Survey Research Methods (2023)
Vol. 17, No. 3, pp. 341-354
doi:10.18148/srm/2023.v17i3.7993
European Survey Research Association

Question Order and Panel Conditioning Analysing Self-Reported Data

Omar Paccagnella and Mariangela Guidolin
University of Padua, Department of Statistical Sciences

Self-ratings might not be directly comparable in socio-economic groups because of the presence of Differential Item Functioning (DIF) across individuals. Survey respondents might interpret, understand or use the response categories for the same question in different ways, due to different perceptions, attitudes, and so on. The instrument of the anchoring vignettes (based on additional questions to be collected in the survey) was introduced in the literature to compute response category DIF-corrected individual assessments. However, self-evaluations can be asked before or after the vignette questions. To take this type of reported heterogeneity into account, in this paper we apply the anchoring vignette approach to investigate the extent of priming effects (due to the order of the questions) and the role of the panel conditioning in measuring the customer satisfaction for some online banking services in Italy. The introduction of the anchoring vignettes induces some priming effects, that occur both in the perceived level of satisfaction and in the response tendencies. However, these question order effects move in two opposite directions that, in some cases, might lead to some forms of compensation. Differences are pointed out comparing respondents who already experienced the administration of the anchoring vignettes from those who answered the vignette questions for the first time.

Keywords: Anchoring vignettes; Differential Item Functioning; online banking; panel conditioning; priming; question order

1 Introduction

The question-order effect is a well-documented phenomenon in the literature: the order in which questions are presented to the respondents may affect their answers because earlier items can change the respondent interpretation of later questions (Tourangeau et al., 2000). Priming effects may be particularly striking when analysing self-assessment questions (McClendon & O'Brien, 1988). When self-reported evaluations are compared across different groups of respondents, individuals might interpret, understand or use the response categories for the same question in different ways, due to different perceptions, attitudes, or propensity to use them, particularly the extreme categories. This implies that self-assessments cannot be directly comparable across these groups. When the same evaluations are asked over time, individuals might recalibrate the measurement scale over time, providing variations in the observed level of the investigated variable, even if the construct of interest is conceptualised in the same way at all time occasions (Golembiewski et al., 1976).

King et al., 2004 introduced in the literature an appealing tool to realise measures free from the interpersonal incomparability resulting from the different uses of response scales, by means of the so-called *anchoring vignettes.* These are additional questions in a survey questionnaire where the experiences of some hypothetical individuals are depicted: adopting the same wording on the final question and the same set of answer categories, respondents are asked to evaluate both themselves and each scenario. When introduced in the literature, the anchoring vignettes were addressed just after the self-reported question, but Hopkins and King, 2010 supported an intentional use of priming of these questions, positioning them prior to the self-evaluation.

Few applications have so far studied the presence and the strength of priming effects working with anchoring vignettes, even less have investigated the extent of such effects longitudinally. The hypothesis is that the knowledge of the anchoring vignette tool, through participation in previous surveys, may weaken or even counteract any question order effects, because respondents may acquire a better understanding of the meaning of these questions due to the repetition of interviews. This hypothesis refers to the occurrence of a panel conditioning effect, the phenomenon under which individual reporting or knowledge changes by repeated participation in the panel survey (Warren & Halpern-Manners, 2012).

In this paper we exploit information collected by a survey that investigated the satisfaction for some online banking services in Italy, carried out in two waves, May and September 2015. The first sample of respondents is composed of people who took part in both waves, and we will term it as "longi-

Contact information: Omar Paccagnella, Department of Statistical Sciences, University of Padua, via Battisti 241 (E-Mail: omar.paccagnella@unipd.it)

tudinal component"; the second is made of the respondents who only took part in one wave, and we will term it "refresher component". Therefore, the purpose of this study is twofold.

First, measuring the extent of priming effects due to the order of the anchoring vignette questions in the above-mentioned dataset, taking the reporting heterogeneity problem into account, by comparing the two components of survey respondents. To do so, we use data from the second wave only, which contain both types of respondents, and try to understand whether reading the anchoring vignettes before or after the self-evaluation may imply different effects and response attitudes in both components. Our hypothesis is that a more instinctive answer (likely affected by the mood of the moment and the experience of recent problems) should be given when the self-assessment is firstly provided, while the opinion on the anchoring vignettes collected before self-rating may cause a more thoughtful answer (for instance, people may have the opportunity to compare their personal experience with the events represented by the anchoring vignette).

Second, Paccagnella, 2021 evaluated panel conditioning on the same dataset of our analysis, using only the longitudinal component. In our paper, we aim at investigating this issue from a different perspective, that is the role of panel conditioning taking priming effects into account, therefore comparing the behaviour of the refresher and the longitudinal components of the sample in a cross-sectional framework. As before, only data from the second wave of the survey will be used.

2 Background

Researchers are usually interested in investigating self-reported evaluations and comparing them between different groups of respondents. However, individuals might interpret, understand or use response categories for the same question in a different way, due to different perceptions, attitudes, optimism, or propensity to use them, particularly the extreme categories. The presence of such heterogeneity across individuals is known as *Differential Item Functioning* (DIF; Holland & Wainer, 1993) and threatens the comparability of their answers. When DIF occurs, people who are equal with respect to the (latent) trