A NEW TOOL FOR MEASURING CUSTOMER SATISFACTION: THE ANCHORING VIGNETTE APPROACH

Omar Paccagnella[1]
Department of Statistical Sciences, University of Padua, Padua, Italy

Received: 28th February 2013/ Accepted: 4th December 2013

Abstract. *The approach of the "anchoring vignettes" has been recently introduced as a new and powerful way to detect the systematic differences in the use of the response scales within countries or socio-economic groups when respondents evaluate themselves. This approach allows to analyse self-reported ordinal survey responses taking into account individual differences in the interpretation of the questions. This paper aims at showing how the marketing literature can be enriched by the application of this new methodology for the measurement of not-observable constructs, such as the customer satisfaction. Exploiting data collected by the LISS panel, this approach is applied for measuring the satisfaction after the purchase of a smartphone.*

Keywords: *Anchoring vignettes, Customer satisfaction, Chopit model, Ordered probit model, Response scales.*

1. INTRODUCTION

In cross-cultural studies, that is the comparisons across countries or socio-economic groups, a key issue is to determine whether scales are actually measuring the same concepts in all countries or groups. The educational testing literature refers to Differential Item Functioning (DIF) when "individuals with the same ability/skills have a different probability to provide a certain answer to a questionnaire/test" (Holland and Wainer, 1993), while the psychological literature defines the Response Style (RS) as "a tendency to respond to questionnaire items independently of item content" (Paulhus, 1991). Regardless of the literature definition, the common evidence is that respondents of different countries or groups often understand the same survey question differently. The presence of individual heterogeneity, that is the inter-personal and inter-cultural differences

[1] Omar Paccagnella, email: omar.paccagnella@unipd.it

in interpreting, understanding or using response categories for the same question, may systematically bias the measurement of the content variables, producing misleading assessments of relative performances in cross-cultural comparisons.

These problems are common also in many fields of the marketing researches, such as customer satisfaction, brand loyalty or product involvement.

Nowadays, customer satisfaction is one of the most important measures in marketing, even though it has a multi-dimensional nature and is not directly observable. Its measurement is mainly based on the discrepancy paradigm (Oliver, 1993), that is an evaluation obtained comparing the perceived performance of the product (or service) after purchase (or use) and the pre-purchase expectations. The instruments that are usually adopted are measurement scales (Kao et al., 2007). Furthermore, there is a branch of the literature that refers to the work of Fornell et al. (1996), who explicitly recognise the non-observable nature of the customer satisfaction and analyse the collected data by means of structural equation models with latent variables, introducing the so-called American Customer Satisfaction Index (ACSI) index.

Recently, marketing research has shown an increasing interest in individual response scales in cross-country comparisons (Baumgartner and Steenkamp, 2001). The main requirement is that the measurement instruments are cross-nationally invariant. The multigroup Confirmatory Factor Analysis (CFA) is usually adopted for testing measurement invariance across countries. See Vandenberg and Lance (2000) for a review. Lack of the invariance of measurement instruments means that cross-national comparisons cannot be performed. To overcome the limitations of the multigroup CFA, new statistical solutions have been recently introduced. De Jong et al. (2007) propose a hierarchical IRT (Item Response Theory) model. De Jong et al. (2008) introduce an IRT-based method to measure extreme response styles. Van Rosmalen et al. (2010) estimate a latent class-bilinear multinomial logit, ignoring the natural ordering of the rating scale categories.

This paper aims at enriching the marketing literature proposing the definition and the application of a new methodology to measure customer satisfaction. This methodology is based on the *anchoring vignettes*, an original and innovative type of questionnaire introduced by King et al. (2004), and generalised by King and Wand (2007), in the field of the social sciences. In this context, it has become quite common to collect and analyse self-evaluations about the own health status, living conditions or satisfaction with respect to some domains of the own life or society. Such self-reported data strongly show heterogeneity in reporting styles.

The anchoring vignettes can help to identify the systematic differences in the use of response scales within countries, groups or market segments. The econometric specification of such model is known as *chopit* (Compound Hierarchical Ordinal Probit) model and it is basically characterised by a joint estimation of some ordinal probit models.

The approach of the anchoring vignettes is conceptually different with respect to the aforementioned literature: in the current marketing literature cross-country comparisons could be performed including a few country-specific and respondent-specific parameters in the model equation, while in the vignette approach DIF is modelled through variations in the thresholds. The thresholds determine the individual response scales and are allowed to vary with individual characteristics and across countries or groups.

Even though a first attempt to exploit the vignette approach to assess consumer ratings of health plans within the CAHPS (Consumer Assessment of Healthcare Providers and Systems) programme was done by Gallagher (2009), to the best of our knowledge this paper introduces the first structural specification of a vignette questionnaire specifically designed to investigate reporting heterogeneity in marketing. The VECS (Vignette Evaluation of Customer Satisfaction) project is a short questionnaire based on anchoring vignettes and developed for the measurement of the consumer satisfaction after the purchase of some specific e-communication systems. This survey was included as a module in the LISS (Longitudinal Internet Studies for the Social sciences) panel, which is a new online panel of about 5,000 Dutch households carried out by CentERdata.

This paper is organised as follows. Section 2 introduces the anchoring vignette methodology, as well as its statistical (parametric) solution. Section 3 describes the questionnaire and the data analysed in this paper. In Section 4 the empirical application of the *chopit* model and a comparison with the results from a standard ordered probit modelling are reported. Section 5 ends the paper, summarising the main conclusions and some caveats in the use of the anchoring vignettes.

2. THE ANCHORING VIGNETTES

Self-assessments are usually mislead by respondents because they provide the actual values on the concept of interest plus DIF. Anchoring vignettes are additional questions to answer by these respondents, to be used precisely for achieving a

DIF-free measurement of this concept of interest. The novelty of this approach is to compare the individual self-evaluations to a gold standard, which is the *anchor* provided by the answers to the vignettes' questions.

Each vignette depicts a hypothetical scenario, where a (hypothetical) person is described in particular situations or conditions. Respondents are asked to evaluate this individual, using the same scale adopted for self-evaluations. These additional answers are exploited for estimating each respondent DIF and correcting for it, providing the anchor to adjust the individual subjective evaluations in the domain of interest.

As a consequence, self-assessments can be compared across countries or socio-economic groups, because all subjective evaluations are now reported to a common and DIF-free scale. DIF is modelled through individual variations in the thresholds, which in turn determine the individual response scales.

There are several and remarkable applications of this approach to various domains and all findings support the ability of this approach to correct for DIF: from political efficacy and visual activity (King et al., 2004) to work disability (Kapteyn et al., 2007), from self-reported health like mobility, cognition, pain (Bago d'Uva et al., 2008) to job satisfaction (Kristensen and Johansson, 2008) or life satisfaction (Angelini et al., 2013).

Comparing data coming from different countries or socio-economic groups, the *chopit* model estimates allow to carry out counterfactual simulations: researchers may define a benchmark (e.g. the scale of a particular country) and compute adjusted distributions of the observed variable based on the benchmark scale instead of the respondent's own scale. These adjusted distributions may be compared across countries or socio-economic groups since they are now reported on a common scale (that is, as if all individuals used the response scales of the benchmark respondents).

2.1. ASSUMPTIONS

Two key measurement assumptions are requested for the validity of the vignette approach to identify reporting heterogeneity: *response consistency* and *vignette equivalence*.

The *response consistency* assumption states that each respondent uses the response categories for a question in the same way when providing a self-evaluation as well as the evaluation of the individual described in the vignettes. This means that each respondent has approximately the same DIF in his/her use of the survey response categories across the two types of questions.

The *vignette equivalence* assumption states that all respondents perceive in the same way and on the same one-dimensional scale the "true" level of the variable represented in any vignette.

Testing the validity of these assumptions is an open issue in the literature, because no formal tests are available so far. Moreover, all proposed solutions rely on some (strong or weak) additional restrictions and this means that possible rejections of the vignette assumptions' validity could be due by the failure of the auxiliary restrictions rather than the failure of the response consistency or vignette equivalence hypotheses.

For investigating the validity of response consistency, most of the contributions rely on the solution suggested by van Soest et al. (2011) of exploiting the availability of an objective measure of the construct of interest. However, given the multidimensional nature of the domains of application of the vignettes, defining an objective measure at the individual level is arguable. According to this approach, the evidence is in favour (van Soest et al., 2011), against (Datta Gupta et al., 2010) or mixed (Bago d'Uva et al., 2011) the validity of such assumption. Mixed evidence arises also in Kapteyn et al. (2011) and van Soest and Voňková (2014). In the former contribution, an interesting experiment is conducted: respondents are first asked to describe and rate their health and, in a subsequent interview, to evaluate some vignettes that are in fact descriptions of their health. In the latter, the authors construct specification tests comparing non-parametric rankings, that come directly from the raw data, with rankings implied by parametric solutions. Angelini et al. (2013) support the response consistency analysing the correlation between the self-report and the vignette evaluations for those respondents having characteristics increasingly similar to those of the person described in the vignette.

The validity of the vignette equivalence assumption has created the most intense debate in this branch of the literature. The main criticism to the vignette approach claims that people in different countries with different institutional settings might not perceive in the same way the scenario depicted by the vignettes. As an example, the same socio-economic situation might be considered less problematic in countries with a more developed welfare state than in countries with less generous states. For this reason, King et al. (2004, p.199) underline that "we still need to be careful of question wording, question order, accurate translation of different items, sampling design, interview length, social background of the interviewer and the respondent", because missing cultural differences across subsets of respondents might threaten this approach (Bago d'Uva et al., 2011). A test based on the global ordering of vignettes may be used as prima facie supporting the vignette equivalence assumption (Angelini et al., 2013; Rice et al., 2011). Peracchi and Rossetti (2013) provide a joint test of the overidentifying restrictions that are implied by the response consistency and vignette equivalence assumptions. In their applications, the vignette assumptions are rejected, even though the overidentifying restrictions are less likely to be rejected using only one among the available vignettes or performing these tests separately by subgroups of respondents. However, since no constraints on the threshold parameters are allowed[2], their model specification is a bit different with respect to the standard *chopit* model specification (see next section).

2.2. THE STATISTICAL MODEL

King et al. (2004) introduce both a non-parametric and a parametric solution to exploit the information collected through the vignettes. The treatment of the non-parametric approach is out of the scopes of this work. However, some interesting applications and evaluations of the non-parametric solutions are provided by Wand (2013) and van Soest and Voňková (2014).

The parametric approach dealing with vignette data is called *chopit* (Compound Hierarchical Ordinal Probit) model[3]. It basically consists of a joint modelling of self-assessed and vignette answers by means of an ordered probit modelling approach.

Let Y_i^* be the (unobserved) *perceived* own level of the concept of interest for respondent i ($i = 1, \ldots, n$). We assume it is the result of a linear specification

$$\begin{aligned} Y_i^* &= X_i\beta + \varepsilon_i \\ \varepsilon_i &\sim N(0,1) \end{aligned} \tag{1}$$

where X_i are exogenous variables, β is the vector of coefficients to be estimated (without constant for identification) and ε_i is an independent and identically distributed error term.

[2] This assumption implies that the monotonicity of the thresholds cannot be ensured.

[3] In the literature, the name hopit model is used in an equivalent way.

Respondent i turns the continuous unobserved level into a reported category Y_i (recorded as an ordered variable), by means of a model with individual-specific thresholds τ_i^k

$$Y_i = k \qquad if \qquad \tau_i^{k-1} \leq Y_i^* < \tau_i^k$$

where $-\infty = \tau_i^0 < \tau_i^1 < \ldots < \tau_i^K = \infty$. Thresholds are modelled as a function of some exogenous variable V_i (which may overlap X_i) and a vector of parameters γ:

$$\begin{array}{rcl} \tau_i^1 & = & \gamma^1 V_i \\ \tau_i^k & = & \tau_i^{k-1} + exp\left(\gamma^k V_i\right) \qquad k = 2, \cdots, K-1 \end{array} \tag{2}$$

where the exponential assumption guarantees that thresholds increase with k.

The reported level of the concept of interest is not comparable across respondents, because different respondents apply different threshold values to turn their perceived levels into a category.

Since the information provided by the self-assessments does not allow to identify the parameter vectors β and γ separately, the answers to the vignettes are exploited to overcome this problem.

Let Z_{ij}^* be the (unobserved) *perceived* level of the concept of interest described in vignette j $(j = 1, \ldots, J)$ for respondent i. According to the *vignette equivalence* assumption, the true level of the variable described in each vignette is perceived in the same way by all respondents. Hence, each vignette equation is defined as a function of a vignette-specific intercept plus an independent and identically distributed error term (independent of ε_i, X_i and V_i):

$$\begin{array}{rcl} Z_{ij}^* & = & \theta_j + u_{ij} \\ u_{ij} & \sim & N(0, \sigma_u^2)\cdot \end{array} \tag{3}$$

As before, respondent i turns the continuous unobserved level into a reported category Z_{ij}, by means of a threshold model with individual-specific thresholds τ_i^k

$$Z_{ij} = k \qquad if \qquad \tau_i^{k-1} \leq Z_{ij}^* < \tau_i^k \cdot$$

According to the *response consistency* assumption, the thresholds τ_i^k are the same as the self-assessment equation. As a consequence, self-assessed and vignette questions are asked on the same scale and this allows to identify threshold and vignette dummy parameters (up to scale and location normalisation) from the vignettes' equation alone and β parameters from the self-assessment equation alone.

In order to control for individual unobserved heterogeneity, Kapteyn et al. (2007) extend the standard version of the *chopit* model including a random individual effect in the thresholds' equation. Equation (2) is then replaced by:

$$\begin{array}{lcll} \tau_i^1 & = & \gamma^1 V_i + \eta_i & \\ \tau_i^k & = & \tau_i^{k-1} + exp\Big(\gamma^k V_i\Big) & \quad k = 2, \cdots, K-1\,. \end{array} \tag{4}$$

The unobserved heterogeneity term η_i is assumed to be independent of the other error terms and of the covariates and normally distributed with 0 mean and variance σ_η^2. The original King's et al. model specification is the special case of this extended solution when $\sigma_\eta^2 = 0$.

Other extensions of the original *chopit* model have been introduced in the scientific literature, for instance to control for sample selection bias (Paccagnella, 2011) or to investigate to what extent individual reporting styles are stable over time (Angelini et al., 2011). Kapteyn et al. (2007) suggest the specification in the vignette equation (3) of a gender dummy variable of the vignette description, arguing that respondents can perceive differently the vignettes when the hypothetical individual is female instead of male[4].

This paper employs the Kapteyn et al. (2007) version of the *chopit* model. As shown by van Soest and Voňková (2014), misspecification problems are reduced substantially just adding to the original *chopit* model a term able to capture unobserved heterogeneity in the thresholds.

Basically, the likelihood for the self-assessment and the vignette component is respectively uni- and J-variate ordered probit with varying thresholds. Exploiting the assumptions that the error terms are independent each other, the complete likelihood of the *chopit* model results from the product of the likelihood functions derived for the self-assessment and the vignette component. Parameter estimates are obtained by using the *gllamm* (Generalized Linear Latent and Mixed Models) procedure (Rabe-Hesketh et al., 2004) of the STATA software. In *gllamm*, the marginal log-likelihood is maximised by means of the STATA version of the Newton-Raphson Algorithm.

[4] Such model extension cannot be applied in our work because the gender randomisation of the hypothetical person described in each vignette was not planned in the VECS questionnaire.

The complete likelihood of the *chopit* model would not be easy to derive in the presence of deviations from the normality distribution and i.i.d. assumptions of the error components. However, van Soest and Voňková (2014) prompt the researchers to develop parametric and semiparametric extensions of the *chopit* model, in order to incorporate individual unobserved heterogeneity in a more flexible way (for instance, error terms could be heteroscedastic or follow semi non-parametric distributions).

3. DATA

Data come from the LISS panel, which is a new online panel of about 5,000 Dutch households carried out by CentERdata. This institute focuses on fundamental longitudinal research and provides a laboratory for the development and testing of new, innovative research techniques.

The LISS panel is a representative sample of Dutch individuals who participate in monthly Internet surveys. The panel is based on a probability sample of households drawn from the population register. Households that could not participate are provided with a computer and Internet connection. A longitudinal survey is fielded in the panel every year, covering a large variety of domains including work, education, income, housing, time use, political views, values and personality (Scherpenzeel and Das, 2010). More information about the LISS panel can be found at: www.lissdata.nl.

3.1. THE QUESTIONNAIRE

Our work involves the development, within the LISS panel, of a survey called VECS (Vignette Evaluation of Customer Satisfaction), specifically designed for the measurement of the consumer satisfaction after the purchase of some specific e-communication systems (laptop, smartphone or LCD TV). In particular, the first wave of this module investigates the experience of purchasing one of these goods in the six months before the interview, carried out in November 2011.

Respondents who report to have experienced at least one of the listed purchases are asked to answer to some standard questions on the overall quality of the product, the expectations on this quality, the extent to which the product meets their expectations, the presence of negative experiences after the purchase, such as a manufacturing defect, a delay in the delivery, and so on. Then, one self-evaluation on the overall satisfaction of this good is asked together with two specific vignettes (for more details on the VECS module and a first evidence on its data see Daschevici, 2012).

