Assessing (and Addressing) Reporting Heterogeneity in Visual Analogue Scales (VAS) with an Application to Gender Difference in Quality of Life[†]

Zhiyong Huang[a] and Fabrice Kämpfen[b,c*]

[a]Southwestern University of Finance and Economics, Chengdu, China
[b]Department of Economics, HEC, University of Lausanne, Switzerland
[c]Population Studies Center, University of Pennsylvania, USA

July 30, 2019

Abstract

In this study, we propose several new methods to account for reporting heterogeneity in self-reported data coming from Visual Analogue Scales (VAS) using corresponding VAS-based anchoring vignettes. Compared to usual Likert scale measures, VAS have the advantage that they lead to more nuanced assessments. Yet, like responses to Likert scale, VAS may suffer from individual-specific reporting heterogeneity. To the best of our knowledge, such reporting heterogeneity and potential solutions to solve this problem in the context of VAS measures have not yet been addressed in the literature. Using VAS-based anchoring vignettes and standard vignettes assumptions (vignette equivalence and response consistency), we show how standard fixed-effect approaches and double-index models can be used to address individual-specific reporting heterogeneity in VAS. We also show that several other methods such as Generalized

[†]We are grateful to the participants of the North American Summer Meeting of the Econometric Society at UC Davis (2018), the Annual Congress of the Swiss Society of Economics and Statistics in St-Gallen Switzerland (2018), the Southwestern University of Finance and Economics seminar in Chengdu China, the Demography Club at the University of Pennsylvania, the 12th Ruhr Graduate School Doctoral Conference in Economics in Germany and to Jürgen Maurer for valuable comments and suggestions that greatly improved our paper. We also would like to thank Katja Schwab-Weis and the HEC Research Fund at the University of Lausanne for their logisitics and financial support.

[*]Corresponding author. E-mail address: kampfenf@sas.upenn.edu, Tel.: +1 215-898-6441. Do not quote or circulate without permission of the authors. The data and codes used in this article can be obtained from the corresponding author upon request. This study was financially supported by the University of Lausanne - HEC Research Fund: the grant was exclusively used to incitivize students to participate in the online survey. The study was approved by the HEC Ethics Committee of the University of Lausanne in February 2017 and informed consent was obtained from all individual participants included in the study. Zhiyong Huang and Fabrice Kämpfen have nothing else to disclose.

Ordered Response models and Hierarchical Ordered Probit (HOPIT) models can be used to meaningfully adjust for potential reporting heterogeneity under the weaker assumption that VAS responses should be interpreted as ordered rather than cardinal data. We then apply our methods to real data assessing gender differences in Quality of Life (QoL) among students in Switzerland. While female students report higher levels of QoL than male students –as commonly found in the literature– we also show that female students tend to rate the QoL of corresponding comparable anchoring vignettes higher than male students. Accounting for these gender differences in response behaviors, we show that female students actually appear to be worse off in terms of QoL than male students. This finding suggests that reporting heterogeneity may be important in assessing gender differences in QoL and that the commonly found female advantage in QoL assessments may at least be partially due to differences in reporting behavior.

Keywords: *Reporting heterogeneity, Visual Analogue Scale, Quality of life, Method*
JEL: *C30, C81, I31, J16*

1 Introduction

In economic, social, medical and psychological studies, Visual Analogue Scales (VAS) are a widely used measurement tool for eliciting subjective assessments such as pain intensity (Ismail *et al.*, 2015; Kelly and Anne-Maree, 1998; Zampelis *et al.*, 2014), distress (Lesage *et al.*, 2012; Obayashi *et al.*, 2016), quality of life (QoL) (Abdel-Fattah *et al.*, 2007; Devesa *et al.*, 2012), happiness (César *et al.*, 2014; Sakamoto *et al.*, 2016) and many other constructs that are often difficult, costly or impossible to measure objectively. As continuous scales, VAS operates through a horizontal or vertical line of some fixed length with labels at the two end-points, called "anchors", representing the best and worst scenarios. VAS are widely used in surveys because they are easy to understand by respondents, simple to implement and have high levels of validity and reliability (Abend *et al.*, 2014; Bailey *et al.*, 2012; Bijur *et al.*, 2001). Compared with discrete Likert scale measures (e.g., a 5-point scale with "poor", "fair", "good", "very good", "excellent" as typical 5-level Likert items), VAS measures are more nuanced and less likely to be biased from discretization, and therefore often have better discriminating power (Studer, 2011).

Despite their apparent advantages, VAS measures can be subject to reporting heterogeneity among different groups of respondents, who may have different interpretations of the two anchors attached to the scale. For example, VAS of pain often takes "no pain" as one endpoint and "worst possible pain" as another. Both of these two anchors can be interpreted very differently by heterogeneous groups of respondents. And because respondents use these anchors to answer the questions that they are being asked, the variations in the interpretations of these anchors across groups of respondents could potentially mean that they use different reporting scales, making comparisons between self-reported measures across these groups difficult to interpret. While it is well-known that measures based on Likert scale suffer from reporting heterogeneity, there exists, to the best of our knowledge, no study on whether reporting heterogeneity is present in VAS measures, and if so, how to correct for it.

To examine potential reporting heterogeneity in VAS measures, we use anchoring vignettes. Anchoring vignettes are short descriptions of hypothetical individuals or situ-

ations evaluated by respondents alongside their self-assessment on the same domain. For example, a vignette on life satisfaction, from the second wave (2006-2007) of the Survey of Health, Ageing and Retirement in Europe (SHARE), is profiled as: "John is 63 years old. His wife died 2 years ago and he still spends a lot of time thinking about her. He has 4 children and 10 grandchildren who visit him regularly. John can make ends meet but has no money for extras such as expensive gifts to his grandchildren. He has had to stop working recently due to heart problems. He gets tired easily. Otherwise, he has no serious health conditions". Respondents are asked to evaluate the life satisfaction of this hypothetical person alongside their own life satisfaction using a five-point Likert scale measure: "very dissatisfied", "dissatisfied", "neither satisfied, nor dissatisfied", "satisfied" or "very satisfied". As anchoring vignettes are pre-defined and invariant across individuals, their ratings are supposed to reflect only the differences in reporting scales. Under testable assumptions of vignette equivalence and response consistency, the effect of reporting heterogeneity can be purged out from self-assessed measures, making the adjusted self-assessment better reflect the actual underlying evaluations of individuals. Anchoring vignettes have been used to correct for reporting heterogeneity in Likert scale measures in many domains such as health (Bago d'Uva *et al.*, 2008; d'Uva *et al.*, 2008; Hanandita and Tampubolon, 2016; Jürges, 2006; Molina, 2016; Mu, 2014; Salomon *et al.*, 2004), healthcare (Malhotra and Do, 2013; Rice *et al.*, 2012), political efficacy (Hopkins and King, 2010; King *et al.*, 2004), job satisfaction (Kristensen and Johansson, 2008), working disability (Kapteyn *et al.*, 2007), social status (Wang, 2016) and subjective well-being (Angelini *et al.*, 2013; Crane *et al.*, 2016; Kapteyn *et al.*, 2013; Ravallion *et al.*, 2016) to name but a few.

In this paper, we employ anchoring vignettes to explore the presence of reporting heterogeneity in VAS measures and develop econometric models to correct for it. While Hierarchical Ordered Probit (HOPIT) models are commonly used to adjust for reporting heterogeneity in self-reported Likert scale measures using anchoring vignettes (King *et al.*, 2004), continuous VAS measures enable us to explore a much richer set of econometric models such as linear fixed-effect models and linear double-index models, which treat

reporting heterogeneity as an index of individual characteristics. We also design models that relax the cardinality assumption implied in linear VAS model specifications and propose models that account for reporting heterogeneity in VAS measures by developing ordered response models that only assume that the VAS data represent a valid ordering of outcomes. More specifically, we explore the possibilities of using Generalized Ordered Probit and HOPIT models to fit (properly discretized) VAS measures and compare the resulting estimates with estimates that are derived from linear VAS models.

We test our econometric models on observations collected from students who have been asked, among other things, to evaluate their QoL using VAS in the context of our online survey. Online surveys are particularly well-adapted for this kind of research question (Kapteyn *et al.*, 2007) and have been used in many other studies to assess reporting heterogeneity in self-reported evaluations (Kapteyn *et al.*, 2007; Studer, 2011; van Soest *et al.*, 2011).

QoL is defined as the general well-being of individuals and societies and includes several factors like physical and mental health, family, education, employment and many other features of life. There has been a long-standing interest in epidemiology, psychology, economics, sociology and other fields in QoL (Abdel-Fattah *et al.*, 2007; Deaton, 2018; Devesa *et al.*, 2012) and its determinants, such as age, gender, education, family income, physical and mental health, employment and social support, to name but a few, are, to a certain degree, well understood in the literature. However, one puzzling result, which is often found in the literature and which we will revisit in our application of VAS-based vignette models, is that despite being disadvantaged in terms of income, education, health and many other social dimensions, women tend to report being happier and having a higher QoL than men (Deaton, 2018; Helliwell *et al.*, 2017; Montgomery, 2016). Similar to these findings, our initial analysis of gender differences in our QoL data without vignette adjustments suggests that this is also the case in our sample. However, after adjusting for reporting heterogeneity using anchoring vignettes, we show that females are actually worse off than males in terms of QoL in our study. These results are robust across various model specifications and econometric assumptions. Our findings therefore suggest that

reporting heterogeneity could be one of the reasons for the puzzling result that has been reported in the literature on QoL and happiness.

The rest of the paper is organized as follows. In the next section, we describe the data that we collected, provide some descriptive statistics of our sample and show some initial evidence of the presence of reporting heterogeneity in our VAS self-reported measure of QoL. Section 3 then develops and explains in detail the various econometric specifications that we put in place to account for reporting heterogeneity. We apply our econometric models to our data and show the results of our estimates in Section 4. Section 5 discusses and presents the fitted values resulting from our benchmark econometric model and the counterfactual distributions of QoL in our sample, once reporting heterogeneity has been purged out from the self-reported evaluations. Section 6 provides the results of tests and robustness checks that we conduct in order to support our analyses and the different assumptions we make in our econometric models. Section 7 concludes the paper.

2 Data and descriptive statistics

The sample of this study was recruited among a pool of students from various schools of higher education in the region of Lausanne in Switzerland[1] in 2017. We created an online questionnaire on Qualtrics[2] and sent out invitation emails to all the students who registered in the University of Lausanne's experiment program, which amounted to a total of 6,578 emails. Students had two weeks to complete the survey and a total of 1,938 observations was collected (response rate= 29.5%). As incentives, students who completed the survey were automatically entered into a lottery for which the highest prize was about USD 300[3].

[1]Participants were students at the Swiss Federal Institute of Technology in Lausanne (EPFL), the University of Lausanne (UNIL) and Ecole hoteliere de Lausanne (EHL).

[2]More information can be found here: https://www.qualtrics.com

[3]The distribution of prizes was the following: 3 × USD 300, 10 × USD 100 and 60 × USD 20.

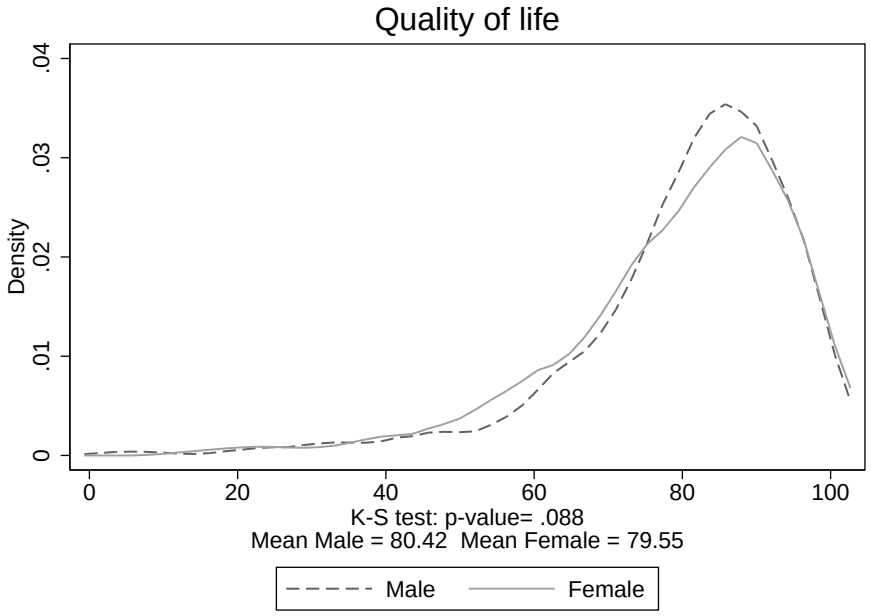

Figure 1: Kernel densities of the QoL of the respondents by sex. 0 corresponds to the label "worst possible QoL" whereas 100 corresponds to the label "best possible QoL".

2.1 Outcome variable of interest and anchoring vignettes

Respondents were asked to evaluate their QoL on a VAS that uses a horizontal line on which respondents had to move a slider from left or right. The VAS we use in our study ranges from 0 ("worst possible QoL") to 100 ("best possible QoL")[4]. Figure 1 shows the kernel distributions of self-evaluated QoL for both males (dashed line) and females (solid line) in our sample. The distributions are of similar shapes, although the distribution for males seems more centered around its mean (80.42), whereas the one for females is more skewed towards the left (mean equal to 79.55). The p-value of a Kolmogorov-Smirnov test of equal distribution is equal to 0.088, and therefore indicates that we cannot reject the null hypothesis (at 5%) that the QoL of males and females in our sample are drawn from the same distribution. Overall, respondents evaluate their QoL rather positively, with only 4.57% of them judging their QoL to be closer to 0 than 100.

To determine the existence of reporting heterogeneity in self-reported QoL, we asked respondents to evaluate the QoL of persons described in three hypothetical scenarios,

[4]Only the two values (and their corresponding labels) at both ends of the horizontal graphic slider, i.e., 0 and 100, were displayed and no tick points in between were used.

also called anchoring vignettes, using identical VAS (also ranging from 0 to 100)[5].

The design of our anchoring vignettes relies heavily on existing anchoring vignettes developed by Kapteyn *et al.* (2007) for the Gallup survey of 2011-2014. Because the Gallup survey does not primarily target student populations, we modify the anchoring vignettes in several ways to match the characteristics of the persons described in the vignettes to the characteristics of our student population. First, as highlighted by the International Report from the 2013/2014 Survey of the Health Behavior in School-aged Children (HBSC) Study, students' life satisfaction is associated with subjective health, family environment, relationship with peers and academic success. Age, gender and family income are also important factors for their life satisfaction. Based on these findings, we define vignettes with factors including school performance, quality of relationship with parents, number of close friends, minor health problems and family income. Second, we require all vignettes to share the same factors, but to various intensities. For example, we create three different degrees ("four", "two" and "none") to the factor "number of close friends" and assigned them to vignette 1, 2 and 3, respectively. Third, two factors, family income and health problems, are tailored to the Swiss context. More specifically, the family income we use in vignette 2 is equal to USD 10,079, which corresponds to the monthly average gross income per household in Switzerland in 2014[6]. Family incomes in vignettes 1 and 3 correspond to twice and half this amount, respectively. Eating disorder is one of the most prevalent diseases among teenagers and young adults in Switzerland and we therefore include this disorder in vignette 3 (Vust and Michaud, 2008). We include perceived obesity in vignette 2 while vignette 1 describes the scenario of someone who has no health problems. To encourage response consistency, we follow King *et al.* (2004), Salomon *et al.* (2004), Grol-Prokopczyk *et al.* (2011) and Au and Lorgelly (2014) and assign gender-specific vignettes to respondents. However, we show in section 6.4 that

[5]As suggested by Hopkins and King (2010), we asked respondents to rate these three anchoring vignettes prior to evaluating their own quality of life. This allows to prime respondents into interpreting the self-assessment question in a similar light and defining the response scale in a common way. Asking respondents to evaluate vignettes first therefore increases the chances that they have a more standardized conception of what quality of life is.

[6]See https://www.bfs.admin.ch in the "Household income and expenditure" section for details.

males' and females' evaluations of the QoL of the persons described in the vignettes do not depend on the sex of these fictitious persons. The three vignettes we use in our study can be found in Appendix A.

Figure 2 shows the kernel densities of the respondents' evaluations of our three vignettes as well as the averages of these three evaluations. The top panel displays the kernel densities of the average score of the three vignette evaluations by sex. The difference across sex is significant (p-value of the Kolmogorov-Smirnov test = 0.000), with females evaluating the three vignettes more positively than males on average. From the three lower panels, we can see that our three vignettes cover the range of possible QoL values very well. Vignette 1 is clearly considered as a description of someone who has a high QoL (mean of 89.74 for males and 92.68 for females) with evaluations averaging close to 100, corresponding to the "best possible QoL" scenario. Respondents consider vignette 2 as being the description of someone with a slightly above average QoL (mean of 63.92 for males and 68.84 for females). By contrast, vignette 3 describes the scenario of someone with a rather low QoL (mean of 24.99 for males and 26.09 for females), where the distributions of the evaluations are much closer to 0, corresponding to "worst possible QoL". Our Kolmogorov-Smirnov tests of equal distributions show that distributions for males and females are significantly different for vignettes 1 and 2, at 99% confidence, whereas we cannot reject the null hypothesis of equal distribution for the third one. The fact that females almost consistently evaluate these vignettes more positively (more probability mass closer to 100), as well as the interesting parallel shift to the right of about five points for females compared to males in vignette 2, and more generally the average of the three vignette evaluations in the first panel, suggest some preliminary evidence of reporting heterogeneity in QoL between the sexes. We will empirically test the presence of reporting heterogeneity and estimate its determinants in Section 4 after we describe how we account for it in our econometric section (Section 3).

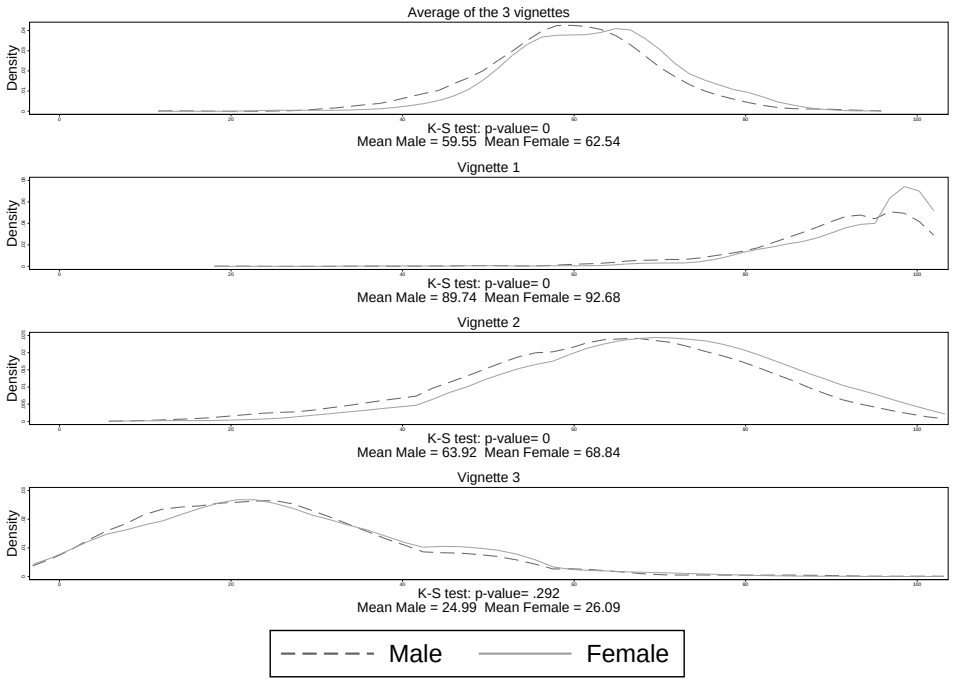

Figure 2: Panel 1: Kernel densities of the average scores of the three vignettes by sex. Panel 2: Kernel densities of vignette 1 by sex. Panel 3: Kernel densities of vignette 2 by sex. Panel 4: Kernel densities of vignette 3 by sex. 0 corresponds to the label "worst possible QoL" whereas 100 corresponds to the label "best possible QoL".

2.2 Control variables

In addition to being asked to evaluate their own QoL and the QoL of the three hypothetical individuals whose life scenarios are described in the vignettes, our questionnaire also covered a wide range of socio-demographic characteristics such as age, nationality, relationship status, quality of the relationship with their parents, their education level (undergraduate or graduate), family income, number of siblings, school performance, number of close friends, whether they think they were obese, whether they have already been diagnosed with eating disorder as well as the standard PhQ-9 questions to determine the presence and severity of depressive symptoms among our respondents (Kroenke *et al.*, 2001).

Table 1 reports the descriptive statistics of our study sample. Out of the 1,938 students who completed our online survey, we only kept the responses from individuals who were 25 or below in order to have a more homogeneous sample with respect to age. After dropping 122 observations that were out of our age range, our final sample consisted of

Table 1: Descriptive statistics

Variables		Male Mean or %	Female Mean or %	Difference significant at 95%†
Age		20.91	21.06	n
Swiss citizenship		47.69	55.44	y
Single		57.60	44.45	y
Number of friends	*None*	2.37	2.48	
	One	3.16	5.06	n
	Two	14.66	16.90	
	Three or more	79.82	75.57	
Relation with parents	*Good*	71.03	61.57	
	Good, but not always	26.16	34.98	y
	Bad	2.82	3.44	
School performance	*Above average*	31.00	23.57	
	Average	56.48	62.76	y
	Below average	12.51	13.67	
Depression (PHQ-9)	*No or minimum symptoms*	82.2	70.5	
	Mild	13.4	18.4	y
	Moderate	3.3	8.6	
	Severe	1.1	2.5	
Think are obese	*Yes*	9.02	13.56	
	No	89.06	83.10	y
	Don't know	1.92	3.34	
Eating disorder	*Yes*	2.25	10.55	
	No	95.38	85.58	y
	Don't know	2.37	3.88	
Family income (in USD)	*..., 4000]*	7.55	8.50	
	(4000,7000]	18.94	23.79	
	(7000,12000]	33.93	33.58	y
	(12000, ...	29.88	23.14	
	Don't know	9.70	10.98	
Number of Observations		887	929	

Note: Unweighted sample characteristics of the students who registered in the University of Lausanne's experiment program and completed our online survey in April 2017. Sample is restricted to respondents who are not older than 25 years of age. Total number of observations: 1,816.
† Differences in categorical variables were determined using Pearson's χ^2 tests. "y" stands for yes and "n" for no.

1,816 observations, of which a bit less than half were male (887, 45,8% of the sample). The average age of the respondents to our online questionnaire was about 21 years old. Females were more likely to be Swiss compared to males (55% vs 48%), less likely to be single (44% vs 58%) and had on average fewer close friends. Only about 3% of our respondents answered that they were in bad terms with their parents. Note also that more than half of our sample of students evaluated their school performance to be average. More females than males were mildly, moderately or severely depressed, thought that they were obese (14% vs 9%) and were diagnosed with eating disorder (11% vs 2%). About 34% of our sample came from a family that was making between USD 7,000.- and 12,000.- per month, which was in line with the monthly average gross income per household in Switzerland in 2014 (USD 10,079). Not reported in the table but also collected in our survey was information on the number of siblings as well as the school

and the program (undergraduate or graduate) the student was enrolled in the year of the online survey.

3 Econometric models

The descriptive statistics in the previous section show some preliminary evidence of the presence of reporting heterogeneity in self-reported QoL between sex as illustrated by the difference in vignette evaluations across sex in Figure 2. We now propose econometric models that can correct for reporting heterogeneity in self-reported VAS measures.

As a starting point, we assume that y_{i0}^*, the *latent* QoL of respondent i, can be modeled as a linear function of a set of factors x_i subject to an independent error term ϵ_{i0} that is not correlated with $\boldsymbol{x_i}$. In other words:

$$y_{i0}^* = \boldsymbol{x_i'} b_0 + \epsilon_{i0} \tag{1}$$

The researchers however do not observe y_{i0}^* but only observe the *reported* QoL of respondents, y_{i0}. The issue is that y_{i0} may not reflect the true underlying y_{i0}^* because of reporting heterogeneity, i.e., respondents with different characteristics $\boldsymbol{x_i}$ report value of y_{i0}^* using different reporting scales. By further assuming that reporting heterogeneity takes the form of an additive and unobserved individual effect c_i, the reported value of QoL of respondent i, y_{i0}, is given by:

$$y_{i0} = y_{i0}^* + c_i \tag{2}$$

As detailed below, reporting heterogeneity c_i may be a function of individual characteristics $\boldsymbol{x_i}$ as well as unobserved variables that are correlated with both $\boldsymbol{x_i}$ and y_{i0}^*.

In addition to these two general assumptions, the use of anchoring vignettes to identify reporting scales and correct for reporting heterogeneity also requires the very common vignette-related assumptions: vignette equivalence and response consistency (Bago d'Uva *et al.*, 2008; Kapteyn *et al.*, 2007; King *et al.*, 2004; Kristensen and Johansson, 2008; Rice

et al., 2012). Vignette equivalence assumes that vignettes are perceived in the same way by all respondents up to an idiosyncratic error term. This means that the characteristics of the respondents $\boldsymbol{x_i}$ do not affect the way respondents interpret the information that is given to them in the vignettes, or in other words, the *latent* QoL of the person described in vignette j, i.e., y^*_{ij} for $j = 1, ..., J$, does not depend on $\boldsymbol{x_i}$. Under this assumption, the perception of the *latent* QoL of the person described in vignette j can be written as:

$$y^*_{ij} = b_j + \epsilon_{ij} \quad \text{for} \quad j = 1, ..., J \tag{3}$$

where b_j represents the location of vignette j on the VAS and ϵ_{ij} the error term. It is important to note here that respondents can make "mistakes" (ϵ_{ij}) while evaluating the latent QoL of the hypothetical person described in vignette j, y^*_{ij} for $j = 1, ..., J$, but these mistakes should be idiosyncratic and hence should not depend on x_i.

On the other hand, response consistency assumes that respondents' evaluations of the anchoring vignettes are subject to the same reporting heterogeneity as their self-reported variable of interest. This implies that respondents use the same reporting scales when they evaluate their own characteristics as when they evaluate the characteristics of the person described in vignettes. Formally, this translates into:

$$y_{ij} = y^*_{ij} + c_i \quad \text{for} \quad j = 1, ..., J \tag{4}$$

This expression mirrors 2 in the sense that, under response consistency, c_i in 2 and 4 enters the expression for y_{i0} and y_{ij} for $j = 1, ..., J$ in the same way. While admittedly strong, vignette equivalence and response consistency are two assumptions that one has to make to be able to identify reporting heterogeneity (Bago d'Uva *et al.*, 2008; Kapteyn *et al.*, 2007; King *et al.*, 2004; Kristensen and Johansson, 2008; Rice *et al.*, 2012). We will test these two assumptions in Section 6 where we show evidence that they seem to hold in our study.

While statistical models that address reporting heterogeneity with anchoring vignettes in Likert scale measures have been extensively discussed in the literature, there are, to

the best of our knowledge, no studies that explore statistical models which can resolve the issue of reporting heterogeneity in VAS measures by using external information provided by anchoring vignettes. Using the set of assumptions described above, this paper proposes several alternative models to fill this gap in the literature. As detailed below, we first describe a model in which we can control for reporting heterogeneity by running a simple linear fixed-effect model (Model 1). This model allows to capture both observed and unobserved heterogeneity that are constant across vignette evaluations and self-assessment. We then define reporting heterogeneity c_i as a linear combination of covariates x_i and run a linear double-index model, which, in addition to controlling for reporting heterogeneity, also allows us to identify what the respondents' characteristics that explain reporting heterogeneity are (Model 2). After that, we relax the cardinality assumption that is implied in any linear models and propose a Generalized Ordered Response model (Model 3) and HOPIT models (Models 4 and 5), where VAS responses are assumed to be ordinal measures. Before diving in, we first follow the current literature on VAS measures and run a naive linear model that ignores reporting heterogeneity but assumes reporting homogeneity instead (Model 0).

3.1 Model 0 – Naive model

We first begin with a naive model in which we assume that there is no systematic reporting heterogeneity, i.e., c_i is not correlated with any x_i. Under this assumption, one can simply replace the expression for y^*_{i0} in 1 into 2 and regress y_{i0} on x_i to get a consistent estimate of the vector b_0.

Under the assumptions of vignette equivalence and response consistency, we can test the presence of reporting homogeneity by running ancillary regressions of vignette reports y_{ij} $(j = 1, ..., J)$ on x_i, where any significant effect of x_i would suggest a misspecification of the model. Indeed, under the vignette equivalence assumption, we know that x_i does not explain y^*_{ij} for $j = 1...J$, and hence plugging in y^*_{ij} from 3 into 4 results in

$$y_{ij} = b_j + c_i + \epsilon_{ij} \quad \text{for} \quad j = 1, ..., J \tag{5}$$

Therefore, if the assumption of vignette equivalence holds, then any effect of x_i on y_{ij} will go through c_i, which indicates the presence of systematic reporting heterogeneity by definition.

