Working Paper Series, N. 1, April 2018

Priming effects and customer satisfaction towards online banking services

Paccagnella Omar
Department of Statistical Sciences
University of Padua
Italy

Guidolin Mariangela
Department of Statistical Sciences
University of Padua
Italy

Basei Chiara
Engineering Ingegneria Informatica S.p.a.
Italy

Abstract: In this paper we apply the anchoring vignette approach for measuring the customer satisfaction of some online banking services. In particular, we investigate the extent of some priming effects when the self-evaluation question is asked after (instead of before) the vignette questions. We show that the satisfaction of respondents who answer the self-assessment question after the vignettes is usually higher than people who evaluate themselves first, when the anchoring vignettes are not known. This finding is due to the application of different response styles by the two groups of respondents. On the other hand, for respondents who have already experienced the anchoring vignettes, some weak priming effects appear only among the low educated people.

Keywords: Anchoring vignettes, Customer satisfaction, Priming, Question order.

Final version (2018-04-04)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Omar Paccagnella
tel: +39 049 827 4154
omar.paccagnella@unipd.it
http://homes.stat.unipd.it/omarpaccagnella/

Priming effects and customer satisfaction towards online banking services

Paccagnella Omar
Department of Statistical Sciences
University of Padua
Italy

Guidolin Mariangela
Department of Statistical Sciences
University of Padua
Italy

Basei Chiara
Engineering Ingegneria Informatica S.p.a.
Italy

Abstract: In this paper we apply the anchoring vignette approach for measuring the customer satisfaction of some online banking services. In particular, we investigate the extent of some priming effects when the self-evaluation question is asked after (instead of before) the vignette questions. We show that the satisfaction of respondents who answer the self-assessment question after the vignettes is usually higher than people who evaluate themselves first, when the anchoring vignettes are not known. This finding is due to the application of different response styles by the two groups of respondents. On the other hand, for respondents who have already experienced the anchoring vignettes, some weak priming effects appear only among the low educated people.

Keywords: Anchoring vignettes, Customer satisfaction, Priming, Question order.

1 Introduction

Customer satisfaction has always been a recurrent theme of marketing and it now represents a key performance indicator within business (Farris et al., 2010). Indeed, companies are addressed to a customer-oriented policy and focus on creating a stable and loyal relationship with consumers. For this purpose, it is crucial to know customers and their needs in detail (Berry and Linoff, 1999; Reinartz et al., 2004). The study and the measurement of customer satisfaction are essential for the acquisition and consolidation of long-term competitive advantages.

In simple and general terms, customer satisfaction may be seen as the overall evaluation that the buyer gives to his/her experience as a consumer, from the initial decision to the final result. Even if this idea seems easy to understand, it is difficult to provide a precise formal definition: this concept is constantly evolving and, referring

to the complete consumption experience, different aspects are involved. Therefore, there is not a single, commonly used definition. However, all the definitions which have been proposed share some common elements. In particular, Giese and Cote (2000) identify three general components: i) customer satisfaction is a response that involves both the emotional and the cognitive sphere; ii) customer satisfaction is related to a clear focus (expectations, product or service, consumption experience, and so on) and usually entails a comparison between the product's performance and some specific or general standards; iii) customer satisfaction is referred to a specific time when a certain good is chosen, when it is used or after extended experience.

Moreover, defining the quality of a service, and consequently the customer satisfaction, is much more difficult than defining the quality of a product: while there are many physical characteristics and objectively measurable data for judging the quality of goods (style, colour, label, package, durability, numbers of defects, and so on), services have mainly intangible perceptions. In most cases, the tangible evidence is limited to the service provider's physical structures, equipment and staff. Often, the quality of a service is not only based on the final result, but (especially) on the way the service is provided. Three characteristics of services must be acknowledged for a full understanding of service quality: intangibility, heterogeneity and inseparability (Parasuraman et al., 1985, 1988). Further researches by Hill and Alexander (2000) confirm Parasuraman et al. theory of service gaps.

Because of this multidimensional and unobservable structure, measuring customer satisfaction in practice is a very difficult task. Many methods have been so far introduced in the literature, but there are no universal criteria, apart from the disconfirmation paradigm (Oliver, 2010).

Overall, a satisfaction analysis may be detected by means of two different methods: the direct and the indirect ones. The first group consists in a family of approaches which measure the customer satisfaction by directly asking the consumers about their judgement, generally through an interview or a questionnaire. It is a very simple and intuitive way to collect assessments and potentially allows to record various information. However, the creation of a survey requires a high cost. Moreover, researchers rely too much on the respondents' evaluation: indeed, customers could not be able to properly determine their actual satisfaction or be honest. Indirect approaches are a set of techniques which do not directly contact consumers, but deduce the level of satisfaction through other information used as a proxy. They include indicators more or less correlated to the degree of satisfaction and refer to attitudinal or behavioural consequences, but are usually more complex than the direct ones. Therefore, the direct approach is often preferred.

1.1 Individual heterogeneity

The preference of a direct approach to perform a satisfaction analysis implies that researchers are usually interested in investigating self-reported evaluations (about satisfaction of a product, a service, a job or life, and so on) and then comparing results between different groups of respondents. However, self-assessments are subjective by definition (since people are typically asked to rank themselves on a personal scale); as a consequence, they lack in interpersonal comparability. Indeed,

individuals might interpret, understand or use the response categories for the same survey question in ways that are not the same: they might perceive differently the problem or simply differ in optimism, servility, propensity to use extreme categories and other features.

This fact may occur when comparing people from different countries, but often it can be detected even if respondents are similar according to many economic and non-economic conditions. Hence, self-evaluations are not directly comparable, so that relying on them when assessing subjective matters can be extremely misleading.

Differences across respondents may be, indeed, due to objective diversities in the domain of interest, as well as to different interpretations of the question's categories. The presence of such a heterogeneity across individuals in scale definition is known as *Differential Item Functioning*-DIF (Holland and Wainer, 1993) or *Response Style* (Paulhus, 1991). Briefly, the output of self-assessments can be seen as the sum of a real, but unobserved, evaluation and a DIF, which has to be removed in order to compare results between countries or socio-economic groups.

King et al. (2004) develop an innovative approach to deal with the DIF problem when writing survey questions: it is called *anchoring vignettes*, thus generalised by King and Wand (2007). Practically, the experiences of some fictitious characters are described in the questionnaire (the so-called anchoring vignettes) and the respondents are asked to evaluate such characters' situations by means of the same proposed categories of the self-assessment. In so doing, researchers have a reference to properly adjust the self-evaluations. Indeed, this method considers the individual heterogeneity identifying the difference in the use of the response scale within respondents.

Anchoring vignettes have found application in a growing number of papers and in different domains, from work disability (Kapteyn et al., 2007) to health (Bago d'Uva et al., 2008), from job satisfaction (Kristensen and Johansson, 2008) to life satisfaction (Angelini et al., 2014). To a less extent, this approach is present in the marketing literature (Gallagher, 2009; Paccagnella, 2011; Paccagnella et al., 2015).

When introduced in the literature, the anchoring vignettes were thought as addressed just after the self-reported question. However, Hopkins and King (2010) support an intentional use of priming of the vignettes (i.e. the set of vignettes immediately prior answering the self-evaluation), because this may help to "clarify the meaning of the self-assessment question and familiarize the respondents with the response scale, further improving measurement" (page 208). In other words, it is more likely that a respondent may understand the concept in the same way as intended by the researcher when vignettes are heard just before answering the self-assessment question.

In a survey experiment, based however on a small sample of German students, Hoffmann (2013) does not confirm the beneficial effects of reversing the vignette question administration order suggested by Hopkins and King (2010). The presence and the strength of priming effects depend upon several factors (Tourangeau et al., 2000), therefore the reversal of the vignettes' order may lead to different effects, for instance according to the context to which the anchoring vignette methodology is adopted (it is reasonable thinking that priming effects may be stronger when people have low familiarity with the research topic under investigation by the question-

naire) or question wording (Grol-Prokopczyk, 2017). Hence, Buckle (2008) claims a complete randomisation of the order of all questions, that is anchoring vignettes and the self-assessment question together. Auspurg and Jäckle (2017) show that in factorial surveys the order in which vignette dimensions are presented plays an important role, stronger or weaker according to the position of the question in the questionnaire (the largest effects occur in the extremes of the vignette sequence).

1.2 Aims and hypotheses

This paper aims at enhancing the literature on the extent of priming effects due to the placement of anchoring vignettes in a questionnaire, investigating customer satisfaction of an online banking service.

More specifically, we analyse two kinds of priming effects:

H1. *General effects*: For respondents who have never experienced the vignette instrument, does the order of the questions (self-evaluation before or after vignettes) affect the reported level of satisfaction of their online banking service?

H2. *Long-term effects*: For respondents who have experienced the vignette instrument in the past (i.e. in a previous survey), does the order of the questions (self-evaluation before or after vignettes) affect the reported level of satisfaction of their online banking service?

2 Data and Methods

2.1 The Anchoring Vignettes

Vignettes have a long history in investigating social phenomena (Nosanchuck, 1972) and may be defined as systematically elaborated descriptions of a concrete situation in the domain of interest. Usually, each vignette describes the same scenario, varying the level or the characteristics of the "the most important factors in the decision-making or judgement-making process of respondents" (Alexander and Becker, 1978).

The Anchoring Vignettes introduced by King et al. (2004) have the same structure of standard vignettes (a brief text where a hypothetical individual is described in a particular condition related to the domain of interest), but differentiate from them because the level of the domain of interest is fixed across respondents in each question. Therefore, in different anchoring vignettes, different scenarios are proposed. Some examples of anchoring vignettes for a particular domain of political efficacy ("say in government" through elections) are provided by King et al. (2004):

Alison lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

Jane lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition

party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

Moses lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

Respondents are then asked to evaluate these scenarios, as well as the own status in the same domain investigated by the anchoring vignettes. According to the previous examples, the question is: "*How much say [does/do] [name of person/you] have in getting the government to address issues that interest [him/her/you]?*. The response categories are: 1. Unlimited Say; 2. A Lot of Say; 3. Some Say; 4. Little Say; 5. No Say at All. Many other examples are presented in https://gking.harvard.edu/vign.

Variations in categorical responses are also caused by the usage of different cut-points between response categories. Vignettes provide an anchor scale, equal for all respondents, that adjusts their self-evaluations: after these corrections, self-assessments can be compared across countries or socio-economic groups, because all subjective evaluations are now reported to a common DIF-free scale.

A practical example of how these vignettes help us to adjust self-evaluations is provided in Figure 1. The illustration displays two respondents, indicated by 1 (on the left) and 2 (in the middle), who answer one self-assessment and three vignette questions, where an evaluation is asked about three fictitious characters (Alison, Jane, Moses). The reported response is drawn as a line, so it is considered as continuous, to simplify the explication. We may immediately see that both individuals rank in the same way the vignettes (Alison with a higher level, Jane in the middle and Moses with a lower level) and it seems that the self-evaluation of political efficacy is higher for the first individual. However, the vignettes' actual level is the same no matter which respondent is taken into consideration and the two respondents evaluate them in a different way, confirming the presence of individual heterogeneity. Thus, it is possible to remove the effect due to DIF and compare the two individuals only by rescaling the second respondent's evaluations, so that vignette assessments for the two respondents match. Relying on the common scale shown on the right, the actual conclusion changes: the first respondent has a lower level of political efficacy than the second one.

