vignettes (AV) have been reported in many studies as a promising tool to correct for DIF (e.g., Mojtabai, 2015; Murray et al., 2002). Despite its promise, studies of the effectiveness of the standard AV (which rely on verbal descriptions of the vignette persons) have yielded mixed results. While some studies have found that text vignettes can effectively correct for DIF (Dowd & Todd, 2011; Van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011), other studies have reported that text vignettes do not necessarily provide comparable results among population groups (e.g., Grol-Prokopczyk et al., 2015). Previous studies have also shown that RC and

VE assumptions can be violated in different domains (Bolt, Lu, & Kim, 2014; Ferrer-i-Carbonell, Van Praag, & Theodossiou, 2011; Kapteyn, Smith, Van Soest, & Vonková, 2011; Rice, Robone, & Smith, 2012).

The assumption violations are likely due to several practical challenges related to the AV design [see also Hu and colleagues (2018)]. The first and most obvious challenge concerns question difficulty (Hopkins & King, 2010). Unlike typical survey questions that ask respondents to rate their own status, AV require respondents to imagine hypothetical persons based on verbal descriptions and to shift their focus from themselves to rate the status of these imagined hypothetical persons, placing greater cognitive burden on respondents. The second challenge is a substantial increase in survey time. Given that vignettes are designed to describe hypothetical situations, one single vignette often contains much more text than other typical survey questions (Hu and Lee, 2016). In addition, because usually more than one vignette is used per domain (e.g., pain), the use of AV may require a nontrivial amount of response time (Hirve et al., 2013; Hopkins & King, 2010; King et al., 2004). Third, the use of AV in cross-cultural research raises yet another issue with text vignettes: measurement inequivalence, where respondents with different cultural background may understand vignette descriptions in systematically different ways. One source that can lead to measurement inequivalence is questionnaire translation. Poor translation can directly influence respondents' interpretation of the vignettes, leading to violation of the VE assumption. Another critical challenge is the specific content to include in vignette descriptions. As acknowledged by Kapteyn et al. (2011), it is difficult to write vignette descriptions that are as "comprehensive" as what respondents know about their own state (Kapteyn et al., 2011). This indicates that respondents may rate themselves using criteria different from those they use for vignettes, resulting in violation of the RC assumption. VE can also be violated if respondents interpret the vignette descriptions in different ways. The potential for this problem is even greater in cross-cultural research where the challenges of designing equivalent and comparable vignettes are increased.

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

Image Anchoring Vignettes

As a potential remedy to the limitations of text AV, we propose in this study to use visual AV with well-designed and carefully-selected images, i.e., image vignettes. With the technical development of internet, image vignettes have gained increasing popularity in survey research, especially in studying attitudes and sensitive questions (Naylor et al., 2014; Groot et al., 2020). To the best of our knowledge, this study is the first research that incorporates visual methodology with AV techniques.

Mechanisms of information processing of visual vs. verbal stimuli have been discussed in previous studies but there are no consensus conclusions. Some studies report similar processing of visual and verbal information in "a functional unitary system that is directly accessed by both visual objects and words" (Caramazza, 1996). In contrast, some other studies have shown that visual and verbal information are processed differently and "creating separate semantic representations" (Glaser, 1992; Glaser & Glaser, 1989; Schlochtermeier et al., 2013). For example, information processing of images is reported to be connected to activation of the right brain hemisphere (Grady et al., 1998; Naspetti et al., 2016), and activation of the left hemisphere is found to be associated with text information processing (Sevostianov et al., 2002). Despite the inconclusive results of the mechanisms of information processing, a common finding reported in previous studies is the "processing superiority" of images as compared to text information (Azizian et al., 2006, Schlochtermeier et al., 2013). As reported in Schlochtermeier et al. (2013), images lead to faster and a more direct access to meaning. In comparison, texts require "additional translational activity at the representational level" to access the semantic system (Schlochtermeier et al., 2013).

Given the reported processing superiority of image processing, the image AV strategy may lead to several potential advantages. First, images may require less cognitive effort to process than do text descriptions. 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 AV (which require imagining hypothetical persons), the use of images is advantageous for both low-literacy respondents and those who are unable to create mental images based on text vignettes. For these respondents, the saying "A picture is worth a thousand words" is particularly relevant considering the challenge of reading through the lengthy text descriptions to understand the vignette scenario (Hibbing & Rankin-Erickson, 2003).

In addition to ease of understanding, because respondents can process information shown in image vignettes relatively quickly, we expect that the use of image vignettes will reduce respondents' cognitive burden and overall survey time. In turn, these two aspects could contribute to improving survey data quality by reducing survey break-offs and respondents' satisficing behavior.

A second potential advantage of image vignettes is that they might help satisfy the measurement assumptions. For example, it has been found that first names used in text 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 may be of less concern in well-designed image vignettes where

the physical characteristics of the vignette person are clearly presented, limiting the possibility of different interpretations. Note that the performances of image vignettes can largely depend on how they are designed. Some design features may be associated with different interpretations of the vignette person, e.g., respondents with different age and gender may view a vignette person with tattoos, piercings, and unnaturally colored hair differently. While it is true that not all image vignettes will help satisfy the measurement assumptions, in this study, we aim to investigate: with carefully designed image vignettes on health domains, whether image vignettes could help with measurement assumptions, compared to text vignettes.

Because there are no prior studies on the use of image anchoring vignettes, it remains an open question whether this approach can remedy limitations of current text vignettes. To fill this gap, this paper aims to evaluate the use of image AV as an alternative to text vignettes and to compare the performance of image and standard text vignettes in terms of response time, how well they satisfy the RC and VE assumptions, and their ability to reduce measurement errors in a confirmatory factor analysis (CFA) framework. In this paper, we 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). We have three research questions (RQ).

RQ1: Will image AV reduce response time, compared to text AV? This research question will be addressed by analyzing survey time associated with text and image AV using time stamp data.

RQ2: Will image AV better meet AV measurement assumptions compared to text AV? This research question will be addressed by examining both VE and RC assumptions for text and image AV.

RQ3: In a confirmatory factor analysis (CFA) framework, a.) we will investigate whether a model of latent health based on image or text AV-adjusted scores will show better fit compared to a model based on unadjusted self-reported scores, and b.) whether a model based on image AV-adjusted scores will have similar or better fit compared to a model based on text AV-adjusted scores, i.e., will image AV adjustment achieve similar or better measurement error-reduction, compared to text AV?

Methods

Design of Image Vignettes

Prior to designing the image vignettes, we established criteria for image selection or creation. A three-step approach was used to develop these criteria: specifically, we 1) thoroughly examined critical elements of the four health domains, 2) identified common elements applicable across groups (e.g., arm pain) based on the litera-

ture review, and 3) based on the elements identified, we selected or designed images with these elements at different intensity levels for each domain (e.g., from no pain to extreme pain). Based on the developed criteria, images were then selected from commercial websites of images and photos (e.g., www.istockphoto.com/). In situations where, for a given health domain, no images meeting the criteria were found on those websites, we 1) recruited volunteers from different platforms (e.g., friends or family members) to serve as models in the photos, 2) obtained each volunteer's consent to take a photo and to use it in this study, and 3) took the photo and edited them. To remove potential confounding effects of various image elements, such as background, size, resolution, and color balance, the selected images or photos were further edited by students with expertise in image-editing.

The ultimate goal of the image vignette design for the current study was to have three well-designed image vignettes per domain. For the purpose of selecting the most comparable images across cultures, we first designed six images for each characteristic: two images for each intensity level (e.g., two no/low pain, two moderate pain and two extreme pain vignettes) per design condition, and eventually selected three out of the six for each condition in the pretest. The selected images (see Appendix 1) were then used in the web survey experiment as described below[1].

Pretesting

The pretest was conducted through Amazon Mechanical Turk (MTurk), where we 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 were 18 or older. A $0.45 incentive was offered for each completed survey. To recruit respondents of all age groups, toward the end of the data collection, we posted a HIT open only to older respondents with the same incentive. In total, 201 respondents completed the pretest survey, about half of them aged 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 World Health Organization (WHO) in their pretesting of anchoring vignettes (Murray et al., 2003). For the two sets of image options, the image with the higher correct ranking rate (the percentage of respondents who correctly

---

1   In designing image vignettes, two different conditions (e.g., male and female) were designed for each domain. Respondents assigned to the image vignette conditions were randomly assigned to the two design conditions. This paper focuses only on the comparison between text and image vignettes, and evaluations on how image vignette design features influence anchoring vignette methodology are discussed elsewhere.

ranked the vignette series) was selected. The final correct ranking rates ranged from about 80% to 97% across all health domains.