This paper examines the experience of purchasing a smartphone. Therefore, the self-evaluation question is "*How satisfied are you with your smartphone?*", while the vignette questions are:

Mark needs a mobile phone for his work. He went to a specialised shop and bought the last version of a smartphone because its features are suitable for his work. He had to wait for four days in order to receive his smartphone. Reading the user guide, he was able to learn its main features in a couple of days. He has never experienced any manufacturing defects.

Anne works part-time, makes various sports and has a lot of friends. She thinks a smartphone can meet her needs. In a shopping centre she immediately bought what she liked. Unfortunately, after two months she experienced a problem in the phone book. She came back to the shopping centre and the smartphone was withdrawn for the assistance. After ten days the phone was delivered to Anne and the problem was solved. She has not experienced any other manufacturing defects.

How satisfied is [Mark/Anne] with [his/her] smartphone?

The same answer categories as the self-evaluations are available in all vignette questions (1=Very Satisfied; 2=Satisfied; 3=Neither Satisfied, nor Dissatisfied; 4=Dissatisfied, 5=Very Dissatisfied).

3.2. DESCRIPTIVE STATISTICS

The final sample is composed of 243 Dutch panel members, who experienced the purchase of a smartphone between May and November 2011 and completed the VECS module. About 12% of them experienced at least one problem after the purchase of the good: a manufacturing defect; a delay in the delivery; a good with different features with respect to the purchase order; a price different with respect to the purchase order.

Respondents are equally distributed with respect to gender, 35.5 years old on average (the median is 33 years and the interquartile range is equal to 25 years) and do not live alone (73%). About 28% of them have high education. More than 62% of respondents have a paid job, while about 25% attend school. The average personal net monthly income is about 1385 Euros (the median is 1300 Euros), even if the third quantile is close to 1900 Euros. About 3/4 of the individuals are home owners.

Figure 1 summarises the answers to the self-assessed and the vignette evaluations on the smartphone satisfaction.

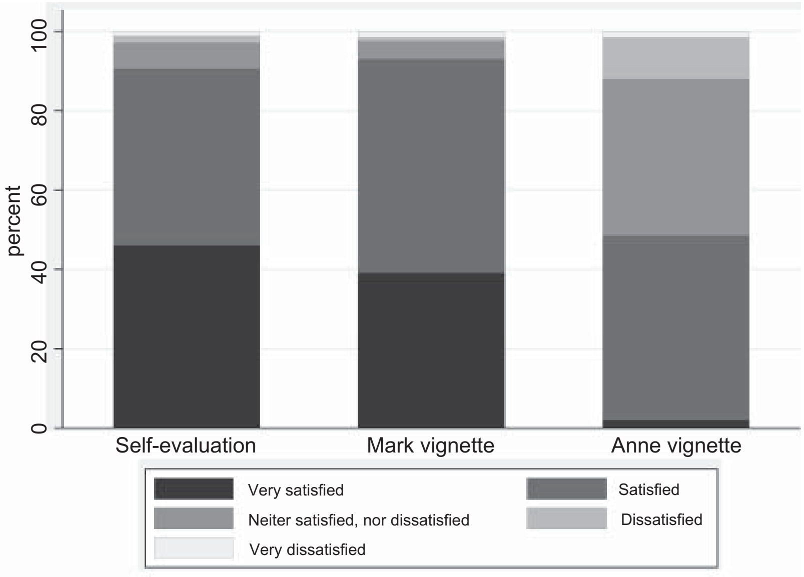

Figure 1: Self-assessed and vignette evaluations on the smartphone satisfaction

In the left panel, respondents rate their level of satisfaction after the smartphone purchase: more than 90% of them state to be satisfied or very satisfied with the product, while the percentage of dissatisfied or very dissatisfied consumers is lower than 3%. The middle and the right panels show how respondents rate the level of satisfaction of the persons depicted in the two vignettes (Mark and Anne, respectively). On average, these consumers describe Mark more satisfied than Anne: while respondents are more likely to consider Mark as either satisfied or very satisfied with his smartphone, they are much more reluctant to use the most extreme (positive) labels evaluating Anne vignette.