3.2 Model 1 – Linear fixed-effect model

Model 0 is based on the assumption that c_i is not correlated with any $\boldsymbol{x_i}$. If this assumption does not hold, Model 0 will yield inconsistent estimates of the vector b_0. However, one can obtain consistent estimate of b_0 by recognizing that c_i can be treated as a fixed effect. Indeed, by response consistency, we know that c_i is invariant in both self- and vignette evaluations, which implies that one can control for reporting heterogeneity by simply running a linear fixed-effect model. By plugging in 1 into 2 and 3 into 4, we obtain the following system of equations:

$$y_{i0} = \boldsymbol{x_i'} b_0 + c_i + \epsilon_{i0} \tag{6}$$

$$y_{ij} = b_j + c_i + \epsilon_{ij} \quad \text{for} \quad j = 1, ..., J \tag{7}$$

This model specification is equivalent to stacking our outcome variables together, considering the vignette evaluations as new observed outcome variables. Expressed at the individual level, the specification in matrix form with J, the total number of vignettes, equal to 3 can be written as follows:

$$\begin{pmatrix} y_{0i} \\ y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_i'} & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 0 \\ \mathbf{0} & 0 & 1 & 0 \\ \mathbf{0} & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{b_0} \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} c_i \\ c_i \\ c_i \\ c_i \end{pmatrix} + \begin{pmatrix} \epsilon_{0i} \\ \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \end{pmatrix} \tag{8}$$

where b_0 is the vector of coefficients of interest that represents the real effects of x_i on y_{0i}, net of reporting heterogeneity c_i. In a more compact form, 8 can be written as:

$$\boldsymbol{y_i} = \boldsymbol{x_i'} \boldsymbol{b} + \boldsymbol{c_i} + \boldsymbol{\epsilon_i} \tag{9}$$

By running a fixed-effect model, this linear specification allows to capture both observed and unobserved heterogeneity that are constant across vignette evaluations and self-assessment. This model is thus particularly appealing because it allows reporting behaviors c_i to be correlated with any observed ($\boldsymbol{x_i}$) or unobserved individual characteristics that determine the outcome variable of interest $\boldsymbol{y_i}$. In other words, c_i is purged out as it is the case in any standard linear fixed-effect panel estimation models where the researcher obtains several observations of the same individual over time. Another advantage of this model is that linear fixed-effect estimators are readily available in most software packages such that data containing VAS and corresponding VAS-based anchoring vignettes can be easily analyzed without a lot of programming work.

Note that for identification purposes, we assume that the coefficient associated with the constant term in the vector $\boldsymbol{x_i}$ is equal to zero.

3.3 Model 2 – Linear double-index model

The information provided by the anchoring vignettes however allows us to do more than just running linear fixed-effect models. Indeed, it is possible in our setting to identify what the characteristics of the individuals $\boldsymbol{x_i}$ that explain reporting heterogeneity c_i are and to quantify the magnitudes of these effects. To do so, one can assume that c_i can be modeled as a linear function of $\boldsymbol{x_i}$, i.e., $c_i = \boldsymbol{x_i'}\gamma$[7]. Taking the system of equation we have seen before (6 and 7) and replacing c_i by $\boldsymbol{x_i'}\gamma$ lead to the following expressions:

$$y_{i0} = \boldsymbol{x_i'}b_0 + \boldsymbol{x_i'}\gamma + \epsilon_{i0} \tag{10}$$

$$y_{ij} = b_j + \boldsymbol{x_i'}\gamma + \epsilon_{ij} \quad \text{for} \quad j = 1, ..., J \tag{11}$$

[7]One can equally assume that $c_i = \boldsymbol{x_i'}\gamma + d_i$, where d_i captures unobserved heterogeneity, because d_i would cancel out, as it was the case for c_i in the previous model.

As previously, expressed at the individual level, the specification of the above system in matrix form with $J = 3$ can be written as follows:

$$\begin{pmatrix} y_{0i} \\ y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_i'} & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 0 \\ \mathbf{0} & 0 & 1 & 0 \\ \mathbf{0} & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{b_0} \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} \boldsymbol{x_i'} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x_i'} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{x_i'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x_i'} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\gamma} \\ \boldsymbol{\gamma} \\ \boldsymbol{\gamma} \end{pmatrix} + \begin{pmatrix} \epsilon_{0i} \\ \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \end{pmatrix} \tag{12}$$

which can be more compactly rewritten as:

$$\boldsymbol{y_i} = \mathbf{x_i'}\boldsymbol{b} + \mathbf{x_i'^c}\boldsymbol{\gamma} + \boldsymbol{\epsilon_i} \tag{13}$$

It is worth noting here that the vector of coefficient b_0, which represents the real effects of $\boldsymbol{x_i}$ on y_{0i}, net of reporting heterogeneity, will be identical to the one in Model 1, as the value of $\boldsymbol{x_i'\gamma}$ is the same for self- and vignette evaluations in 12 by response consistency[8]. The advantage of this specification however is that, in addition to estimate b_0, we explicitly specify reporting heterogeneity c_i with a linear function of $\boldsymbol{x_i}$, which enables us to estimate the vector of coefficients γ, representing the effects of the characteristics of the individuals $\boldsymbol{x_i}$ on reporting heterogeneity c_i. This therefore allows us to acquire a better understanding of the source of reporting heterogeneity along with its magnitude. Again, we assume for the purpose of identification that the coefficient associated with the constant term in $\boldsymbol{x_i}$ is zero.

Another advantage of Model 2 is that it does not require anchoring vignettes to be evaluated by all respondents in a given survey, as opposed to Model 1. In cases where vignettes are evaluated only by a random subsample of respondents, one can use Model 2 to obtain consistent estimates of γ from that subsample and then predict the reporting heterogeneity of all the other respondents. The drawback is that it relies, as in all parametric specifications, on the correct specification of $\boldsymbol{x_i\gamma}$ for c_i, which can call for caution in some empirical studies.

[8] This is an application of Mundlak's results for individual effects in panel data models.

3.4 Model 3 – Generalized Ordered Response model

The models so far have relied on the assumption that VAS measures have the supposed interval and ratio properties[9] and can therefore be used in linear specifications. While the cardinality assumption of our VAS measure seems to hold in our application, as suggested by our test in section 6.3, this assumption may not hold in other applications (Craig *et al.*, 2009; Torrance *et al.*, 2001) and one may want to compare the results from Models 1 and 2 with results in models in which VAS measures are considered as ordinal measures. In the following models, we therefore depart from the cardinality assumption of our VAS measures, which was implied in our previous econometric specifications, and consider models in which only the ordinality of our VAS measure is assumed[10]. Model 3 corresponds to a Generalized Ordered Response model in which reporting heterogeneity is controlled for by allowing the thresholds that characterize the distance between y_{ij} and y^*_{ij} for $j = 0, ..., J$ to be individual-specific. More specifically, as we have seen in 1 and 3, the latent outcome variables of our specification are:

$$y^*_{i0} = \boldsymbol{x}'_i b_0 + \epsilon_{i0} \tag{14}$$

$$y^*_{ij} = b_j + \epsilon_{ij} \quad \text{for} \quad j = 1, ..., J \tag{15}$$

Ordinal observed responses are obtained by comparing the latent outcomes of interest with individual-specific cutoff points as follows:

$$y_{ij} = k \quad if \quad \boldsymbol{x}'_i\boldsymbol{\delta} + \alpha_{k-1} < y^*_{ij} \leq \boldsymbol{x}'_i\boldsymbol{\delta} + \alpha_k \quad \text{for} \quad k = 0, 1, 2, ..., K \quad \text{and} \quad j = 0, ..., J \tag{16}$$

with $\alpha_{-1} = -\infty$, $\alpha_K = \infty$ and where $\boldsymbol{x}'_i\boldsymbol{\delta}$ is the individual-specific index that controls for reporting heterogeneity, J being the number of vignettes at disposal and K the number of response categories, which is equal to 100 in our VAS measure. We further assume for

[9] VAS measures contain the properties of interval scale because VAS are defined as numerical scales for which the distance between two points can be interpreted as an interval. VAS measures also contain the ratio properties as VAS have a starting point, that is a zero value, which allows ratios to be calculated.

[10] As we did in the linear case, we also provide results of a naive model –a model that does not control for reporting heterogeneity– in a ordinal setting. Mirroring Model 0, we define Model 0' as a model in which we estimate y_{i0} with a simple ordered probit model.

the purpose of identification that the coefficient associated with the constant term in $\boldsymbol{x_i'}$ is zero. Under the assumption that ϵ_{ij} for $j = 0, ..., J$ is distributed as $N(0, \sigma^2)$, we can write the corresponding likelihood function as:

$$\begin{aligned} & \mathcal{L}(\boldsymbol{b_0}, b_1, \ldots, b_J, \alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{K-1}, \boldsymbol{\delta}, \sigma) \\ = & \prod_{k=0}^{K} \left(\Phi\left(\frac{\boldsymbol{x_i'\delta} + \alpha_k - \boldsymbol{x_i'b_0}}{\sigma} \right) - \Phi\left(\frac{\boldsymbol{x_i'\delta} + \alpha_{k-1} - \boldsymbol{x_i'b_0}}{\sigma} \right) \right)^{\mathbb{I}(y_{i0}=k)} \quad (17) \\ & \times \prod_{j=1}^{J} \prod_{k=0}^{K} \left(\Phi\left(\frac{\boldsymbol{x_i'\delta} + \alpha_k - b_j}{\sigma} \right) - \Phi\left(\frac{\boldsymbol{x_i'\delta} + \alpha_{k-1} - b_j}{\sigma} \right) \right)^{\mathbb{I}(y_{ij}=k)} \quad (18) \end{aligned}$$

where $\Phi(.)$ is the cumulative distribution function (cdf) of the standard normal distribution. Line 17 corresponds to the contribution of the self-evaluation into the likelihood function and line 18 corresponds to the contribution of each of the vignettes j.

Note that parameters of the real effects of $\boldsymbol{x_i'}$ on y_{0i} ($\boldsymbol{b_0}$) as well as the effects of $\boldsymbol{x_i}$ on reporting heterogeneity ($\boldsymbol{\delta}$) are only identified up to scale with scale parameter (σ) being unknown. To compare results across models, we therefore report estimates of b_0 and δ by rescaling them using our estimate of σ from Model 1[11].

Using a single index to control for reporting heterogeneity as we do in this specification is referred to as index shift (Lindeboom and van Doorslaer, 2004) and allows for parallel shift in the cutoff of our ordered response model. As can be seen in 16, our model controls for reporting heterogeneity by allowing the cutoff points to be individual-specific, as represented by $\boldsymbol{x_i'\delta}$. However, this specification limits these individual cutoffs to be parallel, i.e., for all individuals i, the distance between adjacent cutoffs is fixed and equal to $\alpha_k - \alpha_{k-1}$[12]. This model specification is therefore equivalent to adding the single index $\boldsymbol{x_i'\delta}$ into the latent outcome equations and keeping the cutoffs constant, as assumed in an ordinary ordered response model. In fact, this model specification is similar to the one in Model 2, except that the coefficients in our equations are here estimated with an ordered response model instead of a linear regression.

As just mentioned, Model 3 defines the thresholds in our ordered response model to

[11] This is sufficient to ensure comparability across models since Ordered Response models are identified up to scale. Note also that the normalization of location is not necessary as it only affects the estimates of the constant terms.

[12] The distance between adjacent cutoffs is equal to $\boldsymbol{x_i'\delta} + \alpha_k - (\boldsymbol{x_i'\delta} + \alpha_{k-1}) = \alpha_k - \alpha_{k-1}$.

be parallel, that is to say, we allow for reporting heterogeneity to be uniform along the distribution of y_{ij} for $j = 0, ..., J$. We therefore depart from this restriction in Model 4 and allow for non-uniform reporting heterogeneity by defining shifts in the cutoff points that are individual-specific and not parallel.

3.5 Model 4 – Linear HOPIT model

While a Generalized Ordered Response model such as Model 3 can control for reporting heterogeneity by allowing the cutoffs in the relation between y_{ij} and y^*_{ij} for $j = 0, ..., J$ to be individual-specific, it does however not allow the cutoff shifts to be non-parallel even though they very well may be as a result of non-uniform reporting heterogeneity (Lindeboom and van Doorslaer, 2004). To account for non-parallel cutoff shifts, we first use a linear Hierarchical Ordered Probit (HOPIT) model which allows each cutoff to be modeled as different linear functions of $\boldsymbol{x'_i}$.

Specifically, we assume the same latent response indexes as in Model 3 (14 and 15) but allow the distance between the cutoffs to depend on $\boldsymbol{x'_i}$. The observed responses are therefore characterized as:

$$y_{ij} = k \quad if \quad \tau_i^{k-1} < y^*_{ij} \leq \tau_i^k \quad \text{for} \quad k = 0, 1, 2, ..., K \quad \text{and} \quad j = 0, ..., J \tag{19}$$

with K the number of cutoffs and J the number of vignettes. The cutoff points τ_i^k are given by:

$$\tau_i^k = \boldsymbol{x'_i \zeta_k} \quad \text{for} \quad k = 0, 1, 2..., K-1 \tag{20}$$

with $\tau_i^{-1} = -\infty$, $\tau_i^K = \infty$ and where ζ_k represents the vector of the effects of x_i on reporting heterogeneity at the cutoff k[13].

[13]The distance between adjacent cutoffs is now equal to $\tau_i^k - \tau_i^{k-1} = \boldsymbol{x'_i \zeta_k} - \boldsymbol{x'_i \zeta_{k-1}}$, which depends on $\boldsymbol{x'_i}$ in general.

The corresponding likelihood function for this specification is then:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{b_0}, b_1, \ldots, b_J, \boldsymbol{\zeta_0}, \boldsymbol{\zeta_1}, \boldsymbol{\zeta_2}, \ldots \boldsymbol{\zeta_{K-1}}, \sigma) \\
= &\prod_{k=0}^{K} \left(\Phi\left(\frac{\tau_i^k - \boldsymbol{x_i'b_0}}{\sigma} \right) - \Phi\left(\frac{\tau_i^{k-1} - \boldsymbol{x_i'b_0}}{\sigma} \right) \right)^{\mathbb{I}(y_{i0}=k)} \qquad (21) \\
&\times \prod_{j=1}^{J} \prod_{k=0}^{K} \left(\Phi\left(\frac{\tau_i^k - b_j}{\sigma} \right) - \Phi\left(\frac{\tau_i^{k-1} - b_j}{\sigma} \right) \right)^{\mathbb{I}(y_{ij}=k)} \qquad (22)
\end{aligned}$$

As we did previously, we assume for the purpose of identification that the coefficient associated with the constant term in the vector x_i is zero and we report the rescaled estimates using σ from Model 1[14].