Web Survey Procedure

The main data collection was based on a web survey using a non-probability online panel. Respondents from four different racial/ethnic groups – Non-Hispanic (NH) white, NH black, English-speaking Hispanic and Spanish-speaking Hispanic – were recruited through Qualtrics' online survey panel, which partners with over 20 Web-based panel providers to supply diverse, quality respondents (more information about Qualtrics survey panel, see also Holt & Loraas, 2019; Ibarra et al., 2018). The reason for including these groups is that race/ethnicity and language are proxies of cultures (Davis et al., 2019; Lee et al., 2014; Lee et al., 2017) and are known to influence respondents' self-reporting of their health status (McCarthy, Ruiz, Gale, Karam, & Moore, 2004; Lee et al., 2014). For example, Hispanics have been shown to conceptualize health differently than non-Hispanic Whites as they "include non-medical aspects, such as spiritual and social wellbeing, in addition to medical conditions that non-Hispanic Whites consider the most critical element for assessing health" (Lee et al., 2014). Language can also influence respondents' reporting of their health status, e.g., Lee and colleagues examined Hispanics' self-reported health by interview language and found that the difference was primarily due to Hispanics interviewed in Spanish (Lee et al., 2014). Respondents from each racial/ethnic group were randomized into three conditions: the standard text vignette condition and two image vignette conditions that differed in the vignette persons' characteristics (See Appendix 2 for a flowchart of the experimental conditions and assignments). Robustness of randomization was examined, and results show that there are no significant socio-demographic differences across the experimental conditions (Supplemental Table 5), suggesting that the randomization works well.

For the text vignette condition, we adapted the text vignette descriptions from those widely used in many major surveys (e.g., HRS). Each domain had a series of three vignettes, describing different intensity levels of the measured construct: low, moderate and high (e.g., from least to most pain). For the image condition, we used the image vignettes designed and selected in the pretest with three vignettes per condition, depicting three levels of difficulty/intensity of symptoms in each domain (see Appendix 1). The introduction to the vignette questions also followed the standard approach used in earlier surveys such as HRS. We randomized the order of the domains and of the three vignettes per domain presented to respondents in order to isolate question order effects. Besides self-assessment and vignette questions, the study also included responses to objective questions regarding these health

domains, time stamp data, and respondents' demographic and socio-economic information.

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 the Cross-Cultural Survey Guidelines developed by the survey research center at the University of Michigan (Mohler et al., 2016). Translation was conducted by the translation team of HRS. The translated questionnaire was then reviewed and tested by 20 bilingual speakers who are native Spanish speakers and are also fluent in English.

The online survey questionnaire was programmed in Qualtrics. The Qualtrics online panel team sampled respondents from their panel. Except for Hispanics speaking Spanish, around 750 respondents were sampled for each of the three other race/ethnic groups. Each of the three sampled subgroups had nearly equal proportions of 1) male and female, 2) below or equal to high school education and higher than high school education, and 3) respondents aged 18-49 or 50 and over. For Spanish-speaking Hispanics[2], 889 respondents were sampled with about 43% male respondents. Detailed information of the sample profile is presented in Table 1. In conducting this experiment, we implicitly make the stable unit treatment value assumption (SUTVA) that the outcome for one respondent is unaffected by the assignment of treatments to the other units. This assumption is likely to have been met in our study given Qualtrics' large pool of respondents and our duplicate check on respondents' IP addresses.

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 text condition and two image conditions.

---

2   Due to the difficulties in recruiting Spanish-speaking Hispanics, Qualtrics collected more respondents for this group in order to meet the targeted number of male Spanish-speaking Hispanics who were 50 and above and had education equal to high school or below.

Table 1   Respondents' characteristics.

	White (n=760) %	Black (n=750) %	Hispanic-English (n=750) %	Hispanic-Spanish (n=889) %
Male	50.39	50.00	50.00	42.52
Age				
Age 18 – 29	14.34	22.80	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	15.20	16.80	9.34
More than high school	49.47	50.00	50.00	57.82
Married	53.42	36.67	50.93	54.78
Employed	50.92	52.00	56.13	57.14
Income				
Income below $40,000	35.00	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	21.20	25.60	19.57

Analysis Strategy

We first examined the distributions of the self-assessment and vignette questions by vignette type for each domain descriptively. We then examined whether and to what extent the self-assessments were affected by DIF following previous literature studying measurement errors in self-assessed health (Yan & Hu, 2018). Specifically, since self-assessments of health are correlated with objective health conditions (Idler & Kasl, 1995), we take advantage of this relation to gain insights on how DIF affects respondents' uses of the scales. We constructed a measure of objective health for each domain using respondents' own answers to a series of factual questions asking about health conditions for each domain. We then standardized the number of health issues (e.g., the number of mobility issues) within each racial / ethnic group. The resultant standardized score reflects the number of standard deviations above or below the racial/ethnic subgroup mean, where a value of 0 stands for the subgroup average. Negative values of health scores denote better health than the subgroup average (i.e., respondents reported fewer health conditions) whereas positive values indicate worse health than the racial/ethnic subgroup average (i.e., respondents reported more health conditions). For each category selected

on the self-assessment question, we computed the mean of the standardized scores and compared them across different racial / ethnic groups.

We then examined RQ1 to RQ3 as described below. Note that in examining RQ1 to RQ3, the variables were not standardized.

RQ 1. To evaluate whether image vignettes can reduce survey time compared to text vignettes, we analyzed the survey time using time stamp data. The mean response time was compared between the text and image vignette types. To formally test the effects of vignette types on survey time, for each domain, we fit multilevel linear regression models with random intercepts. The log-transformed response time was used as the outcome, given that time is right skewed. In this model, Level 1 corresponds to vignette questions, and Level 2 corresponds to respondents. Level 1 covariate was vignette type (image vs. text vignettes) and Level 2 covariates included respondents' demographic and socio-economic variables. Results of the multilevel model can be found in Appendix 3 (Supplemental Table 6). Given that it is hard to ascertain whether respondents were completing the online survey from beginning to the end in one sitting or took temporarily breaks – e.g., checking emails and browsing other web tabs, we employed a two-step procedure to identify response time outliers. First, based on the response time distribution, we used 15 minutes (i.e., 900 seconds)[3] per vignette question as a threshold to identify those who might took a break during the survey completion. Second, we examined distributions of random effects and residuals of the multilevel models described above. Using histograms and Q-Q plots, outliers on these parameters were inspected visually. In total, the first step identified four response time outliers for pain domain, two outliers each for sleep and mobility domains and six outliers for affect domain were identified and excluded from this analysis. The second step did not identify any outliers.

RQ 2. We compared image and text vignettes in terms of how well they satisfy the two measurement assumptions – VE and RC. Below we describe approaches for each of the two assumption-testing.

RQ 2a (Test for VE). Two tests of VE were conducted. The first one is referred as correct rank ordering test, which examines whether respondents could correctly rank order vignettes based on their intensity level. Several previous studies refer to this test as a weak test for VE, stating that correct rank-ordering is a "necessary but not sufficient" condition for VE (e.g., Grol-Prokopczyk et al., 2015; Kristensen & Johansson, 2008), given that if VE is fulfilled through effective vignette design, respondents should agree on the ranking of the vignettes.

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, that respondent may rate the first two vignettes (low and moderate pain) or all

---

3   As a sensitivity analysis, we also performed the analysis with 5 minutes and 10 minutes thresholds to identify response time outliers, which gave consistent results.

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 as having similar intensity levels and rate them identically), this is unlikely given the differences among the intensity levels in the vignette design. Thus, here we only consider two kinds of ties: 1) ties between the first two vignettes (low and moderate intensity) and 2) ties between the last two vignettes (moderate and high intensity).

The second test for VE was a statistical test conducted following Grol-Prokop-czyk (2018). This method was first developed by d'Uva et al. (2011) and applied in many other studies (Grol-Prokopczyk, 2018; Grol-Prokopczyk et al., 2015; 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 dimen-sion should be the same for all respondents (d'Uva, Lindeboom, O'Donnell, & van Doorslaer, 2011). The test is based on a likelihood-ratio (LR) test of two nested models. Both models are variations of the hierarchical ordered probit (HOPIT) model. Below we list the key differences between the two models. The first model, Model (A)[4], predicts a respondent's perceived location of vignettes:

$$V_{ij}^* = \alpha_j + \varepsilon_{ij} \tag{A}$$

where V_{ij}^* is respondent i's perceived location of vignette j on the latent dimension, α_j is a constant term and ε_{ij} is the random error term that is assumed to be normally distributed with mean zero and variance one. For one of the vignettes in a domain (the reference vignette), α is set to 0 for model identification. The cut points (τ) for the vignettes are modeled in the same way as in the HOPIT model. Note that Model A does not include covariates to predict perceived vignette locations on the latent dimension. 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 the less restrictive Model B, a vector of covariates, X_i, is added to Model A to predict the perceived vignette locations. In this study, X_i includes marital status, employment status, age, gender, education, income level, and racial/ethnic group.

$$V_{ij}^* = \alpha_j + \lambda_j X_i + \varepsilon_{ij} \tag{B}$$

Since this model is not identified, one needs a normalization. For one of the vignettes (the reference vignette), both α and λ_j are set to zero for identification. If VE is satisfied, λ_j will be 0 for each j. Model A is nested in Model B and if VE is satisfied, the LR test will not reject Model A. If, however, the LR test rejects Model

---

4   In describing the models, we used the same notation as Grol-Prokopczyk & Carr (2017).

A, it indicates that respondents with different characteristics perceive the severity of the vignettes differently. The estimated coefficient vector λ_j will indicate which covariates are driving the violation of VE.