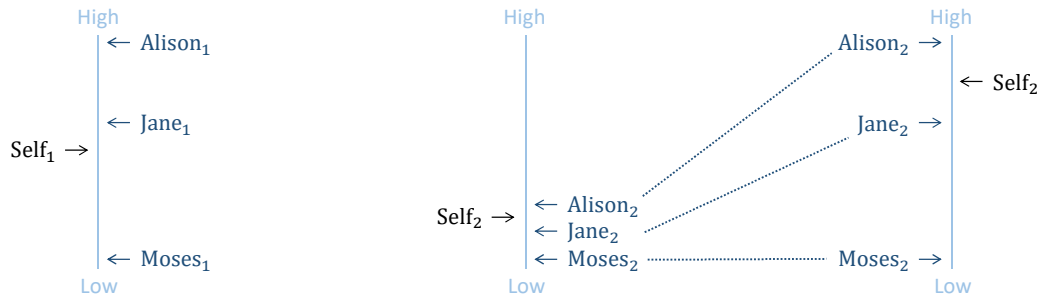

Figure 1: Comparing preferences (King et al., 2004).

When using anchoring vignettes, there are some important matters to discuss:

the characteristics of the hypothetical individuals, the number of vignettes to be included in the questionnaire and their order. First of all, the way a vignette is written is really important: it should be clear and short, and it should be written so that people with different backgrounds and cultures understand it as similarly as possible. Therefore, a particular attention should be paid to question wording, accurate translation of the meaning of different items, and so on. Moreover, vignettes should be highly concrete and with a high discriminatory power, in order to avoid repetitive information. The hypothetical individual described in the vignette should be appropriate to the language and the culture of respondents and, ideally, have the same characteristics as the respondent. Thus, if possible, it is preferred to change the names on the vignettes to match the respondent's gender and age. The number of anchoring vignettes to include in the survey depends on several factors: the sample size, the nature of the DIF, the used model, and so on. The proposed statistical solutions require the collection of just one vignette, but including several vignettes could allow to obtain more information. Empirical applications show that two or three vignettes are enough. In general, finding the right trade-off in bias reduction and survey costs is needed. Indeed, introducing anchoring vignettes in a survey questionnaire adds several sources of additional costs in term of survey design and reduces the time available for collecting other information. Costs can be reduced by an appropriate choice of the number of vignettes and, more importantly, they should be compared with the potential benefits of this approach: the general idea is that anchoring vignettes may provide correction for individual threshold values, which can be applied in following studies without the need to ask again the same questions in the future.

Two fundamental assumptions are needed for the validity of the anchoring vignette approach: *response consistency* and *vignette equivalence*.

According to response consistency, we assume that response categories are the same in the self-assessments as well as evaluating vignettes. The idea is that each person applies approximately the same DIF in answering to both the self-evaluation question and the vignettes. This assumption allows to correct the self-evaluation for interpersonal differences using the vignettes as anchors. If the thresholds applied by each respondent change between questions, it is no more possible to use the evaluation given to the anchoring vignettes as a standard. There could be many cases where response consistency is violated. For example, people who generically feel inferior to others might use different thresholds to evaluate themselves and the vignettes. In particular, they might apply a higher or lower, depending on the domain, response scale to themselves and this would bias the measurements.

The vignette equivalence assumption states that all respondents perceive in the same way the underlying actual level of the variable described in any vignette. In other words, all interviewees agree on the real unobserved level for each vignette and place it at the same location on the latent scale. So, the perception of each vignette does not depend on the individual characteristics of the respondents. Of course, different individuals may apply their own DIF in choosing response categories, even if everybody understands vignettes in the same way. The vignette equivalence is required to obtain a DIF-free measurement to be used as an anchor: thanks to the vignettes, it is possible to measure the thresholds of each respondent because the

differences in the vignette evaluation are only caused by DIF. This assumption would be violated if different respondents understand the vignette in different ways. For example, considering the health condition, an overweight vignette character might be considered unhealthy by residents of developed countries: they might associate overweight with an increase of diseases and risk of diabetes. On the contrary, it might be seen as healthy by citizens of low-income countries, who feel obesity as a sign of good nourishment.

Grol-Prokopczyk (2014) illustrates the idea underlying the anchoring vignette method by means of the example reported in Figure 2. In this case, there are three different groups which evaluate three vignettes and their own health condition. The response categories are: 1. Poor; 2. Fair; 3. Good; 4. Very Good; 5. Excellent. The response consistency essentially means that the thresholds (τ_1, τ_2, τ_3 and τ_4) of the three groups stay in the same position both in the vignette evaluation and in the respondent's rating. The other assumption is indicated by the fact that each vignette is a horizontal line: each hypothetical description has the same actual unobserved level of health for all groups, even if it is reported differently. If this assumption was violated, the vignette lines would cross each health spectrum at different heights.

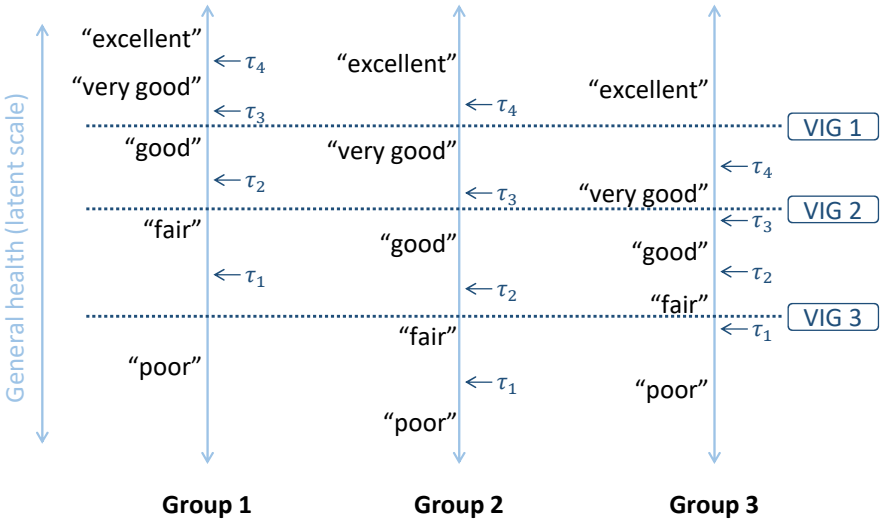

Figure 2: Using anchoring vignettes to estimate reporting heterogeneity (Grol-Prokopczyk, 2014).

The validity of these assumptions has been criticised in the literature. King et al. (2004) affirm that these problems "would have to occur at what would seem to be unrealistically extreme levels to make the unadjusted measures better that the adjusted ones". Anyway, testing the validity of these assumptions is still an open research topic in the literature and no formal tests, without imposing the occurrence of other conditions, have been so far introduced. Indeed, it is very challenging because neither thresholds nor perceived absolute levels can be directly observed. See Paccagnella (2013) for a review.

2.2 The Statistical Solutions

King et al. (2004) propose two statistical solutions to exploit the information collected by the anchoring vignettes: a non-parametric and a parametric approach.

The non-parametric method is a simple and intuitive solution which enables to correct DIF without sophisticated techniques and it is also easy to implement. No new assumptions are required, in addition to response consistency and vignette equivalence. Moreover, this approach does not need explanatory variables. However, the non-parametric solution has two big drawbacks, in addition to the usual problems of the non-parametric approaches. First, it needs the answers of the self-assessment and all anchoring vignettes for every individual. Then, all respondents have to rate the vignettes in the same order (the so-called *natural order*): cases, which this does not happen, are grouped and treated as ties. Ties do not permit to adjust the measure and entail the loss of information, which leads to inefficiency. Because of these weaknesses, the non-parametric method has not been found so far a wide application in the literature, in favour of the parametric approach instead.

The parametric solution proposed to apply the anchoring vignettes is called *chopit* (Compound Hierarchical Ordinal Probit) model, sometimes also labelled as *hopit*. It can be seen as a generalisation of the ordered probit model, as it basically consists in a joint estimation of some ordered probit models. However, the ordered probit is not nested in the *chopit* specification.

Indeed, as in an ordered probit model, a latent variable is observed through an ordinal response variable, defined by means of some cut-points. While these thresholds are not allowed to vary across respondents in the ordered probit solution, in the *chopit* specification the vignettes' information is exploited to modelling the DIF through variations in the thresholds, which are therefore functions of some individual characteristics. After the identification of response scales for each respondent, we can easily correct the self-assessment answers.

The *chopit* model can be divided into two parts with a similar structure: the self-assessment component and the vignette component. Indeed, for each respondent and question, three levels of the variable of interest are present:

- the **actual level**, which represents the real unobserved value and is measured on a continuous scale;
- the **perceived level**, which is the unobserved and unbiased perception of the actual level, measured on a continuous scale, corrected with a noise;
- the **reported level**, which returns an observed value from the perceived level by choosing one of the ordered response categories. Everyone systematically uses different thresholds and, therefore, this is the only incomparable (among respondents) level, due to the presence of DIF.

The *chopit* model overcomes the inefficiencies of the non-parametric solution by recognising that the variable of interest is perceived with a random error, which explains why some respondents do not assess the vignettes with their natural order. Figure 3 shows how these values are linked (each solid arrow evidences a deterministic effect) and summarises the structure of the model.

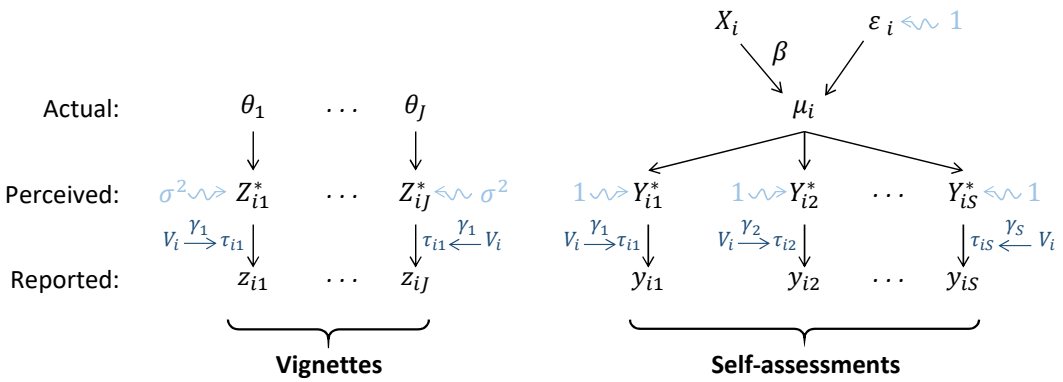

Figure 3: Components of the *chopit* model (King et al., 2004).

In the following description of the statistical model we consider a survey with only a self-assessment question, but the generalisation is simple. Let μ_i be the actual level of respondent i ($i = 1, \ldots, n$) which is perceived by individual i only with a random error, like in the ordered probit model. We denote the unobserved perceived level as Y_i^* and

$$Y_i^* \sim N(\mu_i, 1).$$

The actual level varies over i and is the result of a linear function:

$$Y_i^* = \mu_i = X_i\beta + \epsilon_i, \qquad \epsilon_i \sim N(0, 1)$$

where the X_i's are observed covariates, β is the coefficients' vector and ϵ_i is an independent and identically distributed random effect, independent of the set of exogenous variables ($\epsilon_i \perp X_i$). The vector β does not include a constant for the model identification and, for the same reason, the unit variance of the error term is required. The noise ϵ_i includes reporting error and/or unobserved heterogeneity.

Respondent i is asked to turn his/her continuous perceived level Y_i^* into a reported category y_i by means of this criterion:

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

where τ_i^k is the threshold which divides the $(k-1)$-th and the k-th categories for respondent i, and $-\infty = \tau_i^0 < \tau_i^1 < \ldots < \tau_i^K = \infty$. Thresholds vary across observations as functions of some exogenous variable V_i (which may overlap X_i) and a vector of parameter γ:

$$\begin{aligned} \tau_i^1 &= \gamma^1 V_i, \\ \tau_i^k &= \tau_i^{k-1} + exp(\gamma^k V_i), \qquad k = 2, \ldots, K-1, \end{aligned}$$

where the exponential form guarantees that thresholds increase with k. The variation of cut-points makes the reported level incomparable across respondents, because people apply different threshold values to turn their perceived levels into a category.

It is not possible to estimate only the self-assessment component. Therefore, we add a vignette component to increase the information content. Regarding the vignettes component, each respondent i is characterized by one vignette equation for each anchoring vignette. Let θ_j $(j = 1, \ldots, J)$ denote the actual level for the hypothetical person described in vignette j. According to the vignette equivalence assumption, it is perceived in the same way by all respondents. Indeed, θ_j does not depend on index i. Respondent i perceives this continuous and unobserved value with a random normal error, as Z^*_{ij}:

$$Z^*_{ij} = \theta_j + u_{ij}, \qquad u_{ij} \sim N(0, \sigma^2_u)$$

The error term is independent of ϵ_i, X_i and V_i $(u_{ij} \perp (\epsilon_i, X_i, V_i))$ and its variance is assumed to be the same across respondents and vignettes. However, it is possible to let σ^2_u vary over vignettes and their estimates can be seen as a indicator of how well each vignette is understood.

As before, the perceived value Z^*_{ij} is turned into a categorical answer by means of the same thresholds τ^k_i $(k = 1, \ldots, K)$ described above:

$$Z_{ij} = k, \qquad if \quad \tau^{k-1}_i \leq Z^*_{ij} < \tau^k_i.$$

The unchanged thresholds in both the self-assessment and the vignette component respect the response consistency assumption. The vignette equivalence, as mentioned before, imposes that θ_j does not vary across respondents, and so the differences in vignette evaluations are only a function of DIF. As a consequence of both assumptions, the vignettes allow to identify the type of DIF for each person and, consequently, to estimate individual thresholds. Then, with this information, adjusting the self-assessment is easy, as estimating β parameters.

The *chopit* model is estimated by means of maximisation of the log-likelihood. The self-assessment and the vignette components have their own likelihood functions, which are joined together to obtain the overall likelihood, since the error terms are independent of each other. Basically, the contribution of the self-assessment is an univariate ordered probit with varying thresholds:

$$L_s(\beta, \gamma \mid Y) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\Phi(\tau^k_i \mid X_i\beta, 1) - \Phi(\tau^{k-1}_i \mid X_i\beta, 1)]^{I(y_i=k)}$$

The likelihood function for the vignette component is also an ordered probit, but a J-variate one:

$$L_v(\theta, \gamma \mid Z) = \prod_{i=1}^{n} \prod_{j=1}^{J} \prod_{k=1}^{K} [\Phi(\tau^k_i \mid \theta_j, \sigma^2_u) - \Phi(\tau^{k-1}_i \mid \theta_j, \sigma^2_u)]^{I(z_{ij}=k)}.$$

The complete likelihood is:

$$L(\beta, \theta, \gamma, \sigma^2_u \mid Y, Z) = L_s(\beta, \gamma \mid Y)\, L_v(\theta, \gamma \mid Z)$$

The parameters β can be interpreted just as in an ordered probit: a positive coefficient is associated with a positive relationship with the actual value of interest Y^*_i.

We have seen that the *chopit* model overcomes the inefficiencies of the non-parametric method and allows the thresholds to vary across respondents. Moreover, the answer to all vignettes from each individual is not needed for its estimation. Alongside with its advantages, there are also different criticisms due to the assumptions: the response consistency is required for model identification; thus, the model specification includes other assumptions, like the linear relation, the form of the thresholds and the distribution of errors.

Several extensions of the *chopit* model have been so far introduced in the literature (see Paccagnella (2013) for a review). Among them, we mention the Kapteyn et al. (2007)'s proposal: they extend the standard version of the model by introducing an unobserved individual effect in the threshold equation. In order to control for individual unobserved heterogeneity, the response scale equation is replaced by:

$$\begin{aligned} \tau_i^1 &= \gamma^1 V_i + \eta_i, \\ \tau_i^k &= \tau_i^{k-1} + exp(\gamma^k V_i), \qquad k = 2, \ldots, K-1, \end{aligned}$$

with $\eta_i \sim N(0, \sigma_\eta^2)$ and assumed to be independent of both X_i and of the other error terms in the model. This solution models the thresholds both with a set of observed individual features and with an unobserved individual heterogeneity term η_i. Moreover, this extension entails that different vignettes' assessments are correlated with each other. Indeed, it is obvious to think that individuals tend to use high or low cut-points in all their evaluations. When σ_η^2 is null, the model is equal to the original *chopit* solution. van Soest and Voňková (2014) show that such an extended approach is able to substantially reduce some misspecification problems of the original *chopit* specification.

2.3 The Questionnaire

Data analysed in this paper are collected by means of a questionnaire investigating online banking service in Italy and the corresponding customer satisfaction. It was carried out by a team of researchers from the Department of Statistical Sciences at the University of Padua and submitted by Doxa (Institute for Statistical Research and Analysis of Public Opinion) in two periods: May 2015 and September 2015. Only one member of each family was interviewed, using the CAWI (Computer Assisted Web Interviewing) methodology.

About half of the respondents to the second survey has also participated to the first one. The whole questionnaire is made up of 23 questions divided in 3 sections, which collect information regarding different points of view of the online banking customer experience[1].

The first section is a screening, in order to select people owning at least one bank account which allows also online operations. A key question in this section asks for the types of services experienced in the individual online operations.

Therefore, the questionnaire focuses on the satisfaction related to some online banking operations and is divided in two parts with the same structure (the remaining two sections): people who browse the online bank account (like checking the

[1]The questionnaire is in Italian; in this paper we report translations of the most important questions.

account balance and movements) answer the first five questions; then, individuals who carry out operations (as paying taxes, stamp duties, utilities, etc. or making a bank transfer) assess their satisfaction in the remaining five questions. According to the used type of services, respondents may complete one or both sections.

About browsing the main bank account (the focus of this work), people are first asked to evaluate their expectations and experiences, in a scale from 1 (completely disagree) to 10 (completely agree). Then, the self-assessment question is proposed: *How satisfied are you with the ease of online browsing your main bank account?*. The available answering categories are: 1. Very Satisfied; 2. Satisfied; 3. Neither satisfied, nor dissatisfied; 4. Dissatisfied; 5. Very Dissatisfied.

A brief text now introduces the vignette part: "We will now give you two examples of persons who experienced the online browsing of a bank account. We would like to know how you evaluate their satisfaction regarding the ease of online browsing their bank account. Please, imagine that the persons have the same age and background that you have." As previously explained, anchoring vignettes describe hypothetical customers with different satisfaction levels and respondents have to assess how much these individuals are satisfied. In particular, the proposed scenarios are:

Carlo is an employee and opened an online bank account 3 years ago. Every day he checks the movements in his account, in order to verify the existence of possible irregular movements. Carlo goes in the website, finds the bank account section and then selects Account movements" in the drop-down menu. Then, he clicks on Last ten movements" and checks the list. Even if it takes some seconds to load the list, Carlo needs less than a minute to complete the whole control procedure.

Marina is a housewife who checks the list of her family's expenses with the credit card, more or less every 3 days. One day, she wants to check the expenses of the previous month again, but she does not find the drop-down menu to select the right month. She needs to contact the call-center in order to solve the problem. Thanks to the operator, she succeeds in finding the list of movements she was looking for.

After each description, respondents have to answer using the same response categories adopted for self-evaluation: *How satisfied is Carlo/Marina with the ease of online browsing his/her main bank account?*

Before asking the self-assessment question and the evaluation of the vignettes, the sample is randomly divided in four groups which differ in the question order: the first two groups answer first the self-assessment and then the vignette questions, but with a different order of the Carlo and Marina questions; conversely, the other two groups give an evaluation of the hypothetical scenarios before evaluating their personal experience. Therefore, the question order for every group is:

Group 1: self-assessment, Carlo vignette and Marina vignette

Group 2: self-assessment, Marina vignette and Carlo vignette

Group 3: Carlo vignette, Marina vignette and self-assessment

Group 4: Marina vignette, Carlo vignette and self-assessment

2.4 The dataset

In our datasets, only fully completed questionnaires are included, so that all respondents have at least one bank account which allows online operations.

In May 2015, 1031 household members completed the questionnaire, whereas 1063 individuals participated to the second wave (September 2015). The 52.2% of these respondents has already answered to the questionnaire in the first wave, but 8 of them did not have an online account at the first time. Thus, 515 interviewees have seen the whole questionnaire for the first time in September 2015.

In both waves there is a slight majority of male respondents (57.4% and 54.6%, respectively). The average age is equal to 43 years in the first sample and 44 years in the second one: for both periods the oldest respondent is 85 years old, while the youngest one varies from 18 to 19 years old. The distribution of the age classification is shown in Figure 4. In both cases, respondents mainly belongs to the classes 25-34 and 35-44 years.

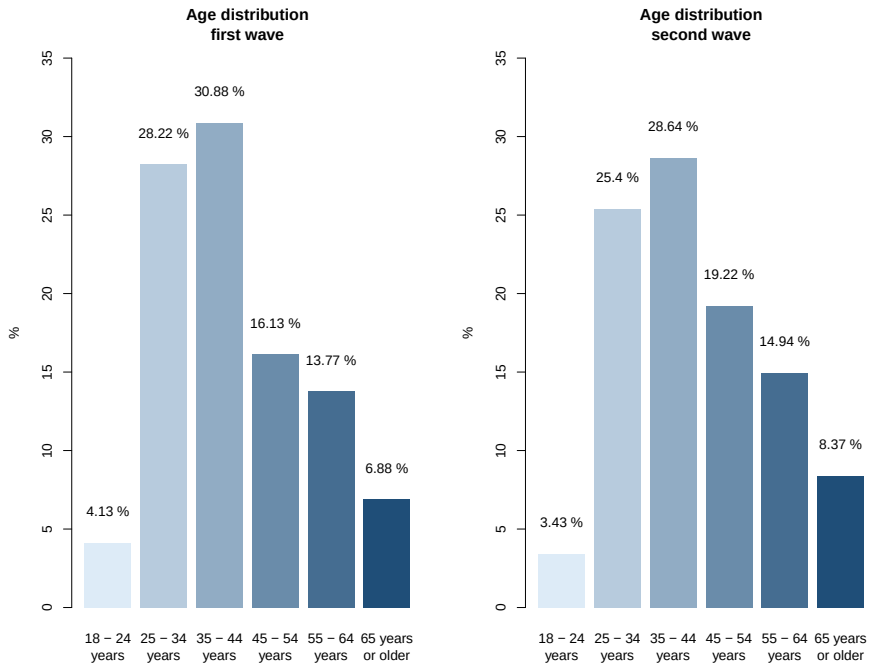

Figure 4: Age distribution of the sample in both waves.

More than half of the respondents has a medium education, while only 10% of them reports a compulsory (or lower) level of education. Most of the people (about the 48% in both surveys) are employed, while the proportion of self-employment is about 16%.

Overall, the explanatory variables that will be specified in the *chopit* model are:

- **gender**: it assumes value 1 if the respondent is a female and 0 if he is a male;
- **age**: it is a categorical variable with three classes and indicates the age of the individual: from 18 to 34 years, from 35 to 54 years, 55 years or older;
- **area**: it corresponds to the geographic area of residence (four categories): North-West, North-East, Central, South & Insular Italy;
- **employment**: it assumes value 0 if the respondent is an employee, 1 if he/she is self-employed and 2 otherwise (i.e. retired, housewives, unemployed and so on);
- **household size**: it is a factor variable with five levels (1, 2, 3, 4, 5 or more) and reports the number of the household members;
- **education**: it is a categorical variable to determine the educational background of the respondent and is divided in low (compulsory school), medium (high school) and high level of education (university degree or above);
- **group**: it assumes value 1 if the person belongs to group B (i.e. the self-assessment question is asked after the vignettes) and value 0 for those belonging to group A (i.e. the self-assessment question is asked before the vignettes);
- **problem**: it is equal to 1 when the respondent had a problem when browsing or managing the main bank account and 0 otherwise.

3 Results

The mostly reported online services are shown in Figure 5. The online bank account is used by almost all respondents in order to check the account balance and movements. In general, all services listed in the questionnaire are used by a large amount of respondents, apart from trading online and loan management (probably, customers prefer to relate in person to an agency contact when they have to handle complex and important transfers). If we compare the waves, the percentage of loan management decreases after the four months; on the other hand, we notice a remarkable increase in the use of all the other services.

Referring to the absolute frequency, the first four categories include 1017 units for the first questionnaire and 1051 units for the second one, respectively. As mentioned above, only these respondents are filtered and will answer the section related to the browsing satisfaction.

About 30% of the respondents found some problems when browsing or managing their online bank account. Most of them contacted the call center (about 11%) or solved the issues by themselves (about 10%).

Figure 6 shows the distribution of the self-assessments in both waves. In general, the respondents are satisfied of the service: both the distributions are right skewed and the categories "Very satisfied" and "Satisfied" include more than 90% of the sample. Even if the two histograms are very similar, people seem a little bit less satisfied in the second survey. This may be caused by a real decrease of

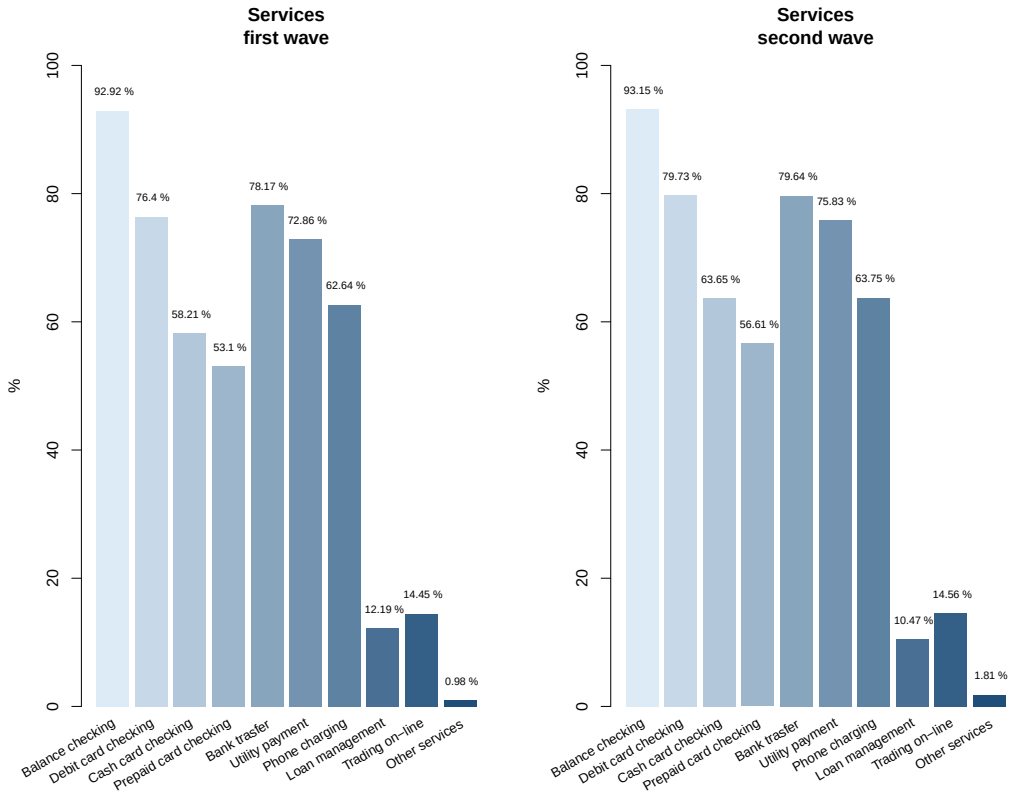

Figure 5: Type of used services in the analysed samples.

the satisfaction, but may also be due to the composition of the sample. As a matter of fact, interviewees of September 2015 can be divided in two groups of similar numbers: new and already-interviewed respondents. The latter already know (or should know) the vignette instrument from the first wave and may answer in a more reasoned way remembering how the questionnaire is structured. This reaction may be due to memory effects.

As explained above, the sample of respondents is divided in four groups, with different question orders. We summarize this division in two classes: *group A*, which includes groups 1 and 2, and *group B*, which gathers groups 3 and 4. People belonging to group A read the self-assessment question before the anchoring vignettes, which entails a more instinctive answer: respondents are likely affected by the mood of the moment and recent problems take a greater weight than past issues. For example, if the respondent had a problem with his/her online account the day before the survey, he would probably answer negatively. Instead, a problem happened in the past may be easily forgotten, so that the answer could be positive. On the other hand, respondents of group B give their opinion first on the vignettes and then on their own satisfaction. The scenarios described in the vignettes prepare people to the self-assessment question, and this causes a more rational and thoughtful answer. Because of this priming effect, respondents may have the opportunity to reflect on their own experience with the service and compare it with the vignette. Briefly,

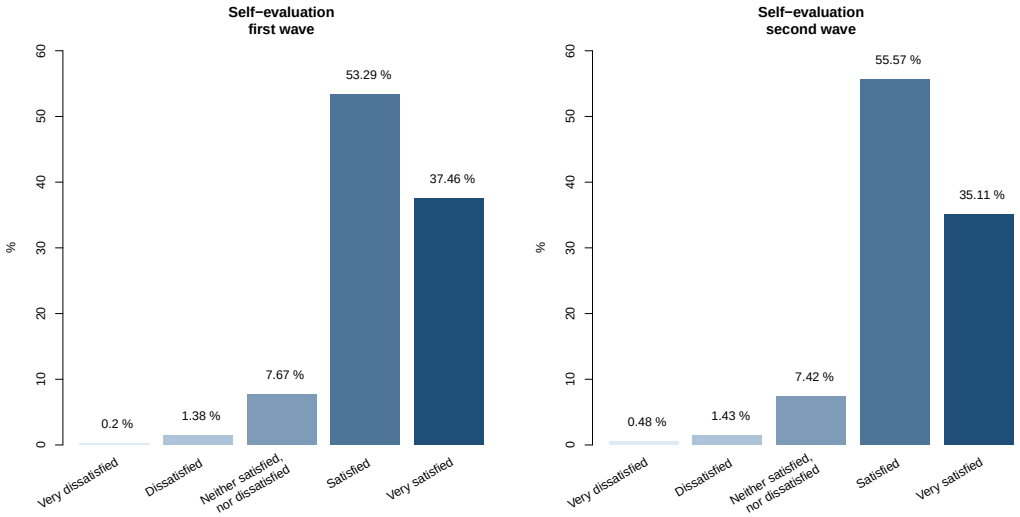

Figure 6: Distribution of the self-reported satisfaction between waves.

Table 1 reports the proportion of each group in the two waves. As we can seen, the distribution is almost equal and neither A nor B stands out.

Table 1: Group distribution in the analysed samples.

Group	First wave	Second wave
Group A	49.85%	49.19%
Group B	50.15%	50.81%

We now study if question order produces some effects on the distribution of the self-reported customer satisfactions. First, Figure 7 compares the satisfaction of the first and second sample in general. In both histograms, the categories "Very dissatisfied", "Dissatisfied" and "Neither satisfied, nor dissatisfied" are less used in group B than in group A. Indeed, in the first wave 12.2% of group A's respondents uses these levels, versus 6.3% of group B. The same happens in the second wave: 12.2% versus 6.6%. Thus, people who answer the vignettes before the self-evaluation question seem to be more satisfied than those who read the hypothetical scenarios after the self-assessment. Maybe, respondents find themselves more gratified of their own condition after reading other probable situations. For example, let us consider a respondent who has never experienced the issue represented in the vignette: when self-evaluating after reading the vignette, he/she would probably express a higher satisfaction about his/her own experience, since he/she realises he/she had never had the issue represented in the fictitious situation. So, his/her self-assessment would be higher with respect to people who answer the self-evaluation first. If we consider only the right plot, the dimensions of the last two categories are a little bit different between groups A and B: the percentage of "Very satisfied" is 38.5% in group A and 36.5% in group B, while the "Satisfied" is the 49.3% in group A and the 57.3% in group B. Respondents belonging to the first group seem to frequently use extreme

levels, whereas answering the vignettes before the self-evaluation leads to a more moderate response. Preliminary results show how anchoring vignettes prime the level of the customer satisfaction of the analysed online banking services.

The differences between the two groups are only partly present in the second wave: the "Satisfied" category is 54.7% in group A and 56.4% in group B, but the higher level is used a little bit less in group A than in group B (33.1% versus 37.1%). The satisfaction distribution is more similar between groups in the right plot and this fact is probably due to the participation of already-interviewed respondents to the survey.

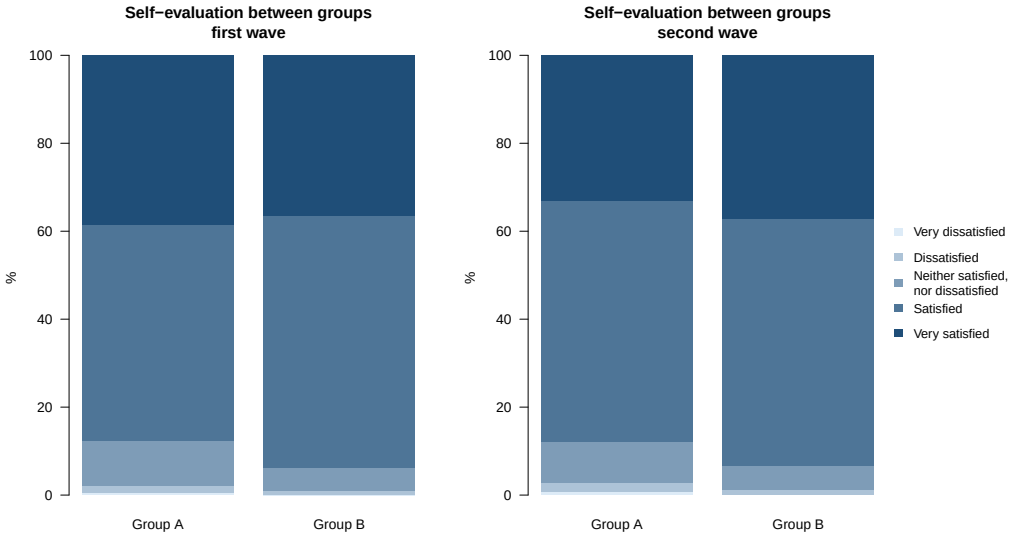

Figure 7: Distribution of the self-reported satisfaction between groups in the two waves.

Comparing the sub-samples of already-interviewed and new respondents in wave 2 (Figure 8), the higher satisfaction of group B clearly appears also in this case: among new respondents, 11.4% of group A is not satisfied, instead of only 5.7% of group B; considering the already-interviewed respondents, the percentage of *not satisfied* (which includes the "Very dissatisfied", "Dissatisfied" and "Neither satisfied, nor dissatisfied" categories) changes from 12.9% in group A to 7.4% in group B. Apart from this, the satisfaction distribution of already-interviewed individuals is similar between the two groups (the differences in the two prevalent levels are smaller than 3%). Hence, people who know the survey (and the anchoring vignettes) appear to be influenced by the question order less than new respondents. Therefore, priming effects seem no longer present when people have experienced at least once the anchoring vignettes' instrument in the previous survey.