4. APPLYING VIGNETTES TO MEASURE CUSTOMER SATISFACTION

The Kapteyn et al. (2007) version of the *chopit* model (i.e. including individual unobserved heterogeneity in the threshold equations) is estimated on the dataset described in the previous section. Results are reported in Table 1 (columns 3 to 7) and compared with those from the estimation of a standard ordered probit model (column 2). In the presence of scale differences across socio-economic groups, the estimates of the model that does not take into account potential differences in reporting styles (i.e. not allowing for any threshold variations across individuals) will reflect both the effects of the true satisfaction with the smartphone and the effects of reporting heterogeneity.

Table 1: *Chopit* and ordered probit model: determinants of the smartphone satisfaction
Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

Variable	Ordered Probit Model	Chopit model				
		Self-assessment	Thresholds			
			τ_1	τ_2	τ_3	τ_4
Female	-0.021	0.073	0.084	-0.044	0.066	1.043
Age	0.098*	0.026	-0.085*	0.009	0.046	0.321
Household size	-0.026	-0.070	-0.068	0.089**	-0.084	-0.810**
High education	-0.022	0.015	-0.068	0.198**	-0.189	0.169
Paid job	-0.463***	-0.424**	0.168	-0.146	-0.031	0.611
Home owner	-0.453***	-0.542***	-0.157	0.166*	0.322	0.044
No problems	-0.625***	-0.914***	-0.291	0.057	0.317	-1.791**
Constant	–	–	-1.282***	0.108	-0.348	2.512*
Cutpoint 1	-1.364	–	–	–	–	–
Cutpoint 2	0.147	–	–	–	–	–
Cutpoint 3	0.807	–	–	–	–	–
Cutpoint 4	1.185	–	–	–	–	–

In the *chopit* solution, the same variables of the self-assessment equation are specified to model the thresholds τ_i^k: we control for demographic characteristics (gender and age, centered to 35 years), household size, socio-economic conditions (paid job, home ownership and education) and experience of any problems after the purchase of the smartphone.

Table 2: *Chopit* model: estimates of the vignette equation parameters
Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

Variable	Estimate
θ_1 (Mark)	-1.427***
θ_2 (Anne)	-0.102
σ_u^2	0.503
σ_η^2	0.227

Table 2 shows the estimates of the vignette equation parameters of the *chopit* model only.

The standard ordered probit is not nested into the *chopit* model. However, a formal likelihood-ratio test that compares *chopit* model against a *chopit* specification not allowing for any threshold variation across respondents (which is a model specification very close to the standard ordered probit solution) results in favour of the model with heterogeneous response scales (χ^2_{29} = 48.86, p-value = 0.012).

In three threshold equations of the *chopit* model, the effects of at least one individual variable are significantly different from 0, signalling that these thresholds depend on some individual characteristics. On the one hand, having experienced any problems after the smartphone purchase affects the respondent definition of the extreme negative ("Very Dissatisfied") reported category. On the other hand, household characteristics and socio-economic conditions like home ownership and education are important determinants for shifting the middle individual-specific thresholds (i.e. the "Satisfied" from the "Neither Satisfied, nor Dissatisfied" categories). Respondent gender and occupation do not play any role to discriminate the answer categories.

Another interesting finding is provided by the determinants of the perceived level of satisfaction. According to the *chopit* modelling, the most important variables are the experience of any problems after the good purchasing, being home owner and having a paid job: the negative sign of these point estimates reveals that respondents who had not experienced any problems after the smartphone purchase, are home owners and have a paid job are more satisfied, ceteris paribus.

The ordered probit estimates mainly differentiate from the *chopit* results because of the role played by age. According to the ordered probit modelling, the lower the age, the larger the smartphone satisfaction. On the other hand, this demographic characteristic is significant only in the first threshold equation of the *chopit* solution: younger respondents are more likely to rank themselves in the extreme positive category ("Very Satisfied") than older respondents, other things being equal. These differences between the two models might be interpreted as a signal of reporting heterogeneity effects, that a model specification not allowing for threshold variation across individuals tries to capture in some different ways with respect to the *chopit* approach.

Results from the vignette equation estimation strengthen the descriptive evidence that Dutch consumers describe Mark more satisfied than Anne, ceteris paribus.

As explained in Section 2, counterfactual simulations can be carried out once model estimates are available. However, vignette data analysed in this work were collected only in one country, where in addition the socio-economic status and standards of living are basically homogeneous across the whole population. On the other hand, from a marketing point of view it is much more interesting to evaluate the behaviour of some particular consumer profiles, in order to be able of planning suitable (one-to-one) marketing strategies.