 RQ 2b (Test of RC). Our test of RC was conducted following Grol-Prokopczyk et al. (2015). This test was based on visual comparisons of two sets of predicted cut points. 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 based on Model C below, which uses objective health measures to predict the self-assessments.

$$Y_i^* = \mu + \beta W_i + \varepsilon_i \tag{C}$$

where Y_i^* is respondent i's true score on the latent dimension in the measured domain, μ is a constant term and ε_i is a random error term that is assumed to be normally distributed with mean zero and variance one. W_i is a vector of covariates consisting of the objective measures. The cut points are modeled in the same way as in Model A. The predicted mean cut points from the two models were then graphed in a figure for visual comparisons. The RC test basically compares the shape (Grol-Prokopczyk et al., 2015) of the two sets of cut points. A similar shape would indicate that respondents had similar standards when rating vignettes and rating themselves (RC). As mentioned in Grol-Prokopczyk (2018), this test can be viewed only as suggestive. The objective measures used in this 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[5], total pain index[6], 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).

 RQ 3. The self-assessments for the health domains have often been used in a confirmatory factor analysis (CFA) framework to measure latent overall health. To examine whether AV-adjustment can reduce measurement errors in self-assessments, following Weiss & Roberts (2018), we compared the model fit of the CFA using original responses with the CFA using text / image AV-adjusted scores. If the use of AV-adjusted scores can correct DIF, we would expect the models with AV-adjusted scores to have better fit (RQ 3a; see also Weiss & Roberts, 2018). To evaluate whether image AV can achieve similar or better DIF-correction compared to text AV (RQ 3b), we also compared the magnitude of improvement compared to CFA with original self-reports, for both image and text AV-adjustment.

---

5 The sleep quality score was constructed based on 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 are unable to fall asleep again.

6 The total pain index was constructed following Ray et al. (2009).

The AV-adjusted scores were calculated using the non-parametric approach, following previous literature (Wand et al., 2011). In situations where respondents have ties in their AV-rating or inconsistent AV orders from researchers' expected order (i.e., order violations), the non-parametric method will result in an interval instead of a number for these respondents. Following the recommendations in previous literature (Kyllonen & Bertling, 2014; Primi, Zanon, Santos, De Fruyt, & John, 2016; Weiss & Roberts, 2018), the lower bounds of the intervals are chosen as the adjusted scores for respondents with ties or order violations. Model fit criteria including Comparative-Fit-Index (CFI), Tucker–Lewis index (TLI), and a Root Mean Square Error of Approximation (RMSEA) and 90% confidence interval (CI) of RMSEA are used to compare the models (Schreiber et al., 2006). A CFI greater than 0.95 and a TLI greater than 0.95 are considered as acceptable model fit (Hu & Bentler, 1999). A RMSEA less than or equal to 0.05 is considered as good fit, and less than or equal to 0.08 is considered as moderate fit (MacCallum, Browne & Sugawara, 1996). For the 90% CI of RMSEA, ideally the lower value should be less than 0.05 and the upper value less than 0.08 (MacCallum, Browne & Sugawara, 1996; Schreiber et al., 2006).

Results

Descriptive Analysis

We first examined the distributions of the self-assessment and vignette questions by vignette type for each domain. Figure 2 shows the distribution for the pain domain. Similar patterns were found for other domains. As expected for a properly randomized design, for each domain, the distributions for the self-assessment questions do not differ by vignette type-text or image vignettes. Comparing vignette distributions by vignette type, in general, the intensity levels of the image vignettes can be better differentiated than those of the text vignettes.
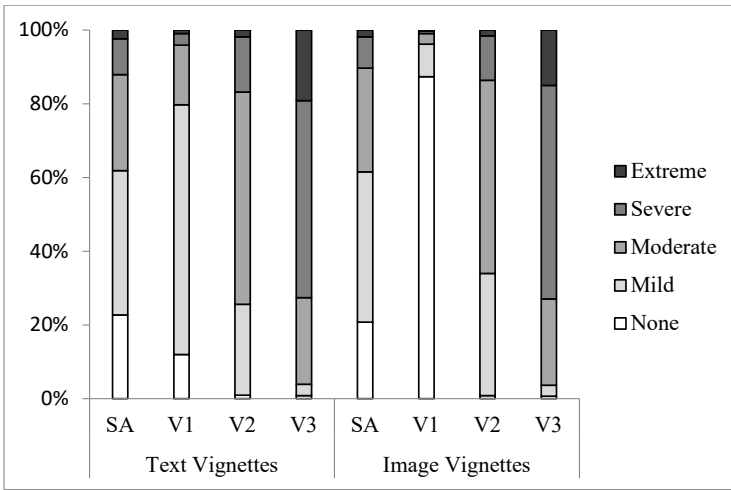
Figure 2 Responses to pain self-assessment (SA) and difficulty/intensity ques-
tions for three vignettes (V1 = none/mild; V2 = moderate; V3 =
severe/extreme).

DIF Evaluation

We then examined whether DIF was present in the self-assessments[7]. Figure 3
displays the mean standardized number of mobility issues by reported response
categories of self-assessed mobility. For all four racial / ethnic groups, the mean
standardized scores are negative for those who selected "none" for mobility, and
positive for those who selected "mild" or "extreme" mobility issues. For White
respondents, the biggest increase of the mean standardized score occurs between
"Moderate" and "Severe", while the change of the score from "Severe" to "Extreme"
is much smaller. Compared to White respondents, for Black and Hispanic speaking
Spanish, the change of the mean scores from "Moderate" to "Severe" is similar to
change from "Severe" to "Extreme". Note that for Hispanics speaking English, the
mean score is lower among those who select "Extreme" compared to those who
select "Moderate" or "Severe", while for all other groups, the standardized score
increases as the severity of the response categories increase. This indicates that
respondents from different racial / ethnic groups use the scales differently, leading
to DIF, and indicates the need to use methods like anchoring vignettes to achieve
cross-cultural comparability.

---

7   We examined DIF across race/ethnicity and other socio-demographic groups, includ-
    ing gender, education and marital status. DIF were found across race/ethnicity groups
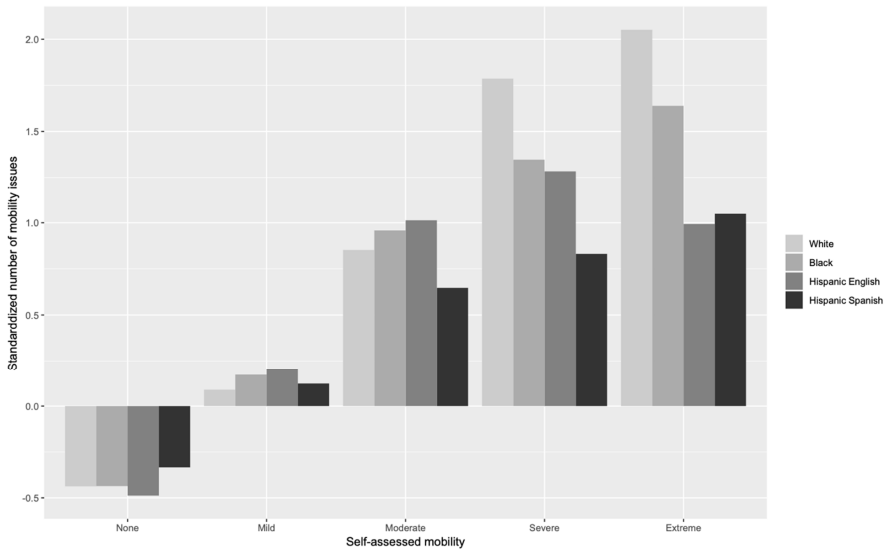   but no other socio-demographic groups.

Figure 3 Mean standardized number of mobility issues by reported response
categories of self-assessed mobility.

RQ 1. Response time

As shown in Table 2, regardless of domain, the average time respondents spent on
a text vignette question is about twice as long as time spent on an image vignette
question. Results for the statistical test of differential response time by vignette
types using multilevel models are presented in Appendix 3 (Supplemental Table 6),
which show consistent results as Table 2.

Table 2   Average time (in seconds) spent on one text or image vignette question
by health domains.

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

RQ 2a. VE Test

Results of two tests of VE, the correct rank ordering test and the VE statistical test, were presented below.

   Correct Rank-Ordering. Table 3 shows the percent 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 around 82%, depending on the domain. It is noted that for each of the four domains, the percentage of consistent rankings is significantly higher for the image than for the text vignette condition. In other words, respondents assigned to the image conditions are more likely to agree on the rank order of the vignettes than those assigned to the text condition. Respondents seem to have difficulty differentiating the rank orders of sleep and mobility *text vignette*s, with less than 20% able to correctly rank vignettes for these domains[8]. We also formally tested the effects of vignette types on the rank ordering of vignettes by fitting logistic regression models for each health domain (Results not shown). Not surprisingly, the odds of correctly ranking vignettes in the image vignette conditions are significantly higher compared to those in text vignette conditions. This is consistent across all four domains. Similar results were found when allowing for ties.

   Statistical test of VE. Table 4 presents the results of statistical test of VE. The VE assumption is rejected in almost all conditions, except for the sleep text vignettes.

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

	Pain		Sleep		Mobility		Affect	
	n	%	n	%	n	%	n	%
Text vignette	1051	47.6	1051	17.7	1051	19.8	1051	67.1
Image vignette	2098	79.7	2098	74.0	2098	43.4	2098	81.8

Table 4   Likelihood ratio tests of vignette equivalence.

	Pain		Sleep		Mobility		Affect	
	df	LR Test	df	LR Test	df	LR Test	df	LR Test
Text vignettes	24	70.4***	24	24.4	24	55.1***	24	110.9***
Image vignette	24	137.4***	24	158.8***	24	67.1***	24	154.3***

*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

---

8   This analysis was also performed when two tie situations were allowed: 1) ties between the first two vignettes (low and moderate intensity) and 2) ties between the last two vignettes (moderate and high intensity). Results of rank order test allowing ties are consistent with Table 3.

Table 5 presents the results for predicting vignette locations (i.e., where it lies on the latent health spectrum) for both text and image vignette conditions of each domain[9]. In Table 5, Vignette 3 is the reference vignette, the one describing the highest pain level. Gender, marital status and racial/ethnic groups are the main predictors that drive the violations of VE for pain text vignettes. As for pain image vignettes, gender, age, income, and racial/ethnic groups are the main predictors that drive the violations of VE.

Those who are married view the first pain text vignette (the vignette with the least pain) as further away from the reference vignette on the latent spectrum, with a positive coefficient of 0.31 ($p = 0.02$). In other words, married respondents view the first pain text vignette as depicting better health (or less pain) than those who are not married. Males view the first pain text vignette as depicting worse health (or more pain) than females, which is consistent for both text and image AV conditions. Note that racial/ethnic group differences are significant for all health domains, suggesting that respondents from different racial/ethnic groups view the vignettes differently. For example, Hispanics interviewed in Spanish view Vignette 1 as depicting more pain than White respondents, regardless of text or image vignette designs.

As shown in Table 5, racial/ethnic group is a predictor that drives violations of VE for all health domains. To further examine this, Figure 4 presents the estimated vignette locations relative to the reference vignette by racial/ethnic group and vignette type for each health domain. If VE is satisfied, 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 5 and Figure 4. As shown in Figures 4A1 and 4A2, Hispanics who completed the Spanish-language survey view the first vignette person (least severity) as having more pain (i.e., closer to 0 line, the reference vignette with the highest severity) compared to White respondents. On the other hand, Hispanics who completed the English-language survey also view the first vignette person as having more pain than do White respondents under the text condition, but not under the image vignette condition. Similar results are found for the affect domain (see Figure 4D1 and 4D2).

Figures 4B1 and 4B2 shows the estimated vignette locations for the sleep domain. As can be seen from Figure 4B1, 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 worth noting that the perceived vignette location for the second vignette is not significantly different from the reference vignette, suggesting that the sleep text vignettes failed to provide a good distinction between the second and third vignettes. As
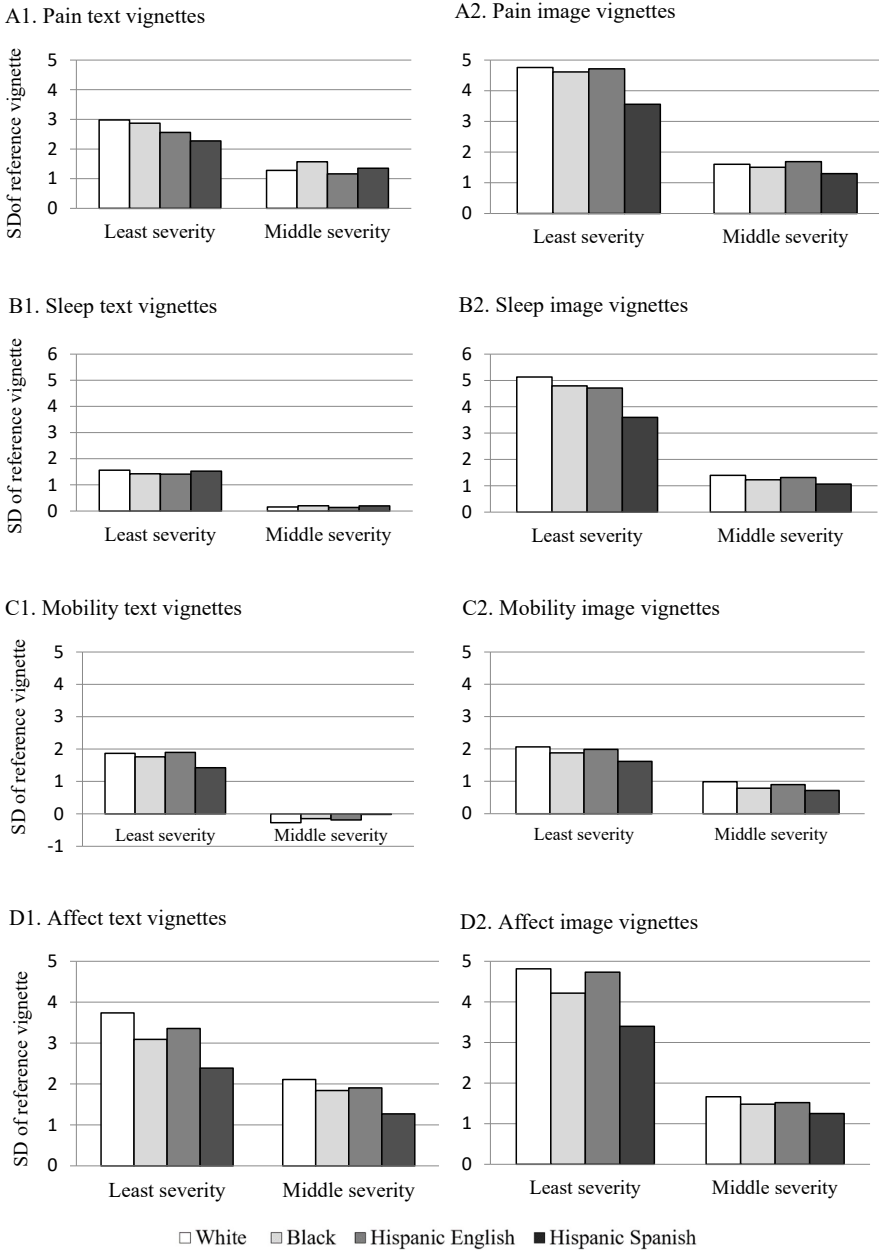
---

9   As a sensitivity analysis, we also fit models combining image and text vignettes in one model for each domain (i.e., treating vignette type as a predictor in the model). Results (shown in Appendix 4) suggests image vignettes perform better in distinguishing the intensity levels of the three vignettes for each domain.

Table 5 Predictors for perceived vignette locations on the latent health spectrum.

	Pain		Sleep		Mobility		Affect	
	Text	Image	Text	Image	Text	Image	Text	Image
Vignette 1 (no/mild difficulty/intensity)								
Constant	3.20***	5.12***	1.48***	5.51***	2.03***	2.04***	4.01***	5.72***
Married	0.31*	0.24	0.02	-0.13	0.27	0.06	0.37**	0.23
Male	-0.49***	-0.36**	-0.29**	-0.15	-0.30**	-0.05	-0.58***	-0.23
Employed	-0.1	0.00	0.06	0.20	-0.05	0.18*	0.06	0.15
More than high school	-0.09	0.09	0.16	0.22	0.02	0.18**	0.10	-0.24
Age 18 - 29	0.25	-0.57*	0.14	-0.53*	-0.08	-0.08	-0.05	-0.90***
Age 30 - 49	-0.17	0.01	0.14	-0.10	-0.10	-0.06	0.07	-0.69**
Age 50 - 64	0.09	-0.50*	0.15	-0.29	0.08	-0.10	0.05	-0.88***
Middle income	0.04	-0.28*	0.05	-0.43**	0.03	-0.02	-0.24	-0.33*
High income	-0.15	-0.15	-0.02	-0.29	-0.34*	-0.32**	-0.61**	-0.41*
Black	-0.12	0.02	-0.17	-0.25	-0.13	-0.19	-0.65**	-0.51**
Hispanic (English)	-0.47**	0.04	-0.21	-0.38	-0.04	-0.09	-0.46*	0.00
Hispanic (Spanish)	-0.82***	-1.21***	-0.13	-1.56***	-0.52**	-0.50***	-1.46***	-1.34***
Vignette 2 (moderate difficulty/intensity)								
Constant	1.38***	1.84***	0.01	1.43***	-0.43*	0.97***	2.48**	1.88***
Married	0.09	0.01	0.03	-0.15	0.20*	-0.03	0.19	0.04
Male	-0.11	-0.19*	0.01	0.04	0.03	-0.03	-0.40**	0.02
Employed	-0.03	0.03	0.14	0.01	0.06	0.15*	0.15	0.07
More than high school	-0.12	0.03	0.02	0.05	-0.01	0.20**	-0.08	-0.09
Age 18 - 29	0.13	-0.20	0.11	0.14	0.01	0.10	-0.17	-0.32*
Age 30 - 49	-0.06	-0.11	0.12	0.24	0.02	-0.01	-0.22	-0.20
Age 50 - 64	0.01	-0.24	0.10	0.04	0.12	-0.19	-0.18	-0.22
Middle income	0.06	-0.09	-0.06	-0.16	0.09	-0.07	-0.18	-0.05
High income	-0.12	0.00	-0.06	-0.21*	-0.19	-0.18	-0.31	-0.12
Black	0.27	-0.05	0.03	-0.19	0.12	-0.22*	-0.25	-0.16
Hispanic (English)	-0.13	0.12	-0.03	-0.08	0.06	-0.11	-0.22	-0.13
Hispanic (Spanish)	0.02	-0.30**	0.02	-0.35**	0.20	-0.31**	-0.87***	-0.39***

Notes: Vignette 3 (highest difficulty/intensity) is the reference vignette. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

A1. Pain text vignettes

A2. Pain image vignettes

B1. Sleep text vignettes

B2. Sleep image vignettes

C1. Mobility text vignettes

C2. Mobility image vignettes

D1. Affect text vignettes

D2. Affect image vignettes

☐ White   ☐ Black   ◼ Hispanic English   ◼ Hispanic Spanish

Figure 4   Estimated vignette locations, compared to the reference vignette (severity 3) on the latent health spectrum (measured in standard deviations of the reference vignette) for each health domain. Zero on the y-axis represents the mean of the reference (most pain or least healthy) vignette; higher numbers represent better perceived health.

shown in Figure 4B2, despite the VE violation (e.g., Hispanics who took the Spanish survey view the first vignette person as having more sleep difficulties compared with White respondents), image vignettes did a much better job differentiating the intensity levels of the three vignettes. Similar results are found for mobility domain (see Figures 4C1 and 4C2).

RQ 2b. RC-Test

As described in the *Analysis* section, the RC assumption test is based on visual comparisons of two sets of predicted mean cut points: one from Model A which has only vignettes (i.e., no self-assessments included in the model) and another from Model C which includes self-assessments and objective measures. Figure 5 shows the estimated cut points for all four health domains. If the vignette-derived cut point patterns are similar to the health measures-derived cut points, this indicates no or only minor violations of RC. For pain domain, both text and image vignettes show minor violations of RC. For all other three domains, image vignette conditions seem better fulfill RC, compared to text conditions.
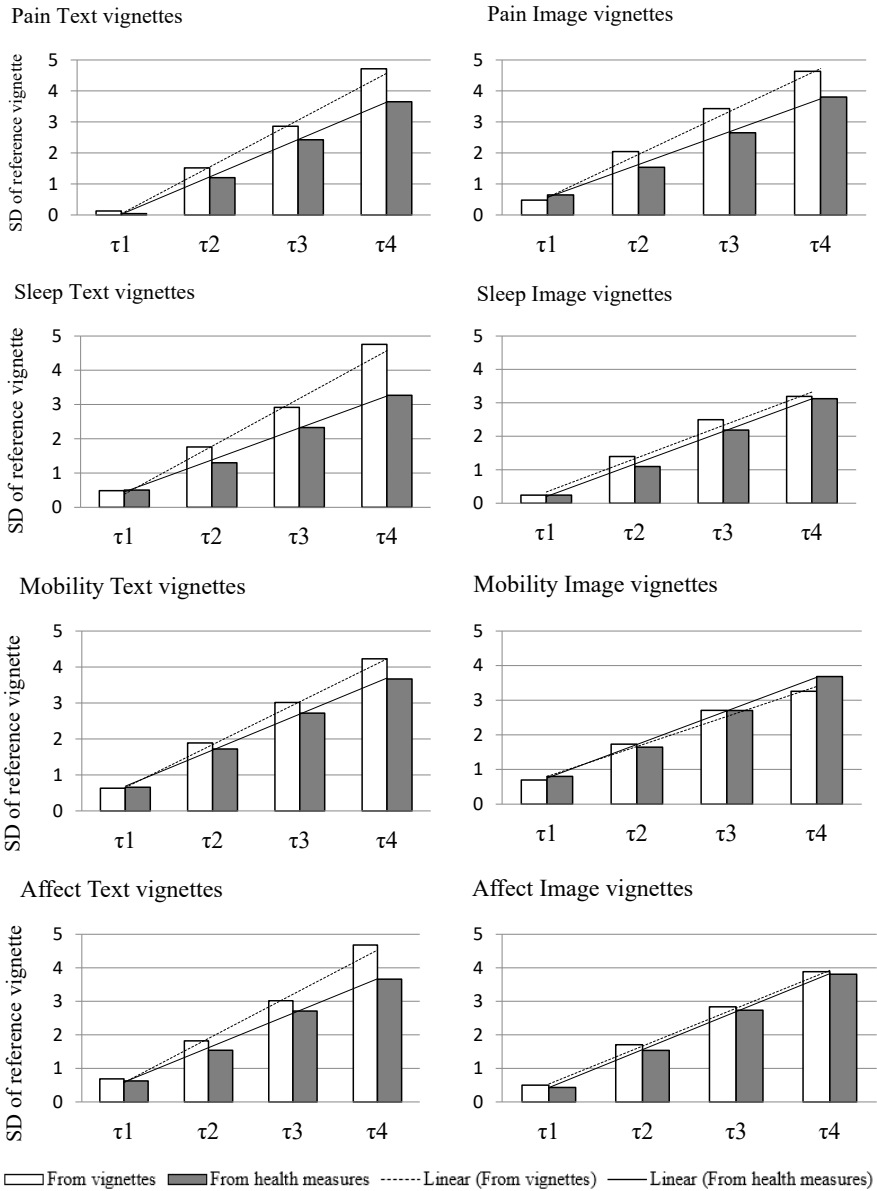
Figure 5 Estimated cut points for health domains based on vignettes and health measures. Evaluations are based on comparisons to the reference vignette [highest severity; measured in standard deviations (SD) of the reference vignette]. $\tau1$–$\tau4$ are cut points for the five-point response scale from "None" to "Extreme" (e.g., $\tau1$ is the cut point between "None" and "Mild").

RQ 3. Confirmatory factor analysis before and after anchoring vignette-adjustments

To test whether image and text AV-adjusted scores perform better than original scores (RQ 2a.), we compared model fit indices in a confirmatory factor analysis (CFA) using both adjusted scores and original scores. The cutoff criteria for acceptable fit are presented in the Analysis Strategy section. As shown in Table 6, CFI are above 0.95 and TLI are around or above 0.95 for all models, indicating that the models fit the data well for all the conditions. Models with AV-adjusted scores lead to better (i.e., higher) CFI and TLI values. For example, for the image condition subsample, the TLI of the model with image AV-adjusted self-assessment scores is 0.977, which is higher than the TLI of the model with original self-assessments – 0.942. RMSEA results shows that using both text and image AV-adjusted scores can greatly improve RMSEA. This suggests that using both text and image-adjusted scores improve CFA model fit.

To test whether image AV-adjusted scores perform similar or better than text AV-adjusted scores in the CFA framework (RQ 2b.), we assessed the model fit indices in CFA with text AV-adjusted scores and CFA with image AV-adjusted scores. As shown in Table 6, both text and image AV-adjustment improves the CFA based on original self-reports with similar improvements in terms of model fit indices. In addition, the CFI, TLI and RMSEA results are similar across the two CFA models with text vs. image AV-adjusted scores. The CFA with image AV-adjusted scores have better 90% CI of RMSEA (which ideally should have the lower value less than 0.05 and the upper value less than 0.08).

Table 6   Confirmatory Factor Analysis model fit estimates based on the original and anchoring vignette-adjusted scores.

Model	N	CFI	TLI	RMSEA	90% CI of RMSEA
CFA with original self-assessments (full sample)	3,149	0.983	0.948	0.158	(0.138, 0.179)
Text condition subsample					
CFA with original self-assessments	1,051	0.986	0.958	0.151	(0.117, 0.189)
CFA with text AV – adjusted self-assessment scores	1,051	0.994	0.982	0.060	(0.026, 0.100)
Image condition subsample					
CFA with original self-assessments	2,098	0.981	0.942	0.162	(0.137, 0.188)
CFA with image AV – adjusted self-assessment scores	2,098	0.992	0.977	0.062	(0.038, 0.089)

Discussion

This study examines the use of image anchoring vignettes (AV) to adjust DIF in self-assessments of health. Despite the fact that text AV have been adopted in many comparative studies, there are several critical challenges associated with text AV. To explore ways to overcome these challenges, this paper proposes the use of image AV, consisting of carefully designed and pre-tested images. In this study, the performances of text and image AV are compared with respect to a number of properties, including response time, tests of assumptions, and CFA model fits. Overall, the results suggest that the image AV methodology can be used as an improved and effective alternative to text AV in cross-cultural research, although the extent to which the VE assumption is satisfied needs further investigation for both text and image AV.

Specifically, the use of image AV can reduce survey time to about half the time of text AV. This result is consistent with previous literature on differences of information processing between text vs. image stimuli (Azizian et al., 2006; Naspetti et al. 2016; Schlochtermeier et al. 2013). Survey time is an important indicator for respondent cognitive burden, which can influence survey data quality and survey response rates. Survey time is also closely associated with survey cost, with shorter time potentially implying lower survey costs. Thus, image AV offers a time and potentially cost-efficient survey option, compared to text AV, especially in studies with many AV items (e.g., Weiss & Roberts, 2018).

Results for comparing how well AV assumptions are satisfied between text and image AV show mixed findings. On the one hand, image AV outperforms text AV in that respondents can better distinguish the different intensity levels in image vignettes (e.g., from no pain to extreme pain) than in text vignettes, indicating that respondents are more likely to perceive the vignettes in similar ways and in the designed order in the image AV condition compared to the text AV condition. This finding is consistent with previous literature showing the information processing advantage of emotional images in terms of larger or more pronounced emotion effects evoked by image stimuli, compared to text stimuli (e.g., Schlochtermeier et al., 2013). One of the reasons may be that image vignettes lead to a stronger activation of relevant information in the cognitive system resulting in more arousal and perceived intensity. Another possible reason is that text AV puts a higher cognitive burden on respondents, potentially resulting in more satisficing behavior including straight-lining (i.e., respondents select the same response option for all the vignette questions) and random selection of responses. For example, we find that respondents assigned to the text vignettes treatment are more likely to straight-line than those assigned to image vignettes (results not shown).

On the other hand, for both text and image AV, it is found that respondents' perceptions of the vignettes can differ by cultural subgroups, a violation of VE.

Similar to text AV, various factors may cause violations of VE for image vignettes. First, like text vignettes, the information in image AV may serve as memory cues that can trigger other related memories, leading to differences in perceptions. Second, although elements included in image AV may be more easily standardized than text AV (e.g., gender of the hypothetical person), the included elements may still weigh differently for different subgroups. For example, an element in the image may be more familiar to one cultural group than to another, resulting in perception differences. The violation of VE implies that designing "universal" anchoring vignettes (Grol-Prokopczyk, 2018), which are familiar to all population groups and reveal the same information to all respondents, is still a challenge for both text and image vignettes.

Despite the VE violations, results of the CFA models indicate that, compared to the model with self-reported data, using vignettes-adjusted scores can greatly improve model fit, which is consistent with Weiss & Roberts (2018)[10]. This shows that, even though VE is not met, it is still better to use text or image AV-adjustments, which can effectively reduce measurement errors. Comparing the two vignette types, text and image vignettes perform similarly in terms of measurement error reduction in the CFA models.

Given the clear advantage of image vignettes in reducing survey time, lowering respondents' cognitive burden and better differentiating intensity levels, we believe there is a potential for the use of image AV to improve text AV methodology.

This study also revealed important findings to deepen our understanding of the vignette methodology, including how different respondents view and rate vignettes. For example, it was found that male respondents view the first pain vignette as describing more pain than female respondents do (as shown in Table 5). This may be because females experience more pain than males (Cepeda & Carr, 2003). They may use themselves as a standard of comparison when rating the vignette person and thus view the first vignette person as depicting minimal pain. Due to space restrictions, this study will not discuss detailed results for all covariates. Future studies can look into this further. In addition, this study generates new insights into the design and use of image AV, and the designed image AV items can be applied to other studies that use anchoring vignettes to adjust self-reported health.

It is worth mentioning that this study is limited in several ways. First, due to resource constrains, our experimental study is based on a non-probability sample, from which the results were not intended to generalize to the full U.S. population. Among the four types of validity of causal inference (statistical, internal, external and construct validity) in Shadish, Cook and Campbell (2002), this paper focused

---

10   We also examined the DIF-adjusting results using HOPIT models. Results are similar for both text and image vignettes. Due to space restrains, results are not shown in this paper and are available upon request.

on the internal and statistical validity with a randomized experiment to compare DIF-adjustment results between vignette types. Per Edgington (1966) and Berk et al. (1995), randomized experiments permit statistical inferences about the experimental factors. However, due to the nature of the sample, we do not claim that our results generalize to the complete U.S. population and beyond. Future studies could replicate this study in probability-based representative surveys to evaluate the effect sizes of the group comparisons in the population. Second, the current RC test is not based on a statistical test and additional evaluations of RC using more stringent RC test are needed (Grol-Prokopczyk, 2018). Third, 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 the questions about objective health are straightforward factual questions (e.g., whether respondent has received doctor diagnosis of certain diseases), for which reporting errors may be less of an issue compared to self-assessing of a health domain. Also, many of the objective measures used in this study are based on widely-used existing scales, and have been successfully applied in previous literature (Kessler et al., 2002; Ray et al., 2009). If available, future studies could use bio-markers (e.g., medical test results and genetic data) in the RC tests. Fourth, this study examined the most commonly used text vignettes that are included in HRS, SHARE, and many other large-scale surveys. It is possible that text vignettes with differently-worded descriptions may perform better in tests of assumptions than the current text vignettes. The same may be true for image vignettes. Possibly, better-designed pictures are less likely to lead to rejection of the VE and RC assumptions. Future research could compare text and image vignettes with different descriptions or designs.

Our research suggests several important directions for future research. First, this study focuses on the comparisons of text and image vignettes in correcting for DIF. Future research could examine in detail how different image vignette designs may influence the performance of image AV. For example, in a related study, we found that when rating image vignettes with average body size vs. obese for the mobility domain, respondents tend to rate the obese vignette person as having more mobility difficulties than a vignette person with average body size. This is not surprising given that obese individuals are more likely to have mobility limitations than non-obese individuals (Koster et al., 2007). In addition, the vignette images showing average body sizes, which match the body size of the majority of respondents, show a higher rate of consistency in the rank-orderings, indicating that respondents may better perceive the image vignettes when the vignette figures match more closely their own characteristics. This could shed light on the future design of image vignettes. For example, it indicates that image vignettes that have a broader applicability and familiarity to the respondents may better satisfy the assumptions. Future research could further evaluate the effects of a wide range of

vignette characteristics on image vignette performance. Second, given budget constraints, 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 the study to cross-national surveys and/or to a wide variety of other racial/ethnic groups, such as Asians, American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander. Third, some domains may be too complex to be expressed using images, such as self-reported political attitudes. In addition, using static image vignettes may not be the best way to present measures related to change over time and location, such as a slow or fast walking speed. Future research can evaluate other visual vignette designs such as using short videos in web surveys (Banuri et al., 2018; Mendelson, Gibson, & Romano-Bergstrom, 2017) and the use of visual vignettes in different domains, including domains that cannot be easily visualized using static images. Fourth, the ways vignettes are presented and their applications can vary by survey mode, which may influence their performance. Verbal vignettes can be delivered orally in telephone and face-to-face interviews or visually as text in mail and web surveys, but image vignettes have to be presented visually in mail and web surveys, or as a picture presented by interviewers in face-to-face surveys. Future research could evaluate mode effects for both text and image vignettes.

In conclusion, this study indicates that using either text or image AV adjustments can reduce measurement errors compared to the analysis without using any AV, and the use of image AV can greatly reduce survey time and respondents' cognitive burden as compared to text vignettes. Improving VE, (in other words, minimizing different interpretations of vignettes by different groups), is critical for both text and image AV and requires further investigation. This study has advanced knowledge of the design and applications of image AV in health surveys and has implications for designing image AV of other domains. Future implementations of AV can use the findings of this study to introduce efficiencies in their survey designs.

References

Azizian, A., Watson, T. D., Parvaz, M. A., & Squires, N. K. (2006). Time course of processes underlying picture and word evaluation: an event-related potential approach. *Brain Topography*, 18(3), 213-222.