Analysing the vignettes answers, Figure 9 shows how much Carlois perceived satisfied with his online banking account. In both waves, more than 90% of the interviewees assesses his condition as "Very satisfied" and "Satisfied". Comparing the two graphs, the highest category decreases from 44.4% to 42.2% in the right plot; on the other hand, the percentage of "Satisfied" changes from 46.1% to 48.9%.

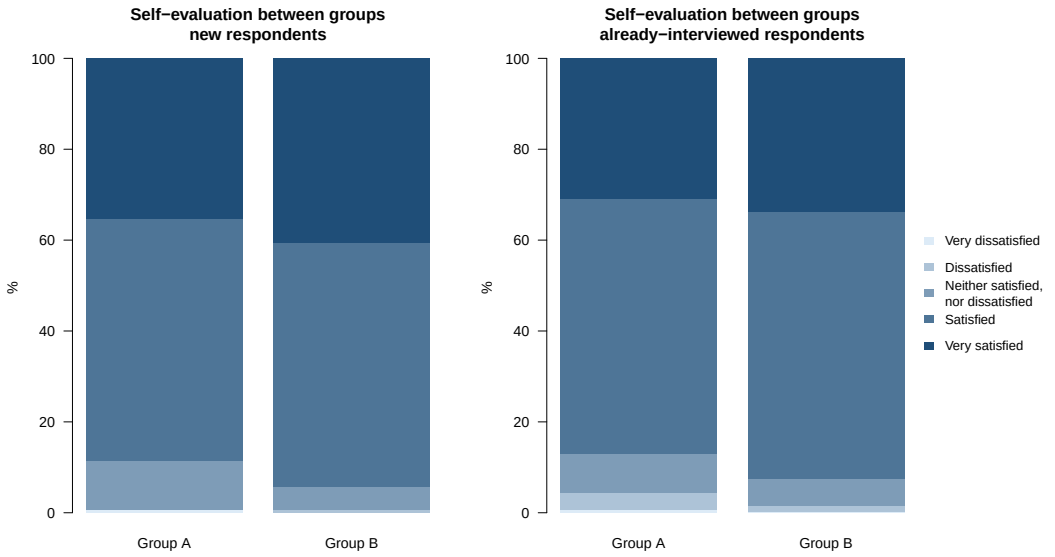

Figure 8: Distribution of the self-reported satisfaction between groups of new and already-interviewed respondents of the second wave.

Figure 10 shows the evaluation of the Marina's vignette. Respondents assess her situations more negatively than Carlo's scenario, which is reasonable since she experienced a difficulty when browsing her online account. Indeed, considering only the first wave, the 30.3% of interviewees chooses the "Satisfied" and 10.1% the "Very satisfied" category. However, the figure shows a clear difference between the distributions in the two waves: in particular, in May 2015 the percentage of "Satisfied" respondents (30.3%) exceeds the "Dissatisfied" ones (28.3%) and the same relationship happens between the extreme levels: people very satisfied (10.1%) are more than the very dissatisfied ones (8.6%). In September 2015 these relationships trade places: the frequency of "Dissatisfied" is 31.8%, more than the 28.6% of "Satisfied" and the "Very dissatisfied" category's percentage is 10.8% against the 7.8% of "Very satisfied". This opposite trend suggests that the evaluation is quite different between the two waves, but it could be also due to the composition of the sample (already-interviewed and new respondents), as mentioned before.

3.1 Study 1

In order to check the extent of general priming effects (hypothesis **H1**) on the reported customer satisfaction, we estimate the Kapteyn et al. (2007) version of the *chopit* model to two different samples of respondents. The first model is applied to all 1017 individuals from the first wave (May 2015) who have answered the questions about the browsing satisfaction. The second model is applied to the subsample of respondents who took part for the first time at the survey in September 2015 (second wave): in this case we consider 510 individuals, who had never answered the questionnaire before, therefore, did not know the anchoring vignette tool at the time

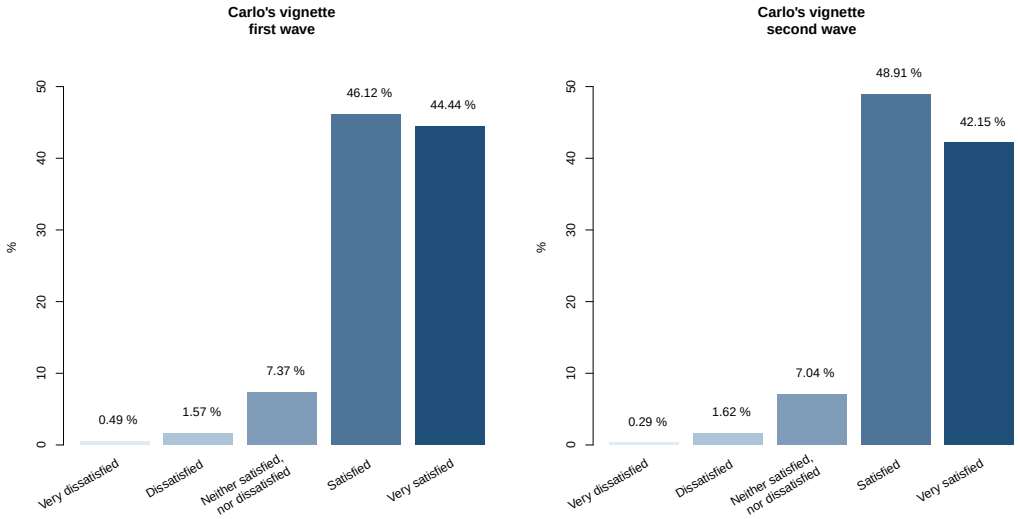

Figure 9: Carlo's vignette evaluation in the two waves.

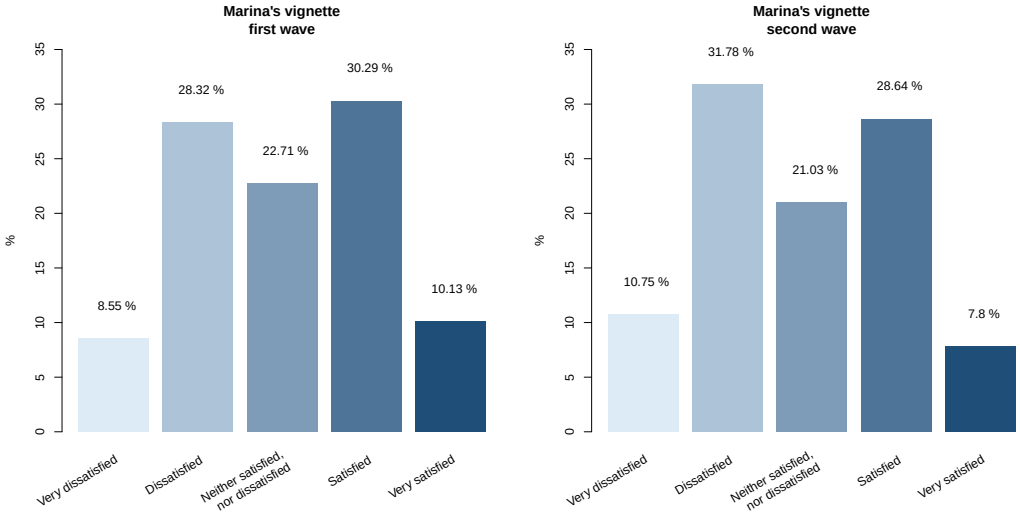

Figure 10: Marina's vignette evaluation in the two waves.

of their interview.

3.1.1 First wave sample

Table 2 displays the estimated parameters of a standard ordered probit regression: the panel on the left lists the estimates of the self-evaluation component, while estimated thresholds values are reported on the right panel. Table 3 shows the results of the *chopit* model estimation: the first column lists the estimates referred to the self-evaluation component, while in the other four columns the estimated coefficients of the threshold equations are shown. The other panels present the estimated parameters characterising the anchoring vignette component and all variance estimates

(in terms of log standard deviation).

Table 2: Ordered probit model estimates for the first wave sample.

Variable		Self-assessment
Gender		-0.048
Age	35-54 years	0.184**
	55 years or older	0.259**
Area	North-East	-0.320***
	Central	-0.107
	South and Insular	-0.137
Employment	Self-employed	-0.125
	Other	-0.154*
Household size	2	-0.019
	3	0.180
	4	0.089
	5 or more	-0.011
Education	Medium	-0.024
	High	-0.123
Group		0.104
Problem		-0.692***

Thresholds	
τ_1	-3.306
τ_2	-2.530
τ_3	-1.641
τ_4	0.125

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

The comparison of these two model estimations allows to first investigate the presence of reporting heterogeneity effects. Indeed, the ordered probit model does not allow a different use of the response categories among people with different characteristics, so all differences are accounted by the coefficients in the main equation.

The appropriateness of the *chopit* model is immediately noticeable. See the last four columns: some variables in the thresholds are significantly different from zero, meaning that reporting styles significantly vary according to some individual characteristics. Therefore, the role played by the reporting heterogeneity should not be neglected and the anchoring vignettes, and consequently the *chopit* model estimation, are appropriate and needed.

In order to statistically verify that the thresholds vary according to different characteristics, we calculate some Wald tests of linear hypotheses for testing if the parameter estimates, except the intercepts, are jointly different from zero, at least for one threshold equation. The parameters of all the thresholds are jointly significant at a 1% level ($\chi^2_{64} = 132.68$, $p-value = 0.00$), confirming that the reporting categories change according to some individual features. Considering each threshold equation separately, the coefficients of the first threshold are not statistically different from zero ($\chi^2_{16} = 19.02$, $p-value = 0.27$), while the estimates of the second ($\chi^2_{16} = 27.91$, $p-value = 0.03$), the third ($\chi^2_{16} = 30.50$, $p-value = 0.02$) and the fourth ($\chi^2_{16} = 29.22$, $p-value = 0.02$) threshold are significant at a 5% level.

Table 3 also shows the estimates of the vignette equation parameters (on the left) and the variance components in term of log standard deviation (on the right). Only the second vignette is significantly (and negatively) different from zero and

Table 3: *Chopit* model estimates for the first wave sample.

Variable		Self-assessment	Threshold equation coefficients			
			γ^1	γ^2	γ^3	γ^4
Gender		-0.535***	-0.164	-0.111	-0.042	-0.052
Age	35-54 years	0.204	0.376*	-0.171	-0.097	-0.021
	55 years or older	0.280*	0.296	-0.227	0.004	0.008
Area	North-East	-0.174	-0.102	0.075	0.206	0.004
	Central	-0.087	-0.146	0.039	0.048	0.051
	South and Insular	-0.267*	-0.121	-0.001	0.156	-0.082
Employment	Self-employed	-0.524***	-0.185	-0.061	0.082	-0.103
	Other	-0.454***	0.198	-0.338**	0.136	-0.029
Household size	2	-0.162	0.100	-0.033	0.041	-0.152*
	3	-0.078	-0.146	0.195	-0.241	-0.157*
	4	-0.052	-0.126	0.183	-0.174	-0.103
	5 or more	-0.223	-0.056	0.230	-0.346	-0.129
Education	Medium	-0.064	0.946**	-0.232	-0.411***	-0.043
	High	0.234	1.213***	-0.188	-0.421***	0.028
Group		0.195*	0.088	-0.130	-0.057	0.161***
Problem		-0.828***	-0.328	0.079	0.167*	0.056
Constant			-6.188***	0.896***	0.419	0.929***

Vigenttes	Coefficients
θ_1(Carlo)	-0.364
θ_2(Marina)	-2.828***

Log standard deviation	
Vignettes	0.439
Self-assessment	0.000
Thresholds	0.673

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

this means that the evaluation given to Marina's experience takes a lower value and, so, a lower satisfaction than Carlo's situation. People are less satisfied with Marina's vignette and this result strengthens the descriptive evidence. This outcome is reasonable, since a problem in the browsing of the bank account occurs in the second vignette, while Carlo's vignette just describes a standard situation. The panel reporting the log standard deviation estimates shows another interesting result: the unobserved heterogeneity term in the threshold equations has the largest estimated variance among all variance components, supporting the benefit of such specification of the *chopit* model.