For these reasons, based upon the model estimates from Table 2, the prediction of the reporting categories of some consumer profiles are carried out. Consequently, results from the model that does allow for correction of the DIF bias against the one without controlling for heterogeneous response scales can be directly compared (see Table 3).

All profiles reported in Table 3 have the same structure (a male, with a degree and a paid job, home owner and living with two other household members is specified) and their differences aim at showing the role played by two important individual characteristics: age and the experience of at least one problem after the purchase.

Profiles 1 and 2 identify a young individual (30 years old) and differentiate each other by the presence or not of at least a problem after the purchase. Profile 3 describes an old consumer (60 years old).

All examples highlight how DIF affects the use of the individual reporting categories. The model with heterogeneous response scales shows a lower probability of using the "Very Satisfied" category than the standard ordered probit, while the opposite applies to the "Satisfied" category. Regardless of the profile,

Table 3: *Chopit* and ordered probit models: comparison of the predicted distributions of three individual profiles
Profile 1: male, 30 years old, with a degree and a paid job, living with two other household members, home owner and without experienced any problems after the purchase
Profile 2: as Profile 1, but experienced at least a problem after the purchase
Profile 3: as Profile 1, but 60 years old

	Profile 1		Profile 2		Profile 3	
Category	Chopit	Ordered Probit	Chopit	Ordered Probit	Chopit	Ordered Probit
Very Satisfied	37.6%	62.8%	17.4%	38.2%	25.8%	51.3%
Satisfied	56.8%	33.9%	63.2%	50.5%	64.7%	42.5%
Neither Satisfied, nor Dissatisfied	4.8%	2.7%	12.1%	8.2%	8.3%	4.8%
Dissatisfied	0.5%	0.4%	7.3%	1.8%	1.2%	0.9%
Very Dissatisfied	0.3%	0.2%	0.0%	1.2%	0.1%	0.5%

the probability of being either "Satisfied" or "Very Satisfied" is always larger according to the model that does not allow for any threshold variations.

In the presence of at least a problem after the smartphone purchase (with respect to the absence of such experience), consumers reduce their positive evaluations more according to the *chopit* model than the standard ordered probit solution. Apparently, this finding seems to be at odds with the null probability of being "Very Dissatisfied" for the Profile 2 of the *chopit* model. However, it is worth noting that in the original dataset very few respondents (less than 1.5%) rate themselves as "Very Dissatisfied" with their smartphone and none of them had experienced at least a problem after the purchase.

In the end, comparing Profiles 1 and 3 the larger the age, the lower is the product satisfaction. In both cases, a reduction of the probability of being "Very satisfied" follows an increase of the probability of being "Satisfied" or "Neither Satisfied, nor Dissatisfied".

5. CONCLUSIONS

Consumers are heterogeneous individuals, whose attitudes (i.e. the satisfaction deriving from the use of specific goods or services) have a multidimensional nature, are non-observable and, consequently, difficult to measure. The existence of inter-personal and inter-cultural differences in interpreting or using response categories in some survey questions is one of the most important aspects affecting this multidimensional concept.

In order to evaluate more accurately the satisfaction of making use of goods or services by the consumers, the "anchoring vignettes" can be an innovative and powerful approach to detect the systematic differences in the use of the response scales within countries, socio-economic groups or market segments. Indeed, this approach allows to analyse self-reported ordinal responses to survey questions, taking into account individual differences in the interpretation of these questions. This work provides the first attempt of a structural definition of vignette questions specifically designed to investigate reporting heterogeneity in marketing.

Analysing the purchase of a smartphone within a period of six months among Dutch consumers, the socio-economic conditions and the experience of any problems after the purchase are the most important determinants of consumer satisfaction. These variables significantly affect both the perceived level of satisfaction and the individual response scales.

The evidence in favour of the *chopit* against the standard ordered probit model is not so strong. However, findings are encouraging to support the ability of the vignette approach to correct for DIF. First, the estimates of some parameters in the threshold equations of the *chopit* model are statistically significant. Furthermore, a formal test is in favour of the more general model that allows for correction of the DIF bias against the model not allowing for response scale variation. In the end, the comparison of the predicted probabilities for some particular consumer profiles highlights how DIF affects the individual reporting styles: according to the model that takes into account the heterogeneity in the response scales, consumers are less likely to use the extreme categories than the approach that does not take into account such heterogeneity.

Moreover, some important limiting conditions are likely to affect our results. First, these vignette data were collected only in one country (the Netherlands), where the socio-economic status and standards of living are quite homogeneous across the whole population. Then, the distribution of the self-reported satisfaction is highly skewed in this application (more than 90% of respondents declare to be satisfied or very satisfied with their smartphone). Last, since this is the first application of a vignette survey in the field of marketing, question wording and item contents need to be refined.

6. ACKNOWLEDGEMENTS

The author thanks Mariangela Guidolin, three anonymous referees and the journal editor for their helpful comments. This paper was presented at the conference "Il valore della statistica per le imprese e la società. Ricerche di opinione e di mercato" (Milan, February 2013), where it was awarded as the best paper from young statisticians. Financial support from the project "The consumption behaviour of the elderly: effects on the company marketing strategies", supported by the Italian Ministry of Education and Scientific Research, is gratefully acknowledged. This paper draws on data of the LISS panel of CentERdata (Tilburg University, The Netherlands).

References

Angelini, V., Cavapozzi, D., Corazzini, L., and Paccagnella, O. (2013). Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. In *Oxford Bulletin of Economics and Statistics,* (forthcoming). On-line available at "http://onlinelibrary.wiley.com/doi/10.1111/obes.12039/abstract".

Angelini, V., Cavapozzi, D., and Paccagnella, O. (2011). Dynamics of reporting work disability in Europe. In *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 174 (3): 621–638.

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. In *Journal of Human Resources,* 46 (4): 875–906.

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

Baumgartner, H. and Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. In *Journal of Marketing Research,* (38(May)): 143–156.

Daschevici, S. (2012). *La soddisfazione nell'utilizzo dello Smartphone: alcune evidenze dal LISS panel. MSc degree,* Department of Statistical Sciences, University of Padua.

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

De Jong, M., Steenkamp, J., and Fox, J. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. In *Journal of Consumer Research,* (34): 260–278.

De Jong, M., Steenkamp, J., Fox, J., and Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. In *Journal of Marketing Research,* (45): 104–115.

Fornell, C., Johnson, M., Anderson, E., Cha, J., and Everitt Bryant, B. (1996). The American customer satisfaction index: Nature, purpose, and findings. In *Journal of Marketing,* (60): 7–18.

Gallagher, P. (2009). Using anchoring vignettes to assess cross cultural comparability of consumer ratings. In *Proceedings of the annual meeting of the American Association for Public Opinion Association*. Miami Beach.

Holland, P. andWainer, H. (1993). *Dierential Item Functioning.* Erlbaum, Hillsdale, NJ.

Kao, Y., Huang, L., and Yang, M. (2007). Eects of experiential elements on experiential satisfaction and loyalty intentions: A case study of the super basketball league in Taiwan. In *International Journal of Revenue Management,* (1): 79–96.

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

Kapteyn, A., Smith, J., van Soest, A., and Voňková, H. (2011). Anchoring vignettes and response consistency. *Wr-840,* RAND Labor & Population working paper series.

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

King, G. and Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes. In *Political Analysis,* (15): 46–66.

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

Oliver, R.L. (1993). Cognitive, aective, and attribute bases of the satisfaction response. In *Journal of Consumer Research,* (20): 418–430.

Paccagnella, O. (2011). Anchoring vignettes with sample selection due to nonresponse. In *Journal of the Royal Statistical Society: Series A (Statistics inSociety),* 174 (3): 665–687.

Paulhus, D. (1991). Measurement and control of response bias. In J. Robinson, P. Shaver, and L. Wrightsman, eds., *Measures of Personality and Social Psychological Attitudes,* 17–59. Academic Press, San Diego, CA.

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

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). *GLLAMM manual.* Working paper 160, U.C. Berkeley Division of Biostatistics Working Paper Series.

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

Scherpenzeel, A. and Das, M. (2010). True longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, and L. Kaczmirek, eds., *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies,* 77–104. Taylor & Francis, Boca Raton.

van Rosmalen, J.M., van Herk, H., and Groenen, P. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. In *Journal of Marketing Research,* (47): 157–172.

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

van Soest, A. and Voňková, H. (2014). Testing the specification of parametric models by using anchoring vignettes. In *Journal of the Royal Statistical Society: Series A (Statistics in Society),* (177): 115–133.

Vandenberg, R. and Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. In *Organizational Research Methods,* (3): 4–70.

Wand, J. (2013). Credible comparisons using interpersonally incomparable data: Nonparametric scales with anchoring vignettes. In *American Journal of Political Science,* 57 (1): 249–262.