Banuri, S., de Walque, D., Keefer, P., Haidara, O. D., Robyn, P. J., & Ye, M. (2018). The use of video vignettes to measure health worker knowledge. Evidence from Burkina Faso. *Social Science & Medicine*, *213*, 173–180.

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological methodology*, 421-458.

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.

Buskirk, T. D. (2015). Are sliders too slick for surveys? An experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys. *methods, data, analyses*, 9(2), 32.

Caramazza, A. (1996). Pictures, words and the brain. *Nature*, 383(6597), 216-217.

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. (2005). Technology trends in survey data collection. *Social Science Computer Review*, 23(4), 486–501.

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

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2), 227-245.

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, 68(2), 255-266.

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.

d'Uva, T., 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.

Davis, R. E., Johnson, T. P., Lee, S., & Werner, C. (2019). Why do Latino survey respondents acquiesce? Respondent and interviewer characteristics as determinants of cultural patterns of acquiescence among Latino survey respondents. *Cross-Cultural Research*, 53(1), 87-115.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of general internal medicine*, 21(3), 267-275

Dowd, J. B., & Todd, M. (2011). Does self-reported health bias the measurement of health inequalities in US adults? Evidence using anchoring vignettes from the Health and Retirement Study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(4), 478–489.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, 66(6), 485.

Ferrer-i-Carbonell, A., Van Praag, B. M. S., & Theodossiou, I. (2011). Vignette Equivalence and Response Consistency: The Case of Job Satisfaction. Retrieved from http://papers.ssrn.com/abstract=1968870

Glaser W. R. (1992). Picture naming. *Cognition*: 42(1–3), 61–105.

Glaser W. R., Glaser M. O. (1989). Context Effects in Stroop-Like Word and Picture Processing. *Journal of Experimental Psychology: General*: 118(1), 13–42.

Grady, C. L., McIntosh, A. R., Rajah, M. N., & Craik, F. I. (1998). Neural correlates of the episodic encoding of pictures and words. *Proceedings of the National Academy of Sciences*, 95(5), 2703-2708.

Gravelle, T. B. (2021). The Measurement Invariance of Customer Loyalty and Customer Experience across Firms, Industries, and Countries. *methods, data, analyses*, 23

Grol-Prokopczyk, H. (2018). In pursuit of anchoring vignettes that work: Evaluating generality versus specificity in vignette texts. *The Journals of Gerontology: Series B*, 73(1), 54-63.

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, 52(5), 1703-1728.

Groot, T. D., Jacquet, W., Backer, F. D., Peters, R., & Meurs, P. (2020). Using image vignettes to explore sensitive topics: a research note on exploring attitudes towards people with albinism in Tanzania. *International Journal of Social Research Methodology*, 1-7.

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. *The reading teacher*, 56(8), 758-770.

Hirve, S., Gomez-Olive, 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(1), 21064.

Holt, T. P., & Loraas, T. M. (2019). Using Qualtrics panels to source external auditors: A replication study. *Journal of Information Systems*, 33(1), 29-41.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public opinion quarterly*, 74(2), 201-222.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.

Hu, M., & 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., & Xu, H. (2018). Using Anchoring Vignettes to Correct for Differential Response Scale Usage in 3MC Surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methodology*.

Ibarra, J. L., Agas, J. M., Lee, M., Pan, J. L., & Buttenheim, A. M. (2018). Comparison of online survey recruitment platforms for hard-to-reach pregnant smoking populations: feasibility study. *JMIR research protocols*, 7(4), e8071.

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, 21-37.

Idler, E. L., & Kasl, S. V. (1995). Self-ratings of health: do they also predict change in functional ability?. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(6), S344-S353.

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.

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

Krosnick, J. A., & Abelson, R. P. (1992). The case for measuring attitude strength in surveys. In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys* (pp. 177–203). Russell Sage Foundation. Retrieved from https://books.google.com/books?hl=en&lr=&id=8FEiM0gA_wwC&pgis=1

Kyllonen, P. C., & Bertling, J. P. (2014). Anchoring vignettes reduce Bias in noncognitive rating scale responses. *Report Submitted to OECD*.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health*, *25*, 295–306.

Lee, S., Liu, M., & Hu, M. (2017). Relationship between future time orientation and item nonresponse on subjective probability questions: A cross-cultural analysis. *Journal of cross-cultural psychology*, *48*(5), 698-717.

Lee, S., Schwarz, N., & Goldstein, L. S. (2014). Culture-sensitive question order effects of self-rated health between older Hispanic and non-Hispanic adults in the United States. *Journal of aging and health*, *26*(5), 860-883.

Liu, M., Kuriakose, N., Cohen, J., & Cho, S. (2016). Impact of web survey invitation design on survey participation, respondents, and survey responses. *Social Science Computer Review*, *34*(5), 631–644.

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

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130.

McCarthy, M., Ruiz, E., Gale, B., Karam, C., & Moore, N. (2004). The meaning of health: Perspectives of Anglo and Latino older women. *Health Care for Women International*, *25*(10), 950-969

Mendelson, J., Gibson, J. L., & Romano-Bergstrom, J. (2017). Displaying Videos in Web Surveys: Implications for Complete Viewing and Survey Responses. *Social Science Computer Review*, *35*(5), 654–665.

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.

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.

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., 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.

Murray, C. J., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R., & Chatterji, S. (2003). Empirical evaluation of the anchoring vignettes approach in health surveys. In *Health systems performance assessment: Debates, methods and empiricism* (pp. 369–399).

Naspetti, S., Mandolesi, S., & Zanoli, R. (2016). Using visual Q sorting to determine the impact of photovoltaic applications on the landscape. *Land Use Policy*, 57, 564-573.

Naylor, R., Maye, D., Ilbery, B., Enticott, G., & Kirwan, J. (2014). Researching controversial and sensitive issues: using image vignettes to explore farmers' attitudes towards the control of bovine tuberculosis in England. *Area*, 46(3), 285-293.

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.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment*, *32*(1), 39–51.

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*.

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *methods, data, analyses*, 11(2), 28.

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.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, 99(6), 323-338.

Sevostianov, A., Horwitz, B., Nechaev, V., Williams, R., Fromm, S., & Braun, A. R. (2002). fMRI study comparing names versus pictures of objects. *Human brain mapping*, 16(3), 168-175

Schlochtermeier, L. H., Kuchinke, L., Pehrs, C., Urton, K., Kappelhoff, H., & Jacobs, A. M. (2013). Emotional picture and word processing: an fMRI study on effects of stimulus complexity. *PLoS One*, 8(2), e55619

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin

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.

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.

Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software*, 42(1), 1-25.

Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported perso-
    nality: A comparison between countries. *Frontiers in Psychology*, *9*(MAR), 1–17.

Witte, J. C., Pargas, R. P., Mobley, C., & Hawdon, J. (2004). Instrument effects of images in
    web surveys: A research note. *Social Science Computer Review*, 22(3), 363–369.

Yan, T., & Hu, M. (2018). Examining Translation and Respondents' Use of Response Scales
    in 3MC Surveys. *Advances in Comparative Survey Methods*, 501–518.
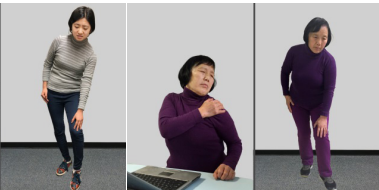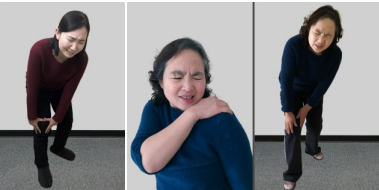
APPENDIX 1

Text and image vignettes used for the web survey for each domain.

Note that in the design of image vignettes, we have two different design conditions per domain. Given that the aim of this paper is to compare text vs. image vignettes, data from different designs of image vignettes are combined in all the analysis. The evaluation the design of features on AV methodology is discussed elsewhere.

Supplemental Table 1   Pain text and image vignettes.

Pain Intensity Level	Text vignette	Image Design One (young adults)	Image 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.		
Moderate 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.		

Supplemental Table 2   Sleep text and image vignettes.

Sleep Difficulty Level	Text vignette	Image Design One (female)	Image 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.		
Moderate 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 him/her to go back to sleep. In the morning she/he 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.		

Supplemental Table 3   Mobility text and image vignettes.

Mobility Difficulty Level	Text vignette	Image Design One (optimal weight/fit)	Image 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.		
Moderate 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.		

Supplemental Table 4   Affect text and image vignettes.

Depression Level	Text vignette	White	Black	Hispanic
No / Low Depression	Matt enjoys his work and social activities and is generally satisfied with his 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.			
Moderate Depression	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 him. 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 to others and that he would be better off dead.			

APPENDIX 2

Randomization conditions and assignments and robustness checks for randomization across text and image conditions.
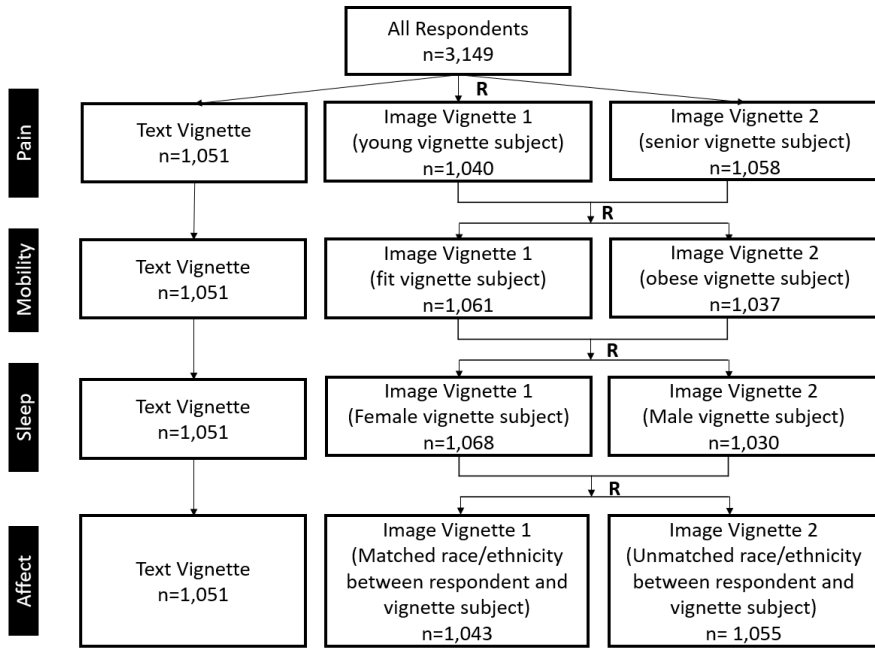

Supplemental Figure 1 Experimental conditions and assignments for each domain. "R" indicates randomization was done.

Supplemental Table 5 Robustness checks for randomization across text and image conditions.

	Text	Image	Chi-square / F statistics
Gender			0.03
Female	52.3	51.9	
Male	47.7	48.1	
Age (mean)	46.9	46.7	0.10
Race			0.47
White	23.8	24.3	
Black	23.8	23.9	
Non-Hispanic White	23.5	24.0	
Non-Hispanic Black	28.9	27.8	
Education			0.01
Below high school	52.2	52.0	
High school and above	47.8	48.0	
Employment status			1.40
Employed	52.6	54.9	
Not employed	47.4	45.1	
Marital status			0.55
Married	50.2	48.8	
Not married	49.8	51.2	
Income			0.84
Low	34.3	34.9	
Middle	42.2	40.6	
High	23.5	24.5	

APPENDIX 3

Distributions of log-transformed response time variable for each domain.
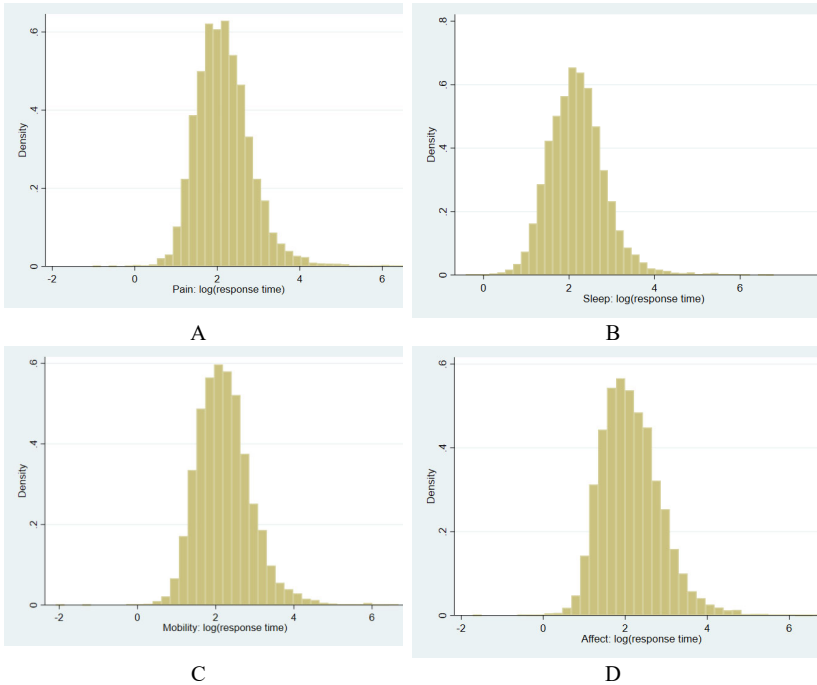
Supplemental Figure 2   Distributions of log-transformed response time variable
for each domain.

To formally test the differential response time by vignette types, for each health domain, we fit multilevel logistic regression models with random intercepts. Given that time is right skewed, we used log-transformed time as outcomes (distributions shown in Appendix 3). In the unconditional model (i.e., no predictors in the model) for each domain, log-transformed response time varied significantly across individuals (the intraclass correlation coefficient [ICC] ranges from 0.43 to 0.50, see Supplemental Table 6), justifying the use of multilevel modeling. Supplemental Table 6 shows the results of the final models which include both question level predictors (i.e., image vs. text vignettes) and respondent level predictors (e.g., demographic and socio-economic variables). As shown in Supplemental Table 6, compared to text vignettes, respondents spent significantly less time answering image vignettes. This is true for all four domains. Compared to non-Hispanic White, respondents

of all other three groups spent significantly longer time in answering the vignette
questions.

Supplemental Table 6   Multilevel linear regression models predicting log-
transformed response time for each health domain.

	Model			
	Pain	Sleep	Mobility	Affect
Image vignettes (ref: Text vignettes)	-0.63***	-0.58***	-0.68***	-0.78***
Age	0.01***	0.01***	0.01***	0.01***
Male (ref: Female)	-0.01	0.00	-0.06**	-0.04*
Above high school education (ref: Below high school)	-0.05**	-0.30	-0.05*	-0.04*
Employed (ref: Not employed)	-0.06**	-0.60**	-0.04*	-0.02
Married (ref: Not married)	-0.05**	-0.45*	-0.04	-0.06**
Respondent Groups (Ref: Non-Hispanic White)				
Non-Hispanic Black	0.18***	0.20***	0.18***	0.18***
Hispanics English	0.06*	0.07**	0.09**	0.07**
Hispanics Spanish	0.16***	0.16***	0.14***	0.17***
ICC (95% confidence interval)	0.50 (0.48, 0.52)	0.45 (0.43, 0.47)	0.49 (0.47, 0.51)	0.43 (0.41, 0.45)

*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

APPENDIX 4

Model results for evaluating VE test for each domain (with both image and text vignettes combined for analysis).

Supplemental Table 7   Predictors for perceived vignette locations on the latent health spectrum.

	Pain	Sleep	Mobility	Affect
Vignette 1 (no/mild difficulty/intensity)				
Constant	3.39***	1.74***	1.90***	4.10***
Image	1.59***	2.84***	0.17**	1.07***
Married	0.21*	-0.07	0.12	0.15
Male	-0.43***	-0.23**	-0.14*	-0.24**
Employed	-0.02	0.15	0.10	0.12
More than high school	-0.03	0.11	0.12	-0.10
Age 18 - 29	-0.20	0.02	-0.12	-0.50**
Age 30 - 49	-0.21	0.10	-0.09	-0.40**
Age 50 - 64	-0.27*	0.09	-0.05	-0.39**
Middle income	-0.14	-0.11	0.00	-0.19*
High income	-0.05	-0.19	-0.31***	-0.46***
Black	-0.08	-0.13	-0.15	-0.44***
Hispanic (English)	-0.17	-0.14	-0.04	-0.17
Hispanic (Spanish)	-0.89***	-0.55***	-0.48***	-1.13***
Vignette 2 (moderate difficulty/intensity)				
Constant	1.58***	0.20	-0.14	2.21***
Image	0.16*	1.03***	0.98***	-0.27***
Married	0.01	-0.07	0.06	0.06
Male	-0.16*	0.00	0.00	-0.08
Employed	0.00	0.05	0.10	0.09
More than high school	-0.04	0.04	0.12*	-0.10
Age 18 - 29	-0.06	0.15	0.07	-0.33**
Age 30 - 49	-0.12	0.18	-0.01	-0.26*
Age 50 - 64	-0.15	0.07	-0.08	-0.24*
Middle income	-0.04	-0.08	-0.03	-0.10
High income	-0.05	-0.14	-0.18*	-0.18*
Black	0.05	-0.07	-0.12	-0.11
Hispanic (English)	0.00	-0.05	-0.05	-0.11
Hispanic (Spanish)	-0.15	-0.19*	-0.15*	-0.46***

Notes: Vignette 3 (highest difficulty/intensity) is the reference vignette. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.