According to the ordered probit model estimates, the only variables that significantly relate to the online browsing satisfaction are *age*, *area*, *employment* and *problem*. In particular, other things being equal, as age increases, people are more satisfied of the service. The other three variables, instead, have an opposite behaviour: *ceteris paribus*, a resident in the North-East Italy is less satisfied than an individual from another area; those who belong to the job class "Other" are less satisfied than employed and self-employed interviewees; respondents who had a problem in their experience are less satisfied than who never experienced an issue,

which is clearly reasonable.

However, results from the *chopit* model estimation show that the ordered probit model estimates may lead to misleading conclusions: 35-54 year old respondents are not more satisfied than the younger ones, they only apply different thresholds; the same remark applies to the residents of North-East Italy compared to the habitants of the North-West; not only the category "Other", but also self-employed individuals evaluate their satisfaction lower than employees, *ceteris paribus*. Self-employed respondents might use the online bank account not only for personal matters, but also to manage their own business. Consequently, they might need some additional options and demand a better quality from the service. On the other hand, individuals who belong to the class "Other" might encounter some extra difficulties when browsing on the bank account, since they might not have enough experience with the service, and thus they are not satisfied of the performance of the service. According to the *chopit* model estimates, the perceived level of the self-reported satisfaction is significantly related to a wider set of individual features with respect to the ordered probit model estimation.

Moreover, other interesting effects stand out, which may not visible according to an ordered probit specification. Indeed, by means of this comparison, we highlight the inadequacy of the previous approach to deal with the DIF, because some variables do not directly affect only the satisfaction level, but also the position of the cut-points. In particular, it is worth noticing that every variable significantly affects the response style, except *gender* and *area*, which only have an impact on the satisfaction. The coefficient γ^1 for people belonging to the class "35-54 years" is estimated positively. Hence, they tend to move the first threshold to the right, comparing to the other categories, thus making the "Very dissatisfied" category larger. As a consequence, they more likely rank themselves in this category, other things being equal. Probably, 35-54 year old people tend to use more the online banking, at home and also at work. They consequently might be more demanding and it might be more difficult to satisfy them. Concerning the variable *employment*, the second threshold of the last class shows a negative value, so, those who do not have a paid employment have less probability to be "Dissatisfied" with respect to the workers: students, unemployed, retired and housewives might make less use of the online bank account and might only perform basic operations. Thus, they might be easy to please, whereas an employed individual might be more dynamic and demanding, because he/she has to manage his/her salary, investments and other payments. The response style significantly changes according to the household size only when the members are two or three, otherwise the thresholds are basically equal, *ceteris paribus*. Respondents with higher levels of education tend to significantly move the first threshold to the right, widening the first category of "Very dissatisfied" and increasing the probability of that response. At the same time, they relocate τ^3 to the left and reduce the "Neither satisfied, nor dissatisfied" level. As for the employment, we expect interviewees with a higher education to be more demanding, so that it might be more difficult to totally satisfy them.

In both estimated models, the order of the questions (the *group* variable) plays an important role in explaining the reported level of satisfaction of the online banking service. However, given the aims of our work, we are going to discuss question order

and priming effects aside from the other variables, in a subsequent section.

3.1.2 Second wave sample

We now evaluate the effects of the same variables in the subsample of respondents who took part for the first time at the survey in September 2015 (second wave). The sample size is now equal to 510 units.

Since these respondents lack knowledge of the vignette tool, we expect a behaviour (in terms of the measurement of their own satisfaction) similar to the one of the respondents of the first wave. However, the descriptive statistics highlight some slight differences in the socio-economic composition of the two samples, especially with respect to gender and age. Indeed, the 57.1% of the sample of the first wave is composed by males, whereas this percentage reduces to 51.1% considering the new respondents of the second wave. In addition, the new respondents are older than the first-wave interviewees: the 32.5% of the first sample and the 26.2% of the new respondents belong to the "18-34 years" class; in addition, the percentage of people with 55 years or older is about 20.5% in the first wave and 26% in the second sample.

Table 4 shows the estimated coefficients of the *chopit* model for this sample. As in

Table 4: *Chopit* model estimates for new respondents of the second wave sample.

Variable		Self-assessment	Threshold equation coefficients			
			γ^1	γ^2	γ^3	γ^4
Gender		-0.264	-0.316	0.104	0.086	-0.081
Age	35-54 years	0.006	0.774**	-0.344*	-0.186	0.083
	55 years or older	0.431*	0.125	0.048	0.079	0.059
Area	North-East	0.448*	-0.026	0.154	0.069	-0.058
	Central	-0.184	0.196	-0.086	0.179	-0.279***
	South and Insular	0.188	-0.216	0.150	0.335**	-0.092
Employment	Self-employed	0.013	0.080	-0.272	0.149	0.055
	Other	-0.497**	0.036	-0.155	-0.002	0.024
Household size	2	-0.335	1.143**	-0.565*	-0.564**	0.488***
	3	-0.397	0.635	-0.248	-0.565***	0.475***
	4	-0.274	-0.065	0.164	-0.408**	0.436***
	5 or more	-0.342	0.511	-0.246	-0.534**	0.535***
Education	Medium	0.057	-0.291	0.049	0.840***	-0.080
	High	0.169	-0.203	0.066	0.451	0.159
Group		0.409**	0.036	0.046	-0.062	0.022
Problem		-0.760***	-0.652**	0.103	0.498***	-0.033
Constant			-5.324***	0.704*	-0.432	0.422**

Vigenttes	Coefficients
θ_1(Carlo)	-0.059
θ_2(Marina)	-2.602***

Log standard deviation	
Vignettes	0.428
Self-assessment	0.000
Thresholds	0.447

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

the *chopit* model analysed for the first wave sample, the thresholds significantly vary between individuals with different characteristics. It is worth noting there are some differences as well as some similarities with respect to the estimated model in Table 3. Both models show that age, area of residence, type of employment, question order and browsing problems directly affect the customer satisfaction, but with some slight differences. Here, residents of the North-East Italy are significantly more satisfied than the others; self-employed respondents do not differ from employees; males have the same satisfaction as females. As for the last four columns, we notice that area of residence, age, level of education and household size impact on the response style as in the previous model, but in a slightly different way. Anyway, most of the estimated coefficients are coherent with the direction of the effects emerged in the first model.

The most important result concerns the variable connected with the order of the questions: as in the previous model, belonging to a specific group directly influences the customer satisfaction. More in detail, the interviewees of group B are more satisfied than the individuals of group A, *ceteris paribus*. However, contrary to the estimates on wave 1 sample, for the new respondents of the second wave the *group* variable is never significant in any threshold equation: answering the anchoring vignettes before the self-evaluation increases the satisfaction, but does not affect the reporting scale. This behaviour is unexpected: wave-2 respondents had never read the vignettes before the interview, so, we would expect an impact of the question order on their response style as for the first wave respondents. However, in this case priming effects may be hidden by both the sample size (it is half with respect the first wave) and the features of the analysed sample. In other words, this difference on the group behaviour and the other dissimilarities are probably due to the number and the socio-demographic composition of the sample taken into consideration. It would be interesting to extract a sample from the first wave with the same features as the new respondents' one and compare the results of the models. If the outputs are similar, we would confirm that the lack of significance in the thresholds of *group* is just due to the composition of the data.

In order to compare this output with a model applied to an equal number of units, we select a sample from the first wave considering only the already-interviewed respondents of the second wave (541 units). Both samples (the new respondents and the already-interviewed respondents in the first wave) do not know the vignette instrument at the time of their interview, consequently, their condition regarding the questionnaire is the same. Table 5 shows the so estimated *chopit* model.

The already-interviewed respondents and the new ones have slightly different features and it is reasonable that also the outputs of the respective *chopit* models might be a little bit different. However, the most important thing to highlight is the effect of the *group* variable on the satisfaction. Indeed, considering two different samples, similar only in numbers, the point estimates of the self-assessment equation coefficient related to the question order (β_{group}) are essentially equal: 0.409 in the model for the new respondents and 0.418 in this last model. Yet, the associated levels of significance are the same, equal to 5%. This particular result shows that the *group* variable acts in the exact same way in the two samples, but in the new respondents the effect does not emerge in the thresholds because of some peculiar characteristics of the individuals. Moreover, it is interesting to notice the similarity

Table 5: *Chopit* model estimates for already-interviewed respondents in the first wave sample.

Variable		Self-assessment	Threshold equation coefficients γ^1	γ^2	γ^3	γ^4
Gender		-0.393**	-0.338	0.011	0.067	-0.042
Age	35-54 years	0.036	0.033	-0.038	-0.131	0.000
	55 years or older	0.375	0.337	-0.195	0.044	0.010
Area	North-East	-0.192	-0.118	0.048	0.216	-0.061
	Central	-0.196	-0.063	0.043	0.027	0.008
	South and Insular	-0.493**	-0.140	-0.075	0.223	-0.172**
Employment	Self-employed	-0.356	-0.056	-0.069	0.210	-0.186**
	Other	-0.552***	0.378	-0.486**	-0.021	0.008
Household size	2	-0.604*	0.242	-0.359	0.055	-0.167
	3	-0.269	-0.275	0.033	-0.109	-0.148
	4	-0.174	0.116	-0.035	-0.152	-0.143
	5 or more	0.010	-1.186	0.631**	-0.304	0.007
Education	Medium	0.304	0.558	0.004	-0.388**	0.082
	High	0.689**	0.602	0.184	-0.371*	0.234**
Group		0.418**	0.044	-0.141	0.025	0.202***
Problem		-0.893***	-0.134	-0.033	0.207	-0.001
Constant			-5.568***	0.717*	0.280	0.851***

Vigenttes	Coefficients
θ_1(Carlo)	-0.262
θ_2(Marina)	-2.712***

Log standard deviation	
Vignettes	0.456
Self-assessment	0.000
Thresholds	0.704

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

between this model and the one proposed in Table 3 regarding the effect of the *group* variable in the threshold equation. In particular, the fourth thresholds are statistically significant at a 1% level in both models and the respective values are similar.

We now join the first wave and the new respondents of the second wave and create a sample of 1527 individuals with different characteristics, but the same background in what concerns the questionnaire knowledge. All individuals have never completed the survey before and we want to detect if there is any difference in the satisfaction due to the different month of participation at the survey. Table 6 displays the results of a *chopit* model applied to this particular pooled sample. The model is specified as the previous ones, with only one extension: we add a dummy variable labelled as *wave 2*, which sorts the respondents of the first wave from the new respondents of the second wave and assumes value 1 if the unit took part at the second wave and 0 otherwise. The output is coherent with the previous models in what concerns the sign of the estimates; the variables, which directly affect the self-assessment, are mostly the same: *ceteris paribus*, women are less satisfied than men, employee

Table 6: *Chopit* model estimates for all respondents of the first wave and new respondents of the second wave sample.