The specification in Model 4, which specifies thresholds as a linear function of covariates as in Ierza (1985), Iburg *et al.* (2001), Peracchi and Rossetti (2013) and Pudney and Shields (2000), assumes monotonicity in the cutoff points $\tau_i^k = \boldsymbol{x_i'\zeta_k}$, that is $\tau^{k-1} \leqslant \tau^k$ for $k = 0, 1, 2..., K$ (see 19). It is possible however that when estimating Model 4, one gets inconsistent cutoffs in the sense that the monotonicity assumption is violated for some combinations of $\boldsymbol{x_i'}$. Although this assumption can be tested *ex post* by computing the estimated cutoffs, it may be preferable to impose such monotonicity assumption prior to estimating the model, as explained in our next model[15]. In the next specification, we modify the linear nature of this HOPIT model and impose an exponential structure to it, such that $\tau^{k-1} < \tau^k$ for $k = 0, 1, 2..., K$ is imposed by definition.

3.6 Model 5 – Exponential HOPIT model

In this model, we consider a HOPIT model which is similar to Model 4 but with an exponential component in its cutoffs. The advantage of this version of the HOPIT model is that the estimated cutoffs are guaranteed to be increasing. In particular, we define the observed responses by:

$$y_{ij} = k \quad if \quad \tau_i^{k-1} < y_{ij}^* \leq \tau_i^k \quad \text{for} \quad k = 0, 1, 2, ..., K \quad \text{and} \quad j = 0, ..., J \qquad (23)$$

[14]This is sufficient to ensure comparability across models since HOPIT models are identified up to scale.

[15]Note that if the assumption of monotonicity is violated, one could also maximize the likelihood function by imposing constraints in the parameters.

which is the same as Model 4. But now we assume the cutoffs to take different functional forms as:

$$\tau_i^{-1} = -\infty \tag{24}$$

$$\tau_i^{K} = \infty \tag{25}$$

$$\tau_i^{0} = \boldsymbol{x_i'\eta_1} \tag{26}$$

$$\tau_i^{k} = \tau_i^{k-1} + \exp\left(\boldsymbol{x_i'\eta_k}\right) \quad \text{for} \quad k = 1, 2, 3..., K-1 \tag{27}$$

This specification clearly imposes $\tau^{k-1} < \tau^{k}$ for $k = 0, 1, 2..., K$. As before, the corresponding likelihood function is defined as:

$$\begin{aligned} & \mathcal{L}(\boldsymbol{b_0}, b_1, \ldots, b_J, \boldsymbol{\eta_0}, \boldsymbol{\eta_1}, \boldsymbol{\eta_2}, ...\boldsymbol{\eta_{K-1}}, \sigma) \\ = & \prod_{k=0}^{K} \left(\Phi\left(\frac{\tau_i^k - \boldsymbol{x_i'b_0}}{\sigma}\right) - \Phi\left(\frac{\tau_i^{k-1} - \boldsymbol{x_i'b_0}}{\sigma}\right)\right)^{\mathbb{I}(y_{i0}=k)} && (28) \\ & \times \prod_{j=1}^{J} \prod_{k=0}^{K} \left(\Phi\left(\frac{\tau_i^k - b_j}{\sigma}\right) - \Phi\left(\frac{\tau_i^{k-1} - b_j}{\sigma}\right)\right)^{\mathbb{I}(y_{ij}=k)} && (29) \end{aligned}$$

For identification purposes, we assume that the coefficient associated with the constant term in the vector x_i is zero and we report the rescaled estimates using σ from Model 1.

Essentially, Model 4 and Model 5 assume different indexes in each cutoff and are two out of many multiple-index models. The clear advantage of the HOPIT model is that it can better capture the non-linear effects of reporting heterogeneity in comparison to the Generalized Ordered Response model that assumes parallel shifts. The HOPIT models are however very data-demanding as they estimate vectors of coefficients ζ or η for each of the k cutoffs of our VAS, therefore complicating direct applications of HOPIT models to VAS measures.

In the empirical application that follows, we discretize our VAS into 5 categories ($K = 5$) to make our estimations of Model 4 and 5 more tractable. In the main results we present below, we partition our VAS in five by using 20, 40, 60 and 80 as thresholds to discretize our VAS measure. As robustness check, instead of discretizing our VAS in 20 points interval, we estimate Model 5 by considering two alternative data-driven ways

to partition the VAS. We first discretize the VAS in quintiles to impose the number of observations to be similar in each of the five groups we create. We also consider the case where we dicretize our VAS measure based on natural groupings using partition clustering methods (k-means clustering) (Jain, 2010).

4 Results

Table 2 column 1 presents the results of our naive model (Model 0) in which we regress our VAS measure of QoL on our set of control variables. Columns 2, 3 and 4 present the regressions of the three vignette evaluations on the same set of control variables. Column 1 shows that on average, females report having higher QoL than males by about 1.4 points on our 0-100 measure scale and this effect is statistically significant at conventional levels. Younger respondents, those who are in couple and those from families with higher income also report having higher QoL, with individuals from families that earns more than USD 12,000 per month reporting their QoL to be about 10 points higher than those from our omitted income category, i.e., from families in which parents were making less than USD 4,000 per month. All these effects are highly statistically significant. Other effects are in the direction one would expect: respondents who perceive themselves as obese, who suffer from eating disorder and depression, who have relatively lower school performances and bad relationships with their parents all report lower QoL. It is also worth noting how large the effect of being depressed on QoL is, with a drop of more than 23 points for those who suffer from severe depression compared to those with no depressive symptoms.

Columns 2 to 4 present the results of the regressions of the vignette evaluations on the same set of control variables. We can see that some of our control variables are statistically significant, suggesting the presence of reporting heterogeneity as explained in the econometric section above. Indeed, under the vignette equivalence assumption, any significant effect of x_i on y_{ij} $(and\ not\ y^*_{ij})$ for $j = 1, ..., J$ must come from reporting heterogeneity c_i. The variable that clearly stands out in these three regressions is sex. Females consistently rate the three vignettes higher compared to males, which is in line

Table 2: Results of the estimation of Model 0 and the regressions of the vignette evaluations on our set of control variables

	(1) Model 0	(2) Vignette 1	(3) Vignette 2	(4) Vignette 3
Female	1.426**	2.993***	4.975***	1.062
	(0.666)	(0.471)	(0.775)	(0.799)
Age	−0.479**	0.067	0.536**	−0.144
	(0.208)	(0.148)	(0.246)	(0.244)
Single	−1.511**	−0.053	−0.497	−0.980
	(0.651)	(0.453)	(0.778)	(0.759)
Family income:				
- (4000 7000]	3.022**	1.567	−0.944	−1.713
	(1.500)	(1.144)	(1.486)	(1.786)
- (7000 12000]	7.087***	1.225	−2.482*	−2.653
	(1.417)	(1.127)	(1.389)	(1.749)
- More than 12000	9.793***	−1.032	−5.077***	−3.560**
	(1.441)	(1.172)	(1.474)	(1.780)
Perceived obese: No	2.520**	0.279	−1.403	1.942
	(1.148)	(0.822)	(1.237)	(1.227)
Eating disorder: No	2.253	1.889*	4.830***	−0.342
	(1.573)	(1.131)	(1.659)	(1.751)
Depression:				
- Mild	−4.951***	−1.196*	−2.277**	0.183
	(0.977)	(0.684)	(1.075)	(1.075)
- Moderate	−9.883***	−1.301	−2.623	−1.978
	(1.881)	(1.178)	(1.668)	(1.756)
- Severe	−23.021***	−0.072	−1.960	−2.390
	(4.285)	(1.596)	(2.865)	(2.303)
School performance				
- Average	−1.299*	−0.228	0.796	0.361
	(0.691)	(0.515)	(0.855)	(0.870)
- Below average	−2.392**	0.834	1.137	0.928
	(1.197)	(0.740)	(1.294)	(1.270)
Relation with parents:				
- Good but not always	−3.735***	0.278	−0.667	−0.274
	(0.727)	(0.488)	(0.855)	(0.823)
- Bad	−13.602***	−0.559	1.223	0.017
	(2.894)	(1.824)	(2.148)	(2.296)
Constant	81.864***	84.670***	55.073***	23.785***
	(5.985)	(4.561)	(6.938)	(6.772)
R-squared	0.233	0.061	0.059	0.020

Note: Robust standard errors are reported in parenthesis * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. In all regressions, we also control for education level (Bachelor, Master, others), number of siblings, number of close friends and origin (coefficients not reported but available upon request). Column 1 corresponds to the estimation of our naive model (Model 0) in which we assume there is no reporting heterogeneity. Columns 2, 3 and 4 correspond to the regressions of vignette 1, 2 and 3 on our set of control variables, respectively.

with what is illustrated in Figure 2, even after controlling for our set of control variables[16]. All in all, our preliminary analysis suggests that females report having a higher QoL (column 1) and at the same time tend to evaluate our three vignettes (columns 2-4) more positively. This indicates that there exists some reporting heterogeneity, notably between the sexes, in self-evaluated QoL[17]. The following estimations of Models 1 to 5 will account

[16]Given the relatively low R^2 in the regression of vignette 3 –a vignette that describes a rather extreme scenario– it is not surprising that most variables including female are insignificant.

[17]It is worth noting that females in our sample suffer more from depression than males (see Table 1). It is therefore possible that the sex gap we find in column 1 comes from the fact that we control for levels of depression in our analysis, which have strong negative effects on QoL, rather than from the reporting heterogeneity. We show however in our

Table 3: Results of the estimation of Model 0 to Model 5

	(1) Model 0	(2) Model 1	(3) Model 2	(4) Model 0'	(5) Model 3	(6) Model 4	(7) Model 5
Female	1.426**	−1.584**	−1.584**	1.438**	−1.705**	−2.181**	−2.259**
	(0.666)	(0.736)	(0.737)	(0.631)	(0.668)	(0.916)	(0.918)
Age	−0.479**	−0.632***	−0.632***	−0.389**	−0.600***	−0.760***	−0.715**
	(0.208)	(0.234)	(0.234)	(0.194)	(0.209)	(0.282)	(0.282)
Single	−1.511**	−1.001	−1.001	−1.659***	−1.178*	−1.209	−1.240
	(0.651)	(0.733)	(0.735)	(0.623)	(0.665)	(0.896)	(0.896)
Family income:							
- (4000 7000]	3.022**	3.386**	3.386**	2.919**	2.612*	2.326	2.260
	(1.500)	(1.678)	(1.682)	(1.264)	(1.479)	(1.704)	(1.703)
- (7000 12000]	7.087***	8.390***	8.390***	6.730***	7.366***	6.529***	6.503***
	(1.417)	(1.574)	(1.578)	(1.235)	(1.402)	(1.682)	(1.680)
- More than 12000	9.793***	13.016***	13.016***	9.575***	12.205***	12.782***	12.736***
	(1.441)	(1.639)	(1.642)	(1.288)	(1.475)	(1.801)	(1.802)
Perceived obese: No	2.520**	2.247*	2.247*	2.324**	2.085**	2.225*	2.297*
	(1.148)	(1.216)	(1.218)	(0.990)	(1.050)	(1.338)	(1.339)
Eating disorder: No	2.253	0.128	0.128	1.895	0.028	−0.714	−0.652
	(1.573)	(1.688)	(1.691)	(1.389)	(1.475)	(1.713)	(1.715)
Depression:							
- Mild	−4.951***	−3.855***	−3.855***	−4.379***	−3.253***	−4.441***	−4.395***
	(0.977)	(1.080)	(1.082)	(0.896)	(0.959)	(1.175)	(1.177)
- Moderate	−9.883***	−7.916***	−7.916***	−8.504***	−6.774***	−7.991***	−8.039***
	(1.881)	(2.024)	(2.028)	(1.470)	(1.689)	(1.765)	(1.769)
- Severe	−23.021***	−21.547***	−21.547***	−15.009***	−15.908***	−14.327***	−14.469***
	(4.285)	(4.369)	(4.378)	(3.117)	(3.476)	(2.987)	(2.988)
School performance							
- Average	−1.299*	−1.608**	−1.608**	−1.492**	−1.669**	−0.986	−0.972
	(0.691)	(0.795)	(0.796)	(0.697)	(0.736)	(1.043)	(1.043)
- Below average	−2.392**	−3.359***	−3.359***	−2.235**	−3.315***	−2.762*	−2.801*
	(1.197)	(1.292)	(1.295)	(1.076)	(1.156)	(1.472)	(1.472)
Relation with parents:							
- Good but not always	−3.735***	−3.515***	−3.515***	−4.035***	−3.663***	−3.515***	−3.578***
	(0.727)	(0.782)	(0.784)	(0.688)	(0.711)	(0.961)	(0.961)
- Bad	−13.602***	−13.829***	−13.829***	−10.827***	−11.847***	−11.756***	−11.681***
	(2.894)	(3.100)	(3.106)	(2.121)	(2.547)	(2.323)	(2.329)
Constant	81.864***	88.432***	81.864***			−0.333	−1.200
	(5.985)	(6.433)	(5.961)			(7.638)	(7.641)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Model 0: Naive linear model. Model 1: Linear fixed-effect model. Model 2: Linear double-index model. Model 0': Native ordered probit model. Model 3: Generalized Ordered Response model. Model 4: Linear HOPIT model. Model 5: Exponential HOPIT model. Robust standard errors are reported in parenthesis for Model 0, 0', 4 and 5. Cluster robust standard errors at the individual level are reported in parenthesis for Model 1, 2 and 3. The estimates of Model 0', 3, 4 and 5 have been rescaled using the standard deviations calculated in Model 1. See the text for more details. In all regressions, we also control for education level (Bachelor, Master, others), number of siblings, number of close friends and origin (coefficients not reported but available upon request).

for the presence of reporting heterogeneity.

Column 1 of Table 3 reports the results of our naive Model 0 in which reporting heterogeneity is not controlled for. Column 2 reports the results of our linear fixed-effect model (Model 1) which exploits the vignette evaluations as additional observations for each individual to control for unobserved individual effects that are constant over evaluations, i.e., individual reporting heterogeneity in our context. As one can see in column 2, the coefficients of most of the control variables are of the same sign and about the same magnitude as the coefficients estimated in Model 0. One exception stands out however: sex. When using information from the vignettes to control for

robustness check analysis that omitting depression in our estimation doesn't change our conclusion on the presence of sex-specific reporting heterogeneity.

reporting heterogeneity, the sign of the coefficient associated with females reverses and becomes negative, going from +1.426 to -1.584. This implies that females appear to have a relatively lower QoL compared to males when reporting heterogeneity is accounted for by using a linear fixed-effect model. Note also that these two coefficients are both statistically significant at 95%.

We then run our estimation of Model 2 (column 3). As explained in the econometric section, the coefficients of the vector b_0 derived from Model 2 are the same as the coefficients in Model 1 (column 2). The advantage of estimating Model 2 is that it allows us to identify the factors that explain reporting heterogeneity, results that we will describe below.

We then depart from the cardinality assumption implied in linear regressions and estimate models that consider the VAS measures as being ordinal. Before estimating Model 3, we estimate another naive model using a standard ordered probit model, Model 0', in which reporting heterogeneity is not controlled for. The results of that naive model are presented in column 4. One can see that once again, females appear to have a higher QoL than males when reporting heterogeneity is not accounted for. These results are very similar to the ones in Model 0 (column 1).

Column 5 of Table 3 reports the results from our Generalized Ordered Response model (Model 3)[18]. Assuming parallel cutoff shifts, one can see that the coefficients we get are very similar to the ones derived in our fixed-effect model (column 2). In this specification, females again appear to have a lower QoL than males, by about 1.7 points, once we control for reporting heterogeneity. Same holds when estimating our HOPIT models (columns 6 and 7), although the coefficient associated with being a female is somewhat more negative than previous estimates. Overall, when imposing females and males to use the same reporting scale and therefore accounting for reporting heterogeneity, females consistently appear to have a lower QoL compared to males by about 1.6-2.3 points on our VAS 0-100 measure, and these effects are all significant at 95% confidence.

[18]In order to make coefficients comparable to previous models, we rescale the estimated coefficients and standard errors of Models 3, 4 and 5 by multiplying them with standard errors we estimate from Model 1. We can proceed to this rescaling because of the assumption regarding the distribution of the error terms we make.

Table 4: Estimates of reporting heterogeneity in Models 2, 3, 4 and 5

	Model 2	Model 3	Model 4				Model 5			
Female	3.010***	3.163***	1.710**	2.670***	2.724***	3.159***	1.725**	0.588	0.539	0.298
	(0.484)	(0.441)	(0.701)	(0.651)	(0.600)	(0.596)	(0.703)	(0.723)	(0.743)	(0.540)
Age	0.153	0.202	-0.335	-0.125	0.298	0.210	-0.237	0.140	0.325	-0.049
	(0.151)	(0.138)	(0.219)	(0.201)	(0.187)	(0.183)	(0.216)	(0.220)	(0.225)	(0.164)
Single	-0.510	-0.398	-0.833	-0.802	-0.875	-0.101	-1.095	0.542	-0.374	0.673
	(0.475)	(0.436)	(0.680)	(0.627)	(0.582)	(0.577)	(0.689)	(0.710)	(0.719)	(0.530)
Family income:										
- (4000 7000]	-0.364	0.231	-0.937	-2.832**	0.090	0.214	-1.338	-1.006	3.014**	-0.056
	(0.984)	(0.906)	(1.431)	(1.274)	(1.158)	(1.144)	(1.467)	(1.510)	(1.465)	(1.013)
- (7000 12000]	-1.303	-0.781	-2.305*	-1.998	-1.130	-0.567	-2.634*	0.929	0.538	0.438
	(0.938)	(0.857)	(1.396)	(1.242)	(1.128)	(1.118)	(1.424)	(1.469)	(1.363)	(0.991)
- More than 12000	-3.223***	-2.915***	-2.490*	-3.456***	-2.617**	-3.253***	-2.863*	-0.147	0.339	-0.485
	(0.982)	(0.892)	(1.451)	(1.302)	(1.181)	(1.178)	(1.501)	(1.580)	(1.482)	(1.053)
Perceived obese: No	0.273	0.144	0.050	1.810*	-0.198	0.054	0.332	1.190	-2.154*	0.291
	(0.779)	(0.740)	(1.065)	(0.985)	(0.909)	(0.892)	(1.059)	(1.049)	(1.166)	(0.791)
Eating disorder: No	2.126**	1.894**	2.006	0.389	3.346***	2.995***	2.021	-1.189	2.403*	-0.148
	(1.046)	(0.961)	(1.362)	(1.227)	(1.138)	(1.143)	(1.401)	(1.521)	(1.314)	(1.059)
Depression:										
- Mild	-1.097	-1.062*	0.198	-1.684**	-1.727**	-0.939	0.306	-1.639*	-0.345	0.825
	(0.668)	(0.608)	(0.953)	(0.857)	(0.785)	(0.780)	(0.979)	(0.962)	(0.991)	(0.725)
- Moderate	-1.967*	-1.845*	-3.928***	-1.226	-0.255	-1.940	-4.280***	3.647**	0.604	-1.372
	(1.065)	(0.984)	(1.402)	(1.275)	(1.205)	(1.231)	(1.436)	(1.685)	(1.511)	(1.042)
- Severe	-1.474	-1.278	-1.022	-4.576**	-1.659	-1.329	-2.174	-1.442	3.275	0.060
	(1.587)	(1.480)	(2.415)	(2.205)	(2.100)	(2.104)	(2.414)	(2.236)	(2.751)	(2.019)
School performance										
- Average	0.309	0.270	-0.075	-0.072	-0.466	-0.063	-0.054	0.146	-0.743	0.457
	(0.528)	(0.483)	(0.781)	(0.726)	(0.668)	(0.664)	(0.782)	(0.799)	(0.832)	(0.602)
- Below average	0.966	1.051	0.124	0.838	0.580	0.884	0.041	0.951	-0.631	0.367
	(0.803)	(0.746)	(1.128)	(1.033)	(0.971)	(0.964)	(1.131)	(1.184)	(1.198)	(0.869)
Relation with parents:										
- Good but not always	-0.221	-0.045	-0.180	0.419	-0.639	-0.442	-0.208	0.313	-0.627	0.010
	(0.506)	(0.469)	(0.743)	(0.690)	(0.640)	(0.627)	(0.751)	(0.788)	(0.773)	(0.567)
- Bad	0.227	0.448	0.458	-1.323	1.078	2.027	0.972	-2.119	2.577	1.208
	(1.369)	(1.263)	(1.910)	(1.678)	(1.602)	(1.603)	(1.928)	(1.751)	(2.175)	(1.526)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Model 2: Linear double-index model. Model 3: Generalized Ordered Response model. Model 4: Linear HOPIT model. Model 5: Exponential HOPIT model. Cluster robust standard errors at the individual level are reported in parenthesis for Models 2 and 3. The estimates of Models 3, 4 and 5 have been rescaled using the standard deviations calculated in Model 1. See the text for more details. In all regressions, we also control for education level (Bachelor, Master, others), number of siblings, number of close friends and origin (coefficients not reported but available upon request). Note that by construction, the parameters associated to reporting behavior in Model 3 (δ), Model 4 (ζ_1 to ζ_4) and Model 5 (η_1 to η_4) have opposite signs to the ones in Model 2. We therefore multiply by (-1) the vectors γ, ζ, η in Models 3,4 and 5 to make the results directly comparable to γ in Model 2.

As described in the econometric section, one of the advantages of Models 2, 3, 4 and 5 is that they not only determine the direction of the reporting heterogeneity, but also its magnitude. Table 4 reports the estimates of reporting heterogeneity itself, i.e., the effects of x_i on c_i (the vectors of coefficients γ, δ, ζ and η[19]). Females, as is expected from the results displayed in Table 3, tend to use different reporting scales when evaluating their QoL compared to males, with a shift of about 3 points to the right (column 1). This also holds true for the reporting heterogeneity we estimate in Model 3 (column 2). These effect are statistically significant at 99% confidence. When looking at reporting heterogeneity that is driven by other individual characteristics, we can see that respon-

[19]It is also worth noting that column 1 in Table 4 corresponds to the difference between the coefficients of Model 1 (or 2) and Model 0.

dents from comparatively rich families report their QoL to be lower than what they truly is. Students from parents who make at least USD 12,000 per month for instance significantly under-report their QoL by about 3 points[20]. Same conclusion holds for those with mild and moderate depression, who under-report their QoL by about 1 and 2 points, respectively, although these effects fail to be significant at conventional statistical levels.

In Table 4, columns 3 and onwards show the results of our HOPIT models in which we report the effects of x_i on reporting heterogeneity for each cutoff (ζ_i and η_i for $i = 1, 2, 3, 4$). That is, we allow for non-uniform reporting heterogeneity and therefore permit the cutoff shifts in the relation between the reported and latent QoL to be non-parallel. Again, females consistently and significantly have higher cutoff estimates by about 1,7-3,2 points compared to males. Although not directly comparable because of the different structures of the specification of the cutoffs, the coefficients associated with sex in the linear and exponential versions of the HOPIT models display very similar patterns[21]. This indicates that once we control for reporting heterogeneity and compare males and females using the same reporting scale, females appear to have a lower QoL than males, and this holds irrespective of the estimated models we consider and their underlying assumptions.

5 Fitted and counterfactual distributions

5.1 Model fit

From the results above, we can compute the fitted values of the QoL that we derive from Model 2 and compare them with the original respondents' QoL evaluation in Figure 1. Figure 3 shows the kernel densities of these fitted values for both males (dashed line) and females (solid line). We can see that our model performs pretty well at reproducing the overall shapes of the distributions displayed in Figure 1. Again, the distribution for

[20] A possible explanation for this under-reporting of individuals coming from high income families is adaptation or habit formation which reflects the fact that respondents with richer parents may need more money to achieve the same level of perceived QoL than respondents whose parents are less well-off.

[21] The insignificant coefficients of η_2, η_3 and η_4 in the exponential HOPIT model (Columns 8-10) imply that there are no incremental effects; the cutoff shift is still significant as indicated by the coefficient associated with sex in η_1 (Column 7).

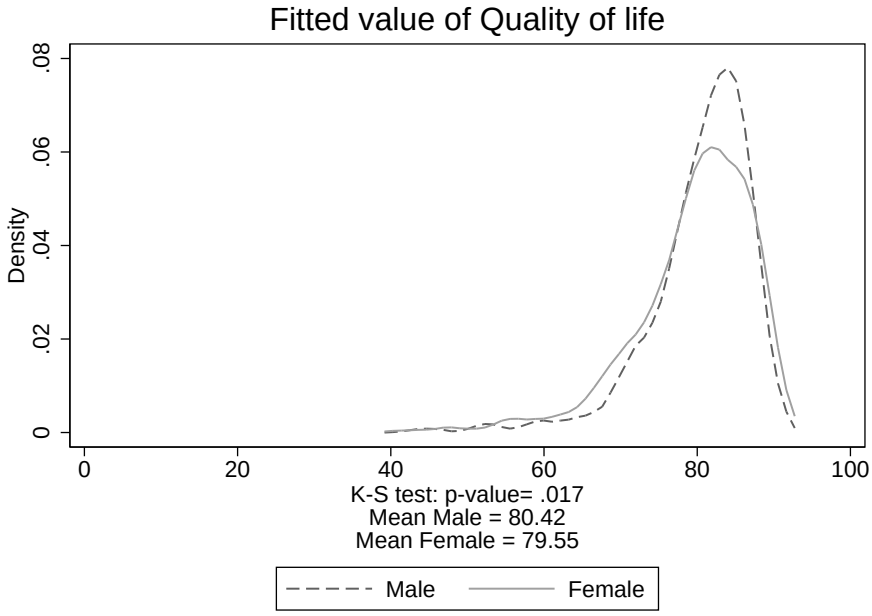

Figure 3: Fitted value of QoL estimated in Model 2

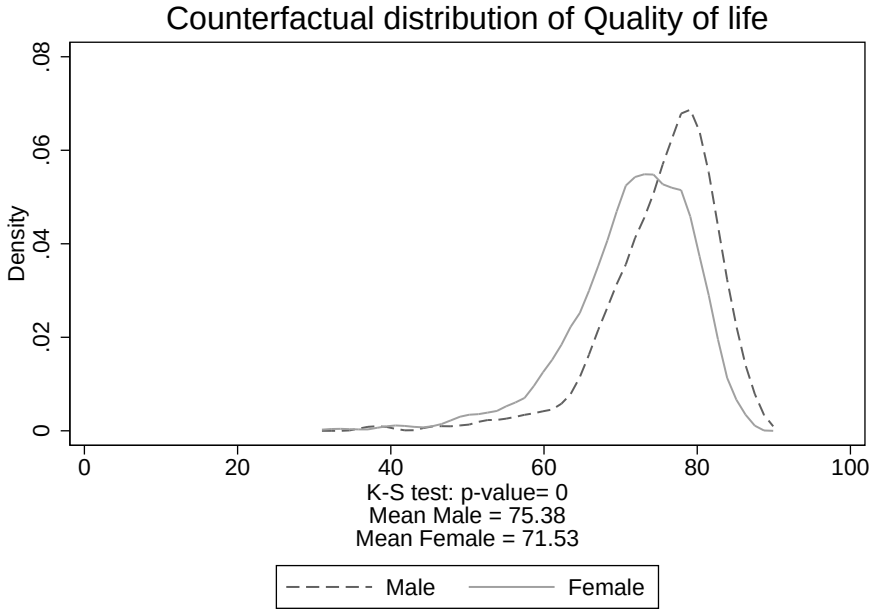

Figure 4: Counterfactual kernel densities of QoL

males seems to be more centered around its mean and the one for females is more skewed towards the left. Although the magnitude of the difference between the two distributions is a bit higher in the range 80-85 compared to the original observations, our model fits nicely with the observations.

5.2 Counterfactual simulations

Based on these estimates, we simulate the counterfactual distribution of QoL of the respondents, had they had the same reporting scale, that is free from any reporting heterogeneity. Figure 4 illustrates the counterfactual distributions of QoL by sex based on the parameters estimated in Model 2. On average, females would have reported their QoL to be nearly 4 points lower than males, had they used the same reporting scale to report their QoL as males. Contrasting these two distributions to the ones in Figure 3 clearly shows a shift to the left for females, again revealing that females are worse off compared to males in terms of QoL. Note also that the sex gap in Figure 4 (about 4 points on average) is larger than the estimate of the sex effect in Model 2 (Table 4 column 2). This is due to the sex differences in the demographic characteristics of our study sample, such as higher prevalence of depression and eating disorder among females, which strongly and negatively affects QoL, as shown in our model estimates.

6 Tests of model assumptions and robustness checks

6.1 Tests of vignette equivalence

In order to be able to test the vignette equivalence assumption, one would ideally need to observe the real (latent) QoL of the persons described in the vignettes, which is impossible. However, there are a couple of tests that one can perform to show evidence that vignette equivalence holds. Following Murray *et al.* (2003), one could for instance look at whether respondents have ranked the vignettes "correctly". Our vignettes were designed in such a way that one can objectively rank them as they share the same factors but in different intensities. Table 5 reports the percentages of "correct" vignette evaluation ranking by all the different subgroups in our sample. Specifically, the first line of the table indicates that more than 98% of our sample ranked the vignettes "correctly" in the sense that $y_{i1} \geqslant y_{i2} \geqslant y_{i3}$ for 98% of our sample. This percentage is very high and remains above 95% for all the different subgroups considered in our analysis. To assess whether there is any systematic variation in the vignette rankings, we follow Bago d'Uva *et al.* (2011)

and regress on our set of control variables a dummy variable that takes the value 1 if a respondent did not rank the vignettes "correctly", i.e., if $y_{i1} \geqslant y_{i2} \geqslant y_{i3}$ does not hold, and 0 otherwise[22]. The F-statistics of joint-significance of this regression has a p-value of 0.331, indicating that there is no evidence of systematic variation in the ordering of our three vignettes, even though respondents from Switzerland and France were significantly more likely to answer "inconsistently" (p-value < 0.01) than others. This, of course, does not mean that the vignette equivalence assumption holds, as this constitutes only an indirect test. It is indeed possible that respondents sharing a common characteristic could have interpreted vignettes differently than others even though the way they have ranked them was "correct"[23].

6.2 Tests of response consistency

In addition to vignette equivalence, our model identification relies on the assumption of response consistency, under which the vignettes are instrumental in revealing heterogeneous reporting scales. This assumption may not hold when respondents use different scales to evaluate anchoring vignettes and their own conditions. For example, respondents might be more "positive" when evaluating their own conditions as compared to when they are asked to evaluate vignettes. That being said, if this "positiveness" is not related to regressors included in the models, estimation results will not suffer from any bias. Falsification tests usually rely on the availability of corresponding objective measures[24]. Unfortunately, objective measures of QoL do not readily exist because of its multi-dimensional and inherently subjective nature. Household income, education and health constitute important aspects of QoL, but these are not its only inputs.

In this study, we test the response consistency assumption by following the method suggested by van Soest *et al.* (2011) and Angelini *et al.* (2013), in which they compare the evaluations of vignettes to self-evaluations for respondents whose characteristics

[22]Results are available upon request.

[23]Note that our econometric models allow for cases where the ranking of the vignettes is "incorrect" due to idiosyncratic errors ϵ_{ij} for $j = 1, ..., J$.

[24]For example, van Soest *et al.* (2011) tested the response consistency assumption by using objective drinking frequency when studying self-reported alcohol consumption. Their results suggest that the response consistency assumption holds in their setting.

Table 5: Ordering - Vignette equivalence

	Percentage	Observations
All Sample	98.02	1816
Male	97.75	887
Female	98.28	929
Parents' income: Less or equal to 4000	96.58	146
Parents' income: (4000,7000]	98.2	389
Parents' income: (7000,12000]	98.69	613
Parents' income: More than 12000	97.71	480
School performance: Better than average	98.18	494
School performance: Average	97.79	1084
School performance: Below average	98.74	238
Relation with parents: Good	97.84	1202
Relation with parents: Good, but not always	98.56	557
Relation with parents: Bad	96.49	57
Close friends: Zero	95.45	44
Close friends: One	98.67	75
Close friends: Two	97.21	287
Close friends: Three or more	98.23	1410
Depression: No or minimum symptoms	98.12	1384
Depression: Mild	97.59	290
Depression: Moderate	97.25	109
Depression: Severe	100	33
Perceived Obese	98.54	206
Perceived not obese	98.02	1562
Eating disorder	96.61	118
No eating disorder	98.05	1641

Note: Percentages of respondents in our sample that "respect" the ordering of the vignettes.

match the characteristics of the fictitious persons described in the anchoring vignettes. If response consistency holds, vignette evaluations and self-evaluation should in theory be close to each other. In addition to this "convergence" test, we also conduct a "divergence" test, where we examine reporting behaviors of respondents whose individual characteristics are different from the ones of the fictitious persons described in the vignettes. The hypothesis is that, given response consistency, respondents whose characteristics are very different from the ones described in the vignettes should evaluate their own conditions and the anchoring vignettes differently.

Figure 5 shows the kernel density of the self-reported QoL (solid line) as well as the evaluation of vignette 1 (lightest dashed line) for the entire sample (N = 1816, Self-evaluation: $Mean$ = 79.976, $S.D$ = 15.231. Vignette 1: $Mean$ = 91.244, $S.D$ = 9.700). On

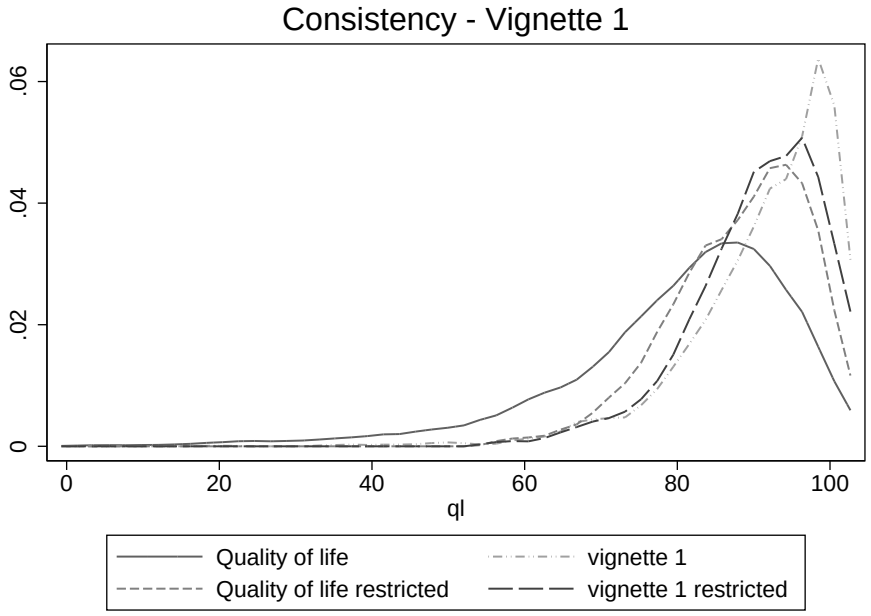

Figure 5: Kernel densities of the evaluation of vignette 1 and self-evaluation based on the entire sample (N = 1816, Self-evaluation: $Mean$ = 79.976, $S.D$ = 15.231. Vignette 1: $Mean$ = 91.244, $S.D$ = 9.700) and based on a subsample that restricts the analysis to individuals whose characteristics match the ones of the fictitious person described in vignette 1 (N = 70, Self-evaluation: $Mean$ = 88.314, S.D=8.519. Vignette 1: $Mean$ = 89.714, $S.D$ = 8.725).

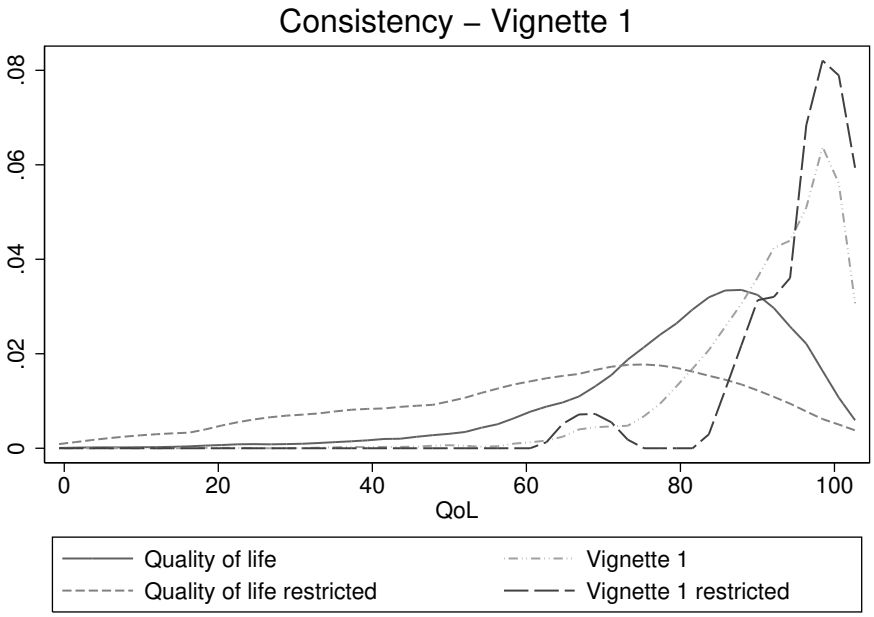

Figure 6: Kernel densities of the evaluation of vignette 1 and self-evaluation based on the entire sample (N = 1816, Self-evaluation: $Mean$ = 79.976, $S.D$ = 15.231. Vignette 1: $Mean$ = 91.244, $S.D$ = 9.700) and based on a subsample that restricts the analysis to individuals whose characteristics do *not* match the ones of the fictitious person described in vignette 1 in all aspects (N = 17, Self-evaluation: $Mean$ = 62.706, S.D=22.019. Vignette 1: $Mean$ = 95.294, $S.D$ = 8.145).

the same graph, we report the same measures but restrict our sample to respondents who closely match the description in vignette 1, i.e., individuals from households in the highest income category, who have better than average grades, who have good relationship with their parents, who have 3 or more friends, who are not obese and do not suffer from any eating disorder (N = 70, Self-evaluation: $Mean$ = 88.314, S.D=8.519. Vignette 1: $Mean$ = 89.714, $S.D$ = 8.725). As can be seen in Figure 5, the two densities of the QoL evaluations for our restricted sample are very similar and almost overlap. In fact, we cannot reject the null hypothesis of equal distribution functions (p-value of the Kolmogorov-Smirnov test = 0.527). This supports the assumption of response consistency, as it indicates that individuals, at least when their characteristics resemble those described in vignette 1, use the same reporting scale to evaluate vignette 1 and their own QoL[25].

When restricting our analysis to respondents who differ in all dimensions from the fictitious person described in vignette 1, unlike in the previous graph, the self-reports and evaluations of vignette 1 are very different to each other compared to the distributions based on the whole sample, as shown in Figure 6. The difference in the distributions for the restricted sample in Figure 6 therefore appears to be mainly coming from intrinsic difference in QoL between the respondents and the person described in the vignette (which we know exist) rather than differences in the reporting scale between self-report and vignette evaluations.

All in all, the stark contrasts in the distributions of self-reports and evaluations of vignette 1 of the restriced samples in Figures 5 and 6 suggest that the response consistency assumption seems to hold in our setting.

[25]We can proceed with the same analysis but focusing on vignette 2 and 3 instead. However, unlike vignette 1, there are few respondents in our sample who match all the characteristics of the persons that are described in vignettes 2 and 3 at the same time. We nonetheless test the response consistency assumption for vignettes 2 and 3 by matching the different dimensions described in the vignettes to respondents one by one. Figures 8 and 9 in Appendix B report the densities of the QoL measure as well as of the evaluations of vignettes 2 and 3, respectively, based on the entire sample and based on the restricted samples. Not surprisingly, evidence is not as clear as in the case of Figure 5 because we have to match characteristics one by one. However, one can see that the overlaps of the restricted densities are more pronounced than the overlaps of the densities based on the entire sample, suggesting that response consistency could hold.

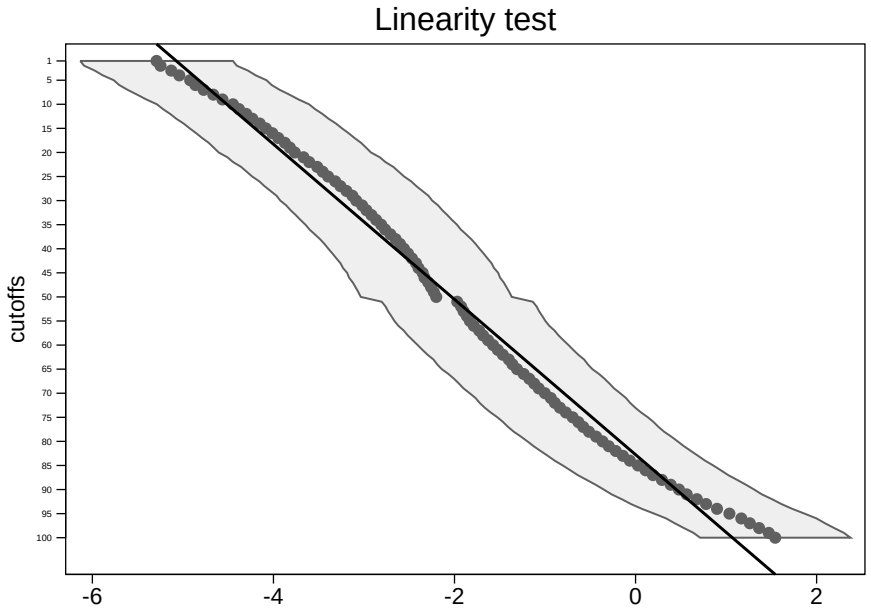

Figure 7: Cutoff points of the Generalized Ordered Response model (Model 3) along with their 95% confidence intervals.

6.3 Test of linearity of VAS

When performing linear estimations, we make the implicit assumption that the measure of our dependent variable has interval property, i.e., that the distances between points in our VAS 0-100 scale correspond to a cardinal representation of differences in QoL. Maintaining the assumption that the error terms are normally distributed, one can assess whether the cardinality assumption holds by looking at the distances between the estimated cutoffs that we derive from Model 3 in which we consider our VAS measure to be ordinal[26]. Figure 7 displays the values of the 100 cutoffs that are estimated from Model 3, along with their 95% confidence intervals.

One can see that the cutoffs are aligned and quite closely match the linear fit (black line), which means that the assumption of interval scale seems to be supported in our case. This implies that trying to fit our data using ordered response models would not be necessary in our setting as the distances between the points in our VAS 0-100 scale are meaningful and our VAS measure can be interpreted as being cardinal.

[26]When VAS measures are assumed to be ordinal, cutoff estimates in Model 3 represent the latent QoL location on a uni-dimensional interval scale. This means that one can assess the cardinality property of the VAS measure by checking whether the estimated cutoffs fall on a straight line.

6.4 Test of sex-specific vignette evaluations

The set of results above relies on the assumption that we can compare QoL across sexes and correct for reporting heterogeneity using sex-specific vignettes. This assumption means that females and males would have evaluated the vignettes in a similar way if the vignettes were describing the hypothetical scenario of someone from the opposite sex. In other words, we are assuming that the different factors described in the vignettes, along with their intensities, do not affect the evaluation of the vignettes differently for the two sexes. This is a strong assumption. Jürges and Winter (2013) shows that vignette evaluations may be sensitive to the sex of the person described in the vignettes. While their analysis is on the perceived disability of older individuals (aged 50+), it is possible that vignettes ratings are sensitive to sex in our setting as well. Although QoL could possibly be more universal and sex-neutral than disability at older ages, one concern in our application for instance could be that obesity (vignette 2) and eating disorder (vignette 3) are evaluated more negatively if these characteristics are associated with females than males. And because vignettes are sex-specific in our setting, this could potentially mean that the sex difference in QoL we obtain in our analysis could only be driven by the fact that some factors in the vignettes influence the evaluation of males and females differently rather than by reporting heterogeneity. In that case, we would not be able to attribute the sex gap in our results to reporting heterogeneity.

We show in Appendix C that this is likely not the case. In a follow-up study, we asked a different set of college students of roughly the same sample size to evaluate the three vignettes used in this study, first matching the sex of the respondents to the sex of the person described in the vignette. Later in the same online survey, we asked the same respondents to evaluate vignettes that were identical in all aspects but sex to the previous vignettes. That is, we ask respondents to evaluate the vignettes describing the situation of someone from the opposite sex. Figures 10 and 11 in Appendix C, which represent males' and females' evaluations, respectively, show that the sex of the person described in the vignettes does not affect the way males and females evaluate the vignettes. All but one vignette show that the evaluations are statistically not different

at conventional level between male and female vignettes. Indeed, as reported in the appendix, female respondents seem to evaluate more positively vignette 1 if the person described in the vignette is a female rather than a male. Again, this test is derived using another set of respondents, but these results suggest that the sex of the persons described in the vignettes is irrelevant and that the sex-gap we find in our analysis can indeed be attributed to reporting heterogeneity[27].

6.5 Robustness check for the inclusion of control variables

We have shown that females are worse off than males in terms of QoL although their self-reports indicate otherwise. However, as shown in our descriptive statistics, females and males differ in many socio-demographic aspects, one of them being the prevalence of depressive symptoms, which clearly have a strong and negative effect on one's QoL. The sex gap in self-reported QoL (Model 0) we get in Table 2 column 1 might therefore be due to spurious correlation resulting from the inclusion of certain covariates such as depressive symptoms.

To show that our results are not sensitive to the inclusion/exclusion of certain covariates, particularly depressive symptoms, we re-estimate both models 0 (naive model without vignette correction) and 1 (linear fixed-effect model with vignette correction) to assess whether our results are robust to the inclusion/exclusion of individual characteristics. We start with a specification that only controls for sex, and then gradually include more covariates into our model until we obtain the benchmark specification we have been estimating until now[28].

In Table 6, columns 1 and 5 correspond to raw differences in QoL between males and females without including any other covariates into the analysis. Under this model specification, there is no statistically significant difference between males and females QoL

[27]It is possible that perceived need of consistency among respondents is driving this result. However, the significant time gap between the vignette evaluations and the fact that our VAS scales do not contain any labels or ticks (except at both ends of our VAS scales) that could be used as referenced points allow to rule this possibility out.

[28]This also allows to indirectly assess to what extent sex-specific sample selectivity affects our results. Indeed, under the assumption that sex difference in sample self-selection is associated with observable characteristics, adding these characteristics in our model would reduce the selection bias.

Table 6: Robustness of our results to inclusion of control variables

	Without vignette correction (Model 0)				With vignette correction (Model 1)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	−0.863	−1.094	−0.426	1.426**	−3.853***	−3.983***	−3.107***	−1.584**
	(0.714)	(0.709)	(0.699)	(0.666)	(0.791)	(0.790)	(0.761)	(0.736)
Age		−0.523***	−0.639***	−0.479**		−0.722***	−0.771***	−0.632***
		(0.187)	(0.227)	(0.208)		(0.214)	(0.248)	(0.234)
Single		−2.371***	−1.840***	−1.511**		−1.832**	−1.336*	−1.001
		(0.718)	(0.692)	(0.651)		(0.806)	(0.768)	(0.733)
Family income:								
- (4000 7000]			3.060*	3.022**			3.396*	3.386**
			(1.627)	(1.500)			(1.757)	(1.678)
- (7000 12000]			8.854***	7.087***			9.987***	8.390***
			(1.505)	(1.417)			(1.620)	(1.574)
- More than 12000			12.121***	9.793***			15.115***	13.016***
			(1.513)	(1.441)			(1.668)	(1.639)
Perceived obese: No				2.520**				2.247*
				(1.148)				(1.216)
Eating disorder: No				2.253				0.128
				(1.573)				(1.688)
Depression:								
- Mild				−4.951***				−3.855***
				(0.977)				(1.080)
- Moderate				−9.883***				−7.916***
				(1.881)				(2.024)
- Severe				−23.021***				−21.547***
				(4.285)				(4.369)
School performance								
- Average				−1.299*				−1.608**
				(0.691)				(0.795)
- Below average				−2.392**				−3.359***
				(1.197)				(1.292)
Relation with parents:								
- Good but not always				−3.735***				−3.515***
				(0.727)				(0.782)
- Bad				−13.602***				−13.829***
				(2.894)				(3.100)
Constant	80.417***	92.727***	88.350***	81.864***	81.947***	98.106***	91.424***	88.432***
	(0.503)	(4.009)	(5.432)	(5.985)	(0.493)	(4.608)	(5.880)	(6.433)

Note: Cluster robust standard errors at the individual level are reported in parenthesis * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. These results are derived using the econometric specification of Model 0 and 1.

when reporting heterogeneity is not taking into account (Column 1). When accounting for reporting heterogeneity however, there is a large and statistically significant difference between males and females (Column 5). One can see that the gender gaps between the models in which we do not control for reporting heterogeneity and the ones in which we do remain rather constant as we include more and more covariates in the specification. Indeed, the difference in the coefficients associated to *Female* between models with and without vignette correction, in other words gender-specific reporting heterogeneity, ranges from 2.681 to 3.010[29] and is therefore robust to the inclusion of covariates.

[29]The range is straightforwardly calculated as:

$$range = \begin{pmatrix} min\{-3.853-(-0.863), -3.983-(-1.094), -3.107-(-0.426), -1.584-1.426\} \\ max\{-3.853-(-0.863), -3.983-(-1.094), -3.107-(-0.426), -1.584-1.426\} \end{pmatrix} = \begin{pmatrix} -3.010 \\ -2.681 \end{pmatrix} \quad (30)$$

6.6 Robustness check for the discretization of VAS in the HOPIT model

In our main results, we discussed several models that account for reporting heterogeneity related to VAS measures, including HOPIT models in which we partitioned our VAS using 20, 40, 60 and 80 as cutoff points. As robustness checks, we estimate Model 5 by partitioning our VAS measure in quintiles and on natural groupings using partition clustering methods (five groups) (Jain, 2010). Table 7 column 1 reports the results we obtained in our benchmark Model 5 when discretizing our VAS in 20 points interval (same as Table 3). Column 2 reports the results of Model 5 when partitioning our VAS by quintiles and column 3 reports the results with clustered discretized responses. As can clearly be seen from these results, irrespective of the way we partition our VAS measure in our HOPIT specification Model 5, the negative effect of being a female on QoL, free from reporting heterogeneity, remains negative and statistically significant[30].

7 Discussion and Conclusion

In this study, we propose several new methods to account for reporting heterogeneity in self-reported data coming from Visual Analogue Scales (VAS) using corresponding VAS-based anchoring vignettes.

Using VAS-based anchoring vignettes and standard vignettes assumptions (vignette equivalence and response consistency), we show how standard fixed-effect approaches and double-index models can be used to address individual-specific reporting heterogeneity in VAS. The advantage of running fixed-effect models is that they allow to control for both observed and unobserved heterogeneity that are correlated with the outcome variable of interest. Moreover, linear fixed-effect estimators are readily available in most software packages, which allows to assess (and address) reporting heterogeneity in data containing VAS and corresponding VAS-based anchoring vignettes in a straightforward way without

[30]Note that we use the same scaling factor σ across HOPIT models to make the coefficients comparable. The normalization of location is not important as it only affects the estimates of the constant terms.

Table 7: Robustness of our results to different VAS discretizations

	HOPIT (Model 5)		
	(1)	(2)	(3)
Female	−2.259**	−3.237***	−1.897**
	(0.918)	(0.908)	(0.878)
Age	−0.715**	−0.779***	−0.693**
	(0.282)	(0.282)	(0.273)
Single	−1.240	−1.561*	−1.322
	(0.896)	(0.885)	(0.860)
Family income:			
- (4000 7000]	2.260	2.248	2.445
	(1.703)	(1.827)	(1.677)
- (7000 12000]	6.503***	7.898***	7.557***
	(1.680)	(1.774)	(1.650)
- More than 12000	12.736***	14.564***	13.385***
	(1.802)	(1.860)	(1.753)
Perceived obese: No	2.297*	2.476*	1.991
	(1.339)	(1.412)	(1.313)
Eating disorder: No	−0.652	−1.169	−0.697
	(1.715)	(1.817)	(1.684)
Depression:			
- Mild	−4.395***	−2.863**	−4.190***
	(1.177)	(1.223)	(1.152)
- Moderate	−8.039***	−5.754***	−7.010***
	(1.769)	(1.973)	(1.765)
- Severe	−14.469***	−8.641**	−15.426***
	(2.988)	(3.757)	(3.115)
School performance			
- Average	−0.972	−2.107**	−1.678*
	(1.043)	(1.013)	(0.996)
- Below average	−2.801*	−2.663*	−2.844**
	(1.472)	(1.482)	(1.422)
Relation with parents:			
- Good but not always	−3.578***	−4.800***	−4.817***
	(0.961)	(0.973)	(0.932)
- Bad	−11.681***	−11.129***	−12.131***
	(2.329)	(2.694)	(2.347)
Constant	−1.200	−1.563	−2.208
	(7.641)	(7.765)	(7.388)

Note: Standard errors are reported in parenthesis * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. In all regressions, we also control for education level (Bachelor, Master, others), number of siblings, number of friends and origin (coefficients not reported but available upon request). Column 1 corresponds to the estimation of Model 5 in which we partitioned our VAS using 20, 40, 60 and 80 as cutoff points. Column 2 reports the results of Model 5 when partitioning our VAS by quintiles and column 3 reports the results where we dicretize our VAS measure based on natural groupings using partition clustering methods (k-means clustering) (Jain, 2010).

a lot of programming work. The double-index models, on the other hand, allow to acquire a better understanding on the source of the reporting heterogeneity, along with its magnitude, because it explicitly defines reporting heterogeneity as a function of individual characteristics.

We also show that several other methods such as Generalized Ordered Response and Hierarchical Ordered Probit (HOPIT) models can be used to meaningfully adjust for potential reporting heterogeneity under the weaker assumption that VAS responses should be interpreted as ordered rather than cardinal data. Compared to Generalized Ordered

Response models that assume parallel shift, the linear and exponential forms of the hierarchical ordered probit model (HOPIT) have the advantage of capturing non-linear effects of reporting heterogeneity.

We then apply our methods to real data assessing gender differences in QoL among students in Switzerland. Using online survey responses, our application shows that under reporting homogeneity, females report having higher QoL than males, at about 1.4 points on a 0-100 VAS scale, which is small but statistically significant. This is in line with previous studies in epidemiology describing the determinants of QoL. We also show that female students tend to rate the QoL of corresponding comparable anchoring vignettes higher than male students. Accounting for these gender differences in response behaviors, we show that female students actually appear to be worse off in terms of QoL than male students, with a QoL that is 1.6 points lower than males. Had other aspects of life disparity been considered, this sex gap would have added up to 3.8 points, as indicated by our counterfactual distribution. These results are robust across all the model specifications we consider. We also show evidence that the usual underlying assumptions that are needed for the proper use of vignettes, i.e., response consistency and vignettes equivalence, seem to hold.

All in all, we believe that reporting heterogeneity could be one of the explanations for why females report having higher QoL than males, despite being disadvantaged in several domains such as education, income and health. This holds true in our empirical application and we speculate that this result generalizes to other populations as well. Using the Gallup World Pull and Likert scale measures of life satisfaction, Montgomery (2016) finds that, on average worldwide, women report higher life satisfaction than men. However, she also finds that women and men systematically use different response scales and that, once differences on response scales across sex have been accounted for, women are less happy than men on average. The results of our application therefore constitute further evidence that reporting heterogeneity may be important in assessing gender differences in QoL and that the commonly found female advantage in QoL assessments may at least be partially due to differences in reporting behaviors.

Appendix A - Vignettes used in this study

Vignette 1

[Firstname1]'s monthly family income is 20,158 CHF. S/He has an above-average school performance compared to her/his classmates, and s/he enjoys school a lot. S/He is very close with her/his parents. S/He has four close friends. Currently, s/he has no health problems.

Vignette 2

[Firstname2]'s monthly family income is 10,079 CHF. S/He has an average school performance compared to her/his classmates, and s/he finds school rather interesting. S/He argues sometimes with her/his parents but has otherwise a satisfactory relationship with them. S/He has two close friends. Currently, s/he has no particular health problem but thinks s/he is rather obese.

Vignette 3

[Firstname3]'s monthly family income is 5,039 CHF. S/He has a below-average school performance compared to her/his classmates, and s/he does not like school at all. S/He is very distant from her/his parents. S/He has no close friend. Recently, s/he has been diagnosed with having eating disorder.

After reading each vignette, the respondents were asked to answer the following question: *How do you feel about the life of [Firstname'i'] as a whole? With* $i = 1, 2, 3$

Appendix B - Test on response consistency

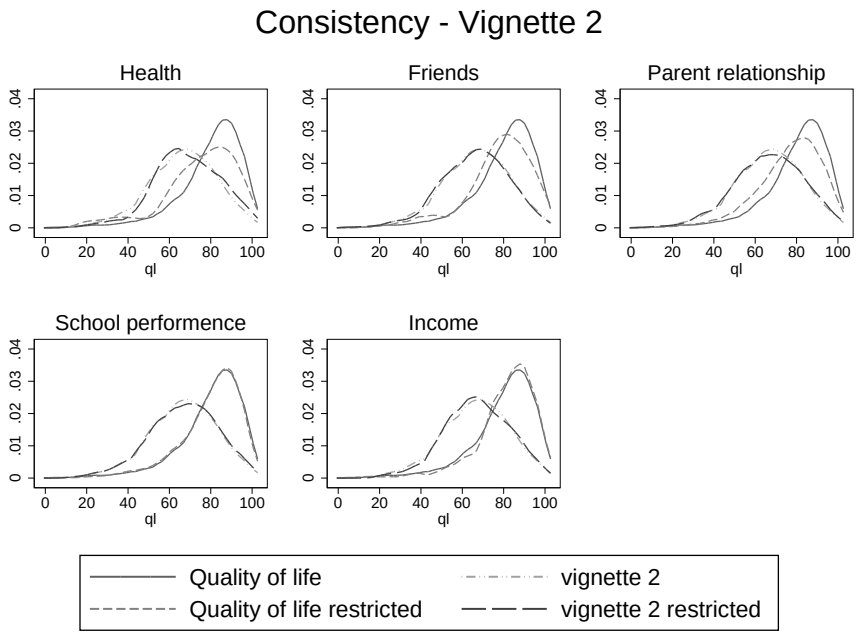

Figure 8: Kernel densities of the evaluation of vignette 2 and self-evaluation of the entire sample and of a subsample that restricts the evaluation to individuals who match the description in vignette 2 (by characteristic).

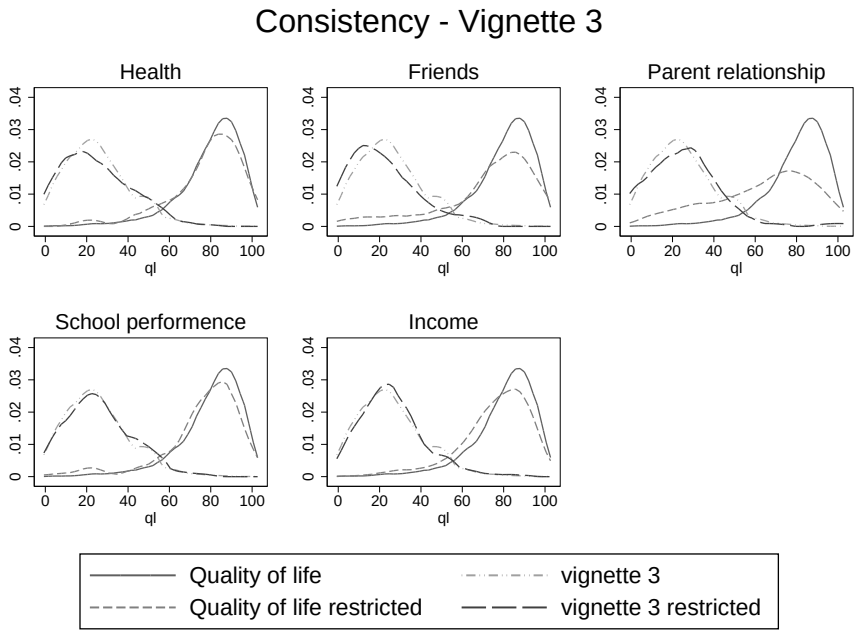

Figure 9: Kernel densities of the evaluation of vignette 3 and self-evaluation of the entire sample and of a subsample that restricts the evaluation to individuals who match the description in vignette 3 (by characteristic).

Appendix C - Vignette evaluations across sex

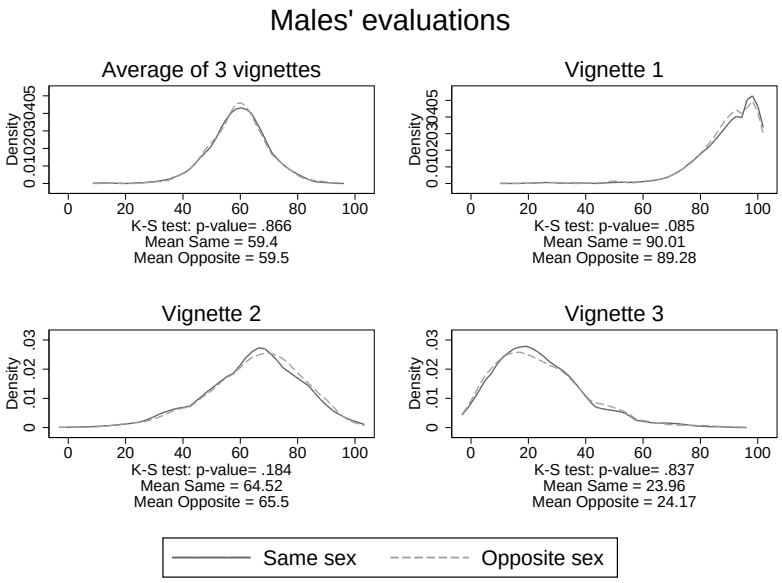

Figure 10: Males' evaluations of the vignettes used in the study using a different set of respondents. Respondents were asked to evaluate the same vignettes twice, identical in all dimensions but sex, once in which the vignette was about a male and once in which the vignette was about a female. Note that there was an interval with several questions in between the two evaluations.

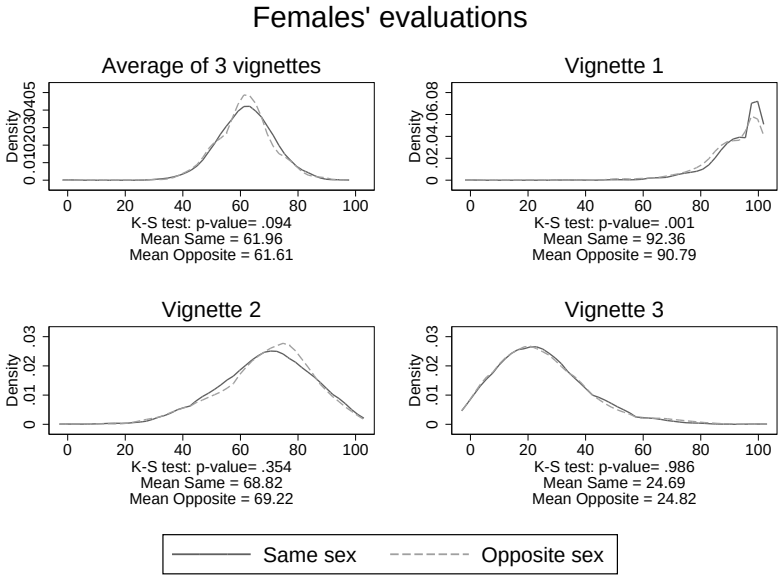

Figure 11: Females' evaluations of the vignettes used in the study using a different set of respondents. Respondents were asked to evaluate the same vignettes twice, identical in all dimensions but sex, once in which the vignette was about a male and once in which the vignette was about a female. Note that there was an interval with several questions in between the two evaluations.

References

Abdel-Fattah, M., Ramsay, I., and Barrington, J. W. (2007). A simple visual analogue scale to assess the quality of life in women with urinary incontinence. *Eur. J. Obstet. Gynecol. Reprod. Biol.*, **133**(1), 86–89.

Abend, R., Dan, O., Maoz, K., Raz, S., and Bar-Haim, Y. (2014). Reliability, validity and sensitivity of a computerized visual analog scale measuring state anxiety. *J. Behav. Ther. Exp. Psychiatry*, **45**(4), 447–453.

Angelini, V., Viola, A., Danilo, C., Luca, C., and Omar, P. (2013). Do danes and italians rate life satisfaction in the same way? using vignettes to correct for Individual-Specific scale biases. *Oxf. Bull. Econ. Stat.*, **76**(5), 643–666.

Au, N. and Lorgelly, P. K. (2014). Anchoring vignettes for health comparisons: an analysis of response consistency. *Quality of Life Research*, **23**(6), 1721–1731.

Bago d'Uva, T., d'Uva, T. B., Van Doorslaer, E., Lindeboom, M., and O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Econ.*, **17**(3), 351–375.

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

Bailey, B., Gravel, J., and Daoust, R. (2012). Reliability of the visual analog scale in children with acute pain in the emergency department. *Pain*, **153**(4), 839–842.

Bijur, P. E., Wendy, S., and John Gallagher, E. (2001). Reliability of the visual analog scale for measurement of acute pain. *Acad. Emerg. Med.*, **8**(12), 1153–1157.

César, K. G., Brucki, S. M. D., Takada, L. T., Nascimento, L. F. C., Gomes, C. M. S., Almeida, M. C. S., Oliveira, M. O., Porto, F. H. G., Senaha, M. L. H., Bahia, V. S., Silva, T. B. L., Ianof, J. N., Lívia, S., Schmidt, M. T., Jorge, M. S., Vale, P. H. F., Cecchini, M. A., Luciana, C., Soares, R. T., Gonçalves, M. R., Jerusa, S., Porto, C. S., Carthery-Goulart, M. T., Yassuda, M. S., Mansur, L. L., and Ricardo, N. (2014). Performance of the visual analogue scale of happiness and of the cornell scale for depression in dementia in the tremembé epidemiological study, brazil. *Dementia & Neuropsychologia*, **8**(4), 389–393.

Craig, B. M., Busschbach, J. J., and Salomon, J. A. (2009). Keep it simple: ranking health states yields values similar to cardinal measurement approaches. *Journal of Clinical Epidemiology*, **62**(3), 296–305.

Crane, M., Rissel, C., Greaves, S., and Gebel, K. (2016). Correcting bias in self-rated quality of life: an application of anchoring vignettes and ordinal regression models to better understand QoL differences across commuting modes. *Qual. Life Res.*, **25**(2), 257–266.

Deaton, A. (2018). What do self-reports of wellbeing say about life-cycle theory and policy? *Journal of Public Economics*.

Devesa, J. M., Vicente, R., and Abraira, V. (2012). Visual analogue scales for grading faecal incontinence and quality of life: their relationship with the Jorge–Wexner score and rockwood scale. *Tech. Coloproctol.*, **17**(1), 67–71.

d'Uva, T. B., O'Donnell, O., and van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older europeans. *Int. J. Epidemiol.*, **37**(6), 1375–1383.

Grol-Prokopczyk, H., Freese, J., and Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, **52**(2), 246–261. PMID: 21673148.

Hanandita, W. and Tampubolon, G. (2016). Does reporting behaviour bias the measurement of social inequalities in self-rated health in indonesia? an anchoring vignette analysis. *Qual. Life Res.*, **25**(5), 1137–1149.

Helliwell, J., Layard, R., and Sachs, J. (2017). World happiness report 2017. *New York: Sustainable Development Solutions Network*.

Hopkins, D. J. and King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opin. Q.*, **74**(2), 201–222.

Iburg, K. M., Salomon, J. A., Tandon, A., and Murray, C. J. (2001). Cross-population comparability of self-reported and physician-assessed mobility levels: evidence from the third national health and nutrition examination survey. *Global Programme on Evidence for Health Policy Series*, (14).

Ierza, J. V. (1985). Ordinal probit: a generalization. *Communications in Statistics-Theory and Methods*, **14**(1), 1–11.

Ismail, A. K., Abdul Ghafar, M. A., Shamsuddin, N. S. A., Roslan, N. A., Kaharuddin, H., and Nik Muhamad, N. A. (2015). The assessment of acute pain in Pre-Hospital care using verbal numerical rating and visual analogue scales. *J. Emerg. Med.*, **49**(3), 287–293.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, **31**(8), 651–666.

Jürges, H. (2006). True health vs response styles: exploring cross-country differences in self-reported health. *Health Econ.*, **16**(2), 163–178.

Jürges, H. and Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics*, **22**(1), 1–13.

Kapteyn, A., Smith, J. P., and van Soest, A. (2007). Vignettes and Self-Reports of work disability in the united states and the netherlands. *Am. Econ. Rev.*, **97**(1), 461–473.

Kapteyn, A., Smith, J. P., and van Soest, A. (2013). Are americans really less happy with their incomes? *Rev. Income Wealth*, **59**(1), 44–65.

Kelly, A.-M. and Anne-Maree, K. (1998). Does the clinically significant difference in visual analog scale pain scores vary with gender, age, or cause of pain? *Acad. Emerg. Med.*, **5**(11), 1086–1090.

King, G., Murray, C. J. L., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and Cross-Cultural comparability of measurement in survey research. *Am. Polit. Sci. Rev.*, **98**(01), 191–207.

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

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, **16**(9), 606–613.

Lesage, F. X., Berjot, S., and Deschamps, F. (2012). Clinical stress assessment using a visual analogue scale. *Occup. Med.*, **62**(8), 600–605.

Lindeboom, M. and van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health. *J. Health Econ.*, **23**(6), 1083–1099.

Malhotra, C. and Do, Y. K. (2013). Socio-economic disparities in health system responsiveness in india. *Health Policy Plan.*, **28**(2), 197–205.

Molina, T. (2016). Reporting heterogeneity and health disparities across gender and education levels: Evidence from four countries. *Demography*, **53**(2), 295–323.

Montgomery, M. (2016). Reversing the gender gap in happiness: Validating the use of life satisfaction self-reports worldwide job market paper.

Mu, R. (2014). Regional disparities in self-reported health: evidence from chinese older adults. *Health Econ.*, **23**(5), 529–549.

Murray, C., Ozaltin, E., Tandon, A., and Salomon, J. (2003). Empirical evaluation of the anchoring vignettes approach in health surveys. *Health Systems Performance Assessment: Debates, Methods and Empiricism*.

Obayashi, Y., Yoko, O., Shigemi, I., and Asuka, S. (2016). The validity and reliability of a scale on postnatal posttraumatic stress symptoms related to childbirth among japanese women: Evaluation of the japanese-language version of the impact of event scale-revised. *Journal of Women's Health Care*, **5**(3).

Peracchi, F. and 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.

Pudney, S. and Shields, M. (2000). Gender, race, pay and promotion in the british nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics*, **15**(4), 367–399.

Ravallion, M., Himelein, K., and Beegle, K. (2016). Can subjective questions on economic welfare be trusted? *Econ. Dev. Cult. Change*, **64**(4), 697–726.

Rice, N., Robone, S., and Smith, P. C. (2012). Vignettes and health systems responsiveness in cross-country comparative analyses. *J. R. Stat. Soc. Ser. A Stat. Soc.*, **175**(2), 337–369.

Sakamoto, R., Okumiya, K., Norboo, T., Tsering, N., Wada, T., Fujisawa, M., Imai, H., Nose, M., Ishimoto, Y., Kimura, Y., Fukutomi, E., Chen, W., and Matsubayashi, K. (2016). Health and happiness among community-dwelling older adults in domkhar valley, ladakh, india. *Geriatr. Gerontol. Int.*

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

Studer, R. (2011). Does it matter how happiness is measured? evidence from a randomized controlled experiment. *Working Paper Series, Working Paper No. 49*.

Torrance, G. W., Feeny, D., and Furlong, W. (2001). Visual analog scales: do they have a role in the measurement of preferences for health states?

van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *J. R. Stat. Soc. Ser. A Stat. Soc.*, **174**(3), 575–595.

Vust, S. and Michaud, P.-A. (2008). Médecine de l'adolescence: Troubles des conduites alimentaires atypiques. *Revue médicale suisse*, **4**(139), 40–43.

Wang, J. (2016). Rural-to-urban migration and rising evaluation standards for subjective social status in contemporary china. *Soc. Indic. Res.*

Zampelis, V., Ornstein, E., Franzén, H., and Atroshi, I. (2014). A simple visual analog scale for pain is as responsive as the WOMAC, the SF-36, and the EQ-5D in measuring outcomes of revision hip arthroplasty. *Acta Orthop.*, **85**(2), 128–132.