Variable		Self-assessment	Threshold equation coefficients γ^1	γ^2	γ^3	γ^4
Gender		-0.408***	-0.269*	0.004	0.006	-0.062
Age	35-54 years	0.126	0.538***	-0.266***	-0.131	0.015
	55 years or older	0.355***	0.286	-0.178	0.069	0.017
Area	North-East	0.028	0.286	0.091	0.120	-0.010
	Central	-0.124	0.068	-0.062	0.056	-0.041
	South and Insular	-0.119	-0.122	-0.040	0.206**	-0.084*
Employment	Self-employed	-0.363***	-0.175	-0.060	0.098	-0.068
	Other	-0.487***	0.248	-0.339***	0.037	-0.011
Household size	2	-0.206	0.381	-0.198	-0.098	0.026
	3	-0.166	0.086	0.022	-0.312**	0.047
	4	-0.099	-0.017	0.077	-0.153	0.057
	5 or more	-0.225	0.082	0.032	-0.348**	0.092
Education	Medium	0.010	0.398	-0.159	0.033	-0.048
	High	0.249	0.518*	-0.070	-0.108	0.073
Group		0.253***	0.105	-0.082	-0.075	0.110***
Problem		-0.795***	-0.413**	0.079	0.278***	0.037
Wave 2		0.070	0.085	-0.059	0.048	0.009
Constant			-5.635***	0.907***	0.066	0.751***

Vigenttes	Coefficients
θ_1(Carlo)	-0.194
θ_2(Marina)	-2.615***

Log standard deviation	
Vignettes	0.414
Self-assessment	0.000
Thresholds	0.586

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

respondents are happier with the service than self-employed respondents and than those who have not a job, encountering a problem reduces the level of satisfaction, etc. Moreover, most of the variables have a direct impact on the response styles in different ways.

The estimated coefficients of the *wave 2* variable are never significantly different from zero, neither in the self-assessment equation, nor in the thresholds. Therefore, participating at the survey in May or in September is exactly the same thing: it does not have an impact neither on the response styles, nor on the general satisfaction. So, the decrease in the satisfaction in the second wave which appears from the descriptive statistics may be due to either differences in individual characteristics or reporting heterogeneity. However, people who are in the same condition of not knowing the vignettes have the same thresholds despite they answer the questionnaire in different times, as shown in Figure 11 (the small differences in the figure are not statistically significant).

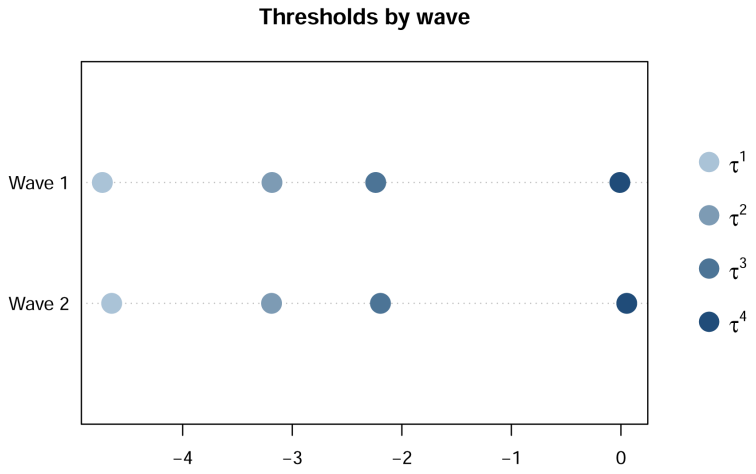

Figure 11: Estimated threshold values of the first wave and the new respondents of the second wave sample.

3.1.3 Discussion on priming effects

According to the estimates of the *group* variable from Table 3 to Table 6, the first interesting result is that question order effect is double: the individuals who answer the anchoring vignettes before the self-assessment (group B) change their response scale if compared to those who first read the self-evaluation (group A), and, in addition, their satisfaction increases. Concerning the threshold equation coefficients, the *group* variable is statistically significant in the last column: the question order affects the response style and, in particular, moves the last cut to the right. As a result, respondents of group B more likely rank themselves as "Satisfied" than "Very satisfied". Figure 12 shows the differences in the threshold values between the reference individual and a respondent with the same characteristics except the *group* covariate. Vignettes prime everybody to a more rational answer and a rarer use of the extreme categories, because the respondents compare their own situation and problems with the hypothetical situations which are presented: reading the scenarios described in the vignettes before the self-evaluation causes a deep reasoning in the respondent, which leads not only to a movement in the thresholds, but also to a higher satisfaction. Respondents might find themselves more satisfied than they thought before, checking over some examples dealing with problems they have never experienced.

This conclusion is supported also by the analysis on the pooled sample (first wave and new respondents of the second wave): answering the anchoring vignettes first and then the self-assessment question affects the respondents' definition of the extreme positive reported category ("Very satisfied"), reducing its area. On the other hand, the "Satisfied" area of group B is bigger than the one of group A, *ceteris paribus*.

Our findings provide evidence of the presence of *general priming effect*, showing that the questionnaire structure, in particular the order of the self-assessment and

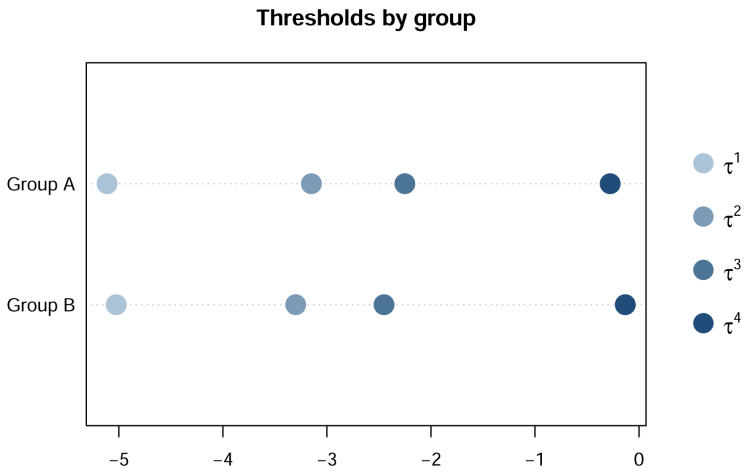

Figure 12: Estimated threshold values of the first wave sample between groups.

the vignettes questions, affects the reported level of satisfaction of the online banking service.

3.2 Study 2

In order to check the extent of long-term priming effects (hypothesis **H2**) on the reported customer satisfaction, in Figure 13 we display the self-assessment satisfaction of the service for the new and the already-interviewed sample of respondents of the second wave. It is worth noting that the histograms are very similar to the ones in Figure 6, in particular in what concerns the left plot. People who have never answered the questionnaire before and, so, never encountered the anchoring vignettes seem to have the same behaviour as the first wave respondents. On the other hand, already-interviewed respondents, who already answered to the first survey, might be no longer affected by the vignettes; indeed, their satisfaction shows a slightly different distribution with respect to the new respondents. This difference is more marked than in Figure 6 and may be caused by the memory effect regarding the instrument.

In September 2015, 541 units were interviewed for the second time. As highlighted by the empirical evidence, it seems that the response style of the already-interviewed respondents is not in influenced by the order of the questions. They have already answered to the same questionnaire about four months before and it is realistic to think that, after this short spell, they have not forgotten the presence of the vignettes tool. Therefore, we expect that the order of the questions does not affect their response scales, because of the presence of this memory effect.

Table 7 reports the estimated coefficients of the *chopit* model for the already-interviewed respondents of the second wave. Most of the explanatory variables significantly affect the response scale, but do not directly influence the satisfaction, like *household size* and *education*. Gender, age and browsing problems affect the self-evaluation: other things being equal, women are less satisfied than men at a 1% significance level; respondents older than 55 years evaluate themselves higher and

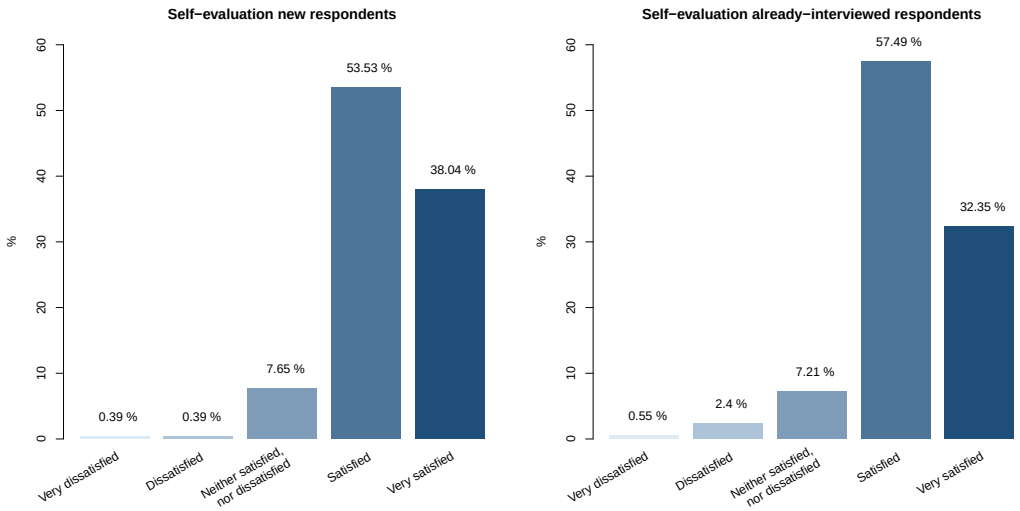

Figure 13: Distribution of the self-reported satisfaction between new and already-interviewed respondents of the second wave.

those who encountered a problem evaluate their satisfaction lower.

Focusing on the *group* variable, this coefficient is significant at 10% level, consistently with the previous estimated models: people who answer first the vignettes and then the self-assessment tend to be more satisfied, *ceteris paribus*. Concerning the thresholds, according to Figure 8 we expect that the coefficients should not be statistically significant and, so, that question order no longer affects the individual response scale. However, γ^2 and γ^3 estimates are statistically significant at 5% and 10% level, respectively. We thoroughly investigate this relationship and find that one particular variable causes the significance of *group*, that is *education*. Including an interaction between the question order and the level of education, labelled as *group*education* in the model, we obtain the result in Table 8.

Adding the interaction does not change the main results concerning the socio-demographic variables, but it affects the *education* and the *group* variable estimates. Overall, the higher satisfaction of the respondents of group B compared to group A no longer emerge. Thus, the second threshold of *group* is no longer significant. Instead, the third threshold is significantly affected not only by the question order, but also by the iteration between *group* and medium level of education. Concerning the iteration between *group* and the high level of education, it is not significant in the third threshold. However, the p-value is approximately 0.1 and, so, the null hypothesis is rejected for a while. This fact is probably due to the sample size (which is equal to 541 observations).

3.2.1 Discussion on priming effects

Figure 14 shows how the response scales change with respect to the question order and the level of education.

People with a high and medium level of education are not affected by the question

Table 7: *Chopit* model estimates for already-interviewed respondents of the second wave sample.

Variable		Self-assessment	Threshold equation coefficients			
			γ^1	γ^2	γ^3	γ^4
Gender		-0.553***	-0.645***	0.083	0.020	0.006
Age	35-54 years	-0.137	-0.006	0.0223	-0.236	-0.134*
	55 years or older	0.686***	-0.188	0.088	-0.133	0.112
Area	North-East	-0.350	0.030	-0.049	0.258	-0.045
	Central	-0.102	0.594*	-0.076	-0.012	-0.012
	South and Insular	-0.414**	0.310	-0.458***	0.220	-0.004
Employment	Self-employed	0.304	0.388	-0.002	-0.317	0.046
	Other	-0.171	-0.238	0.079	-0.079	0.089
Household size	2	-0.025	0.228	0.028	-0.077	-0.55
	3	-0.187	0.271	-0.252	0.094	-0.092
	4	-0.105	0.407	-0.012	-0.244	-0.201*
	5 or more	0.145	-0.402	0.273	-0.533	0.202
Education	Medium	0.369	-0.767**	0.682**	0.152	0.22
	High	0.412	-0.628*	0.814***	-0.140	0.125
Group		0.308*	-0.338	0.265**	-0.239*	0.090
Problem		-0.436**	0.138	0.050	-0.107	0.041
Constant			-3.791***	-0.159	0.133	0.838***

Vigenttes	Coefficients
θ_1(Carlo)	0.200
θ_2(Marina)	-2.457***

Log standard deviation	
Vignettes	0.369
Self-assessment	0.000
Thresholds	0.620

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

order, indeed, the slight difference in their thresholds is not statistically significant. On the other hand, the interviewees with a low educational background show a different position of the cut-points, in particular concerning the category "Neither satisfied, nor dissatisfied". Individuals with a high school or higher degree might better remember the questionnaire answered four months before (the effect of the anchoring vignettes does not vanish after this period). As they remember the structure of the previous survey, they tend to carefully reflect on their situation before answering the self-assessment question of the second wave, whatever the sequence of the questions is; for this reason, their response scales are no longer affected by the question order. This effect may be called as memory effect: during the participation at the survey for the second time, people remember the experience of the first questionnaire and answer accordingly. It would be interesting to learn how long this effect lasts and after how many months the already-interviewed respondents can be joined again with the new ones in order to apply the same response scale.

According to our results, less educated respondents tend to forget the presence of the anchoring vignettes and the memory effect is not strong, but still present. Indeed,

Table 8: *Chopit* model estimates for already-interviewed respondents of the second wave sample with interaction between group and education.

Variable		Self-assessment	Threshold equation coefficients			
			γ^1	γ^2	γ^3	γ^4
Gender		-0.540**	-0.681***	0.095	0.038	0.005
Age	35-54 years	-0.134	-0.014	0.040	-0.254	-0.128*
	55 years or older	0.681***	-0.194	0.104	-0.148	0.108
Area	North-East	-0.348	-0.029	-0.022	0.288	-0.070
	Central	-0.092	0.549*	-0.199	-0.018	-0.015
	South and Insular	-0.415**	0.304	-0.464***	0.263	-0.006
Employment	Self-employed	0.313	0.324	0.016	-0.266	0.045
	Other	-0.180	-0.221	0.079	-0.128	0.103
Household size	2	-0.036	-0.160	0.054	-0.036	-0.072
	3	-0.190	0.253	-0.241	0.111	-0.105
	4	-0.108	0.354	0.020	-0.235	-0.219*
	5 or more	0.153	-0.546	0.344	-0.452	0.165
Education	Medium	0.237	-0.157	0.448	-0.324	0.093
	High	0.274	-0.168	0.607	-0.461	0.033
Group		0.085	0.554	-0.062	-1.228**	0.101
Group*Education	Medium	0.243	-1.154	0.384	1.218**	-0.137
	High	0.247	-0.702	0.256	0.900	0.172
Problem		-0.447***	0.126	0.055	-0.100	0.044
Constant			-4.318***	0.010	0.454	0.842***

Vigenttes	Coefficients
θ_1(Carlo)	0.077
θ_2(Marina)	-2.572***

Log standard deviation	
Vignettes	0.368
Self-assessment	0.000
Thresholds	0.601

Note: *** p-value<0.01, ** p-value<0.05, * p-value<0.1

if we consider only people of the second wave with a low level of education and apply a *chopit* model adding a dummy variable which sorts the already-interviewed respondents from the new ones, this variable is still significant. Thus, although the already-interviewed respondents with a low educational background are still influenced by the question order, their thresholds differ from the new respondents' ones and a memory effect is present, though limited.

Our findings provide evidence of a partial presence of a *long-term priming effect*: the questionnaire structure (in particular, the order of the self-assessment and the vignettes questions) no longer affects the reported level of satisfaction of the online banking service for the middle and high educated people who have already experienced the vignette instrument in the past. This conclusion is not true for the low educated individuals who have already provided answers to some anchoring vignettes in a previous survey.

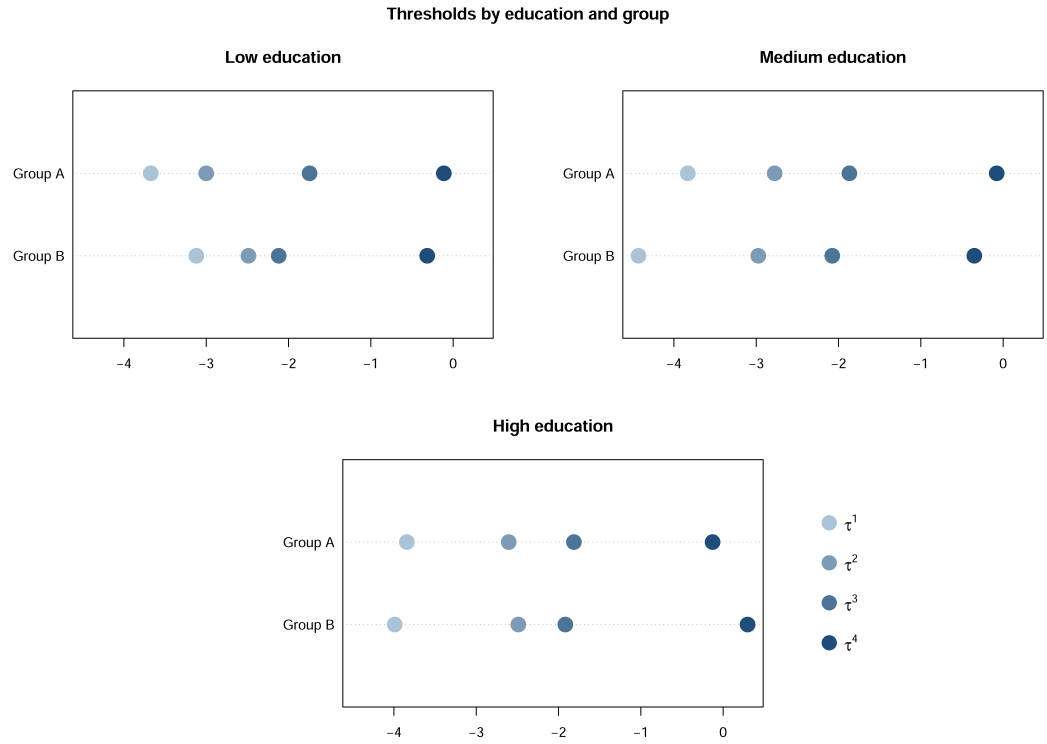

Figure 14: Estimated threshold values of the already-interviewed respondents of the second wave sample between education and group.

4 Conclusions

Our study suggests that the questionnaire structure is a crucial aspect of the research, since it affects the way people evaluate their satisfaction after the use of some online banking services.

Exploiting the instrument of the *anchoring vignettes*, we first provide evidence of the presence of several inter-personal differences in interpreting or using the response categories analysing the self-reported level of satisfaction.

Then, we apply the parametric solution of the anchoring vignette approach to show the presence of some priming effects: people who answered the self-assessment question after the vignettes (without a previous knowledge of this tool) apply different thresholds compared to the respondents who evaluate themselves first. Reading first the vignettes entails a more rational answer and a rare use of extreme categories, since the respondents evaluate more deeply their condition and compare it with the fictitious character's situation provided by the vignettes. Moreover, people may be more gratified of their own situation after reading the vignettes, maybe because they check over some examples dealing with problems they never had. The satisfaction of respondents who answer the self-assessment question after the vignettes is usually higher than the other people.

Priming effects are no longer present when middle and high educated respondents

are re-interviewed with the same questionnaire after a few months, because they remember the anchoring vignette questions asked in the previous wave; for low educated people some weak priming effects are still present, even if the vignettes are not addressed for the first time.

Our findings may also prompt a different application of the anchoring vignettes. While the original idea of this approach is to follow the same respondents over time and ask them the vignettes only during the first survey, we suggest an alternative use, especially evaluating surveys close in time: considering two different samples with the same characteristics which answer the questionnaire in two different periods, it may be possible to ask the vignettes only to the first sample and, then, apply the resulting thresholds also to the other one. Obviously, this idea should be tested, but the presence of a memory effect supports this alternative use. In this way, the benefits of the anchoring vignettes tool would increase: it would be enough to ask the vignettes only to one sample, then it would be possible to apply the same thresholds to every group of people who have never seen the questionnaire before, considering a certain time spell. In this way, the costs of the survey would be reduced and the interview would be shortened. This would not be plausible when the respondents already know the vignette tool.

References

Alexander, C. and H. Becker (1978). The use of vignettes in survey research. *Public Opinion Quarterly 42*(1), 93–104.

Angelini, V., D. Cavapozzi, L. Corazzini, and O. Paccagnella (2014). Do danes and italians rate life satisfaction in the same way? using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics 76*(5), 643–666.

Auspurg, K. and A. Jäckle (2017). First equals most important? order effects in vignette-based measurement. *Sociological Methods & Research 46*(3), 490–539.

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

Berry, M. and G. Linoff (1999). *Mastering Data Mining: The Art and Science of Customer Relationship*. John Wiley & Sons.

Buckle, J. (2008). Survey context effects in anchoring vignettes. Working Paper. Available from http://polmeth.wustl.edu/media/Paper/surveyartifacts.pdf.

Farris, P., N. Bendle, P. Pfeifer, and D. Reibstein (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance* (2nd ed.). Upper Saddle River, NJ: Pearson Education, Inc.

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.

Giese, J. L. and J. A. Cote (2000). Defining consumer satisfaction. *Academy of Marketing Science Review 1*, 1–27.

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. *Survey Research Methods 8*(1), 1–17.

Grol-Prokopczyk, H. (2017). In pursuit of anchoring vignettes that work: evaluating generality versus specificity in vignette texts. *The Journals of Gerontology: Series B 73*, 54–63.

Hill, N. and J. Alexander (2000). *Handbook of customer satisfaction and loyalty measurement*. Gower Publishing, Ltd.

Hoffmann, S. (2013). *Essays on the measurement of economic concepts in surveys*. Ph. D. thesis, Ludwig-Maximilians-Universität (LMU), München.

Holland, P. and H. Wainer (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Kapteyn, A., J. Smith, and A. van Soest (2007). Vignettes and self-reports of work disability in the united states and the netherlands. *The American Economic Review 97*(1), 461–473.

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

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

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

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

Oliver, R. (2010). *Satisfaction: A Behavioral Perspective on the Consumer* (2nd ed.). Routledge.

Paccagnella, O. (2011). A new tool for measuring customer satisfaction: the anchoring vignette approach. *Statistica Applicata - Italian Journal of Applied Statistics 23*(3), 425–442.

Paccagnella, O. (2013). Modelling individual heterogeneity in ordered choice models: Anchoring vignettes and the chopit model. *QdS - Journal of Methodological and Applied Statistics 5*, 69–94.

Paccagnella, O., M. Guidolin, G. Derboni, and T. Bago d'Uva (2015). The anchoring vignette approach to measure customer satisfaction. In *Proceedings of the meeting of the Italian Statistical Society "Statistics and Demography: the Legacy of Corrado Gini"*, Treviso.

Parasuraman, A., V. Zeithaml, and L. Berry (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, 41–50.

Parasuraman, A., V. Zeithaml, and L. Berry (1988). Servqual: a multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing 64*(1), 12–40.

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*, pp. 17–59. San Diego, CA: Academic Press.

Reinartz, W., M. Krafft, and W. Hoyer (2004). The customer relationship management process: Its measurement and impact on performance. *Journal of Marketing Research 41*(3), 293–305.

Tourangeau, R., L. Rips, and K. Rasinski (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

van Soest, A. and H. Voňková (2014). Testing the specification of parametric models by using anchoring vignettes. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 177*(1), 115–133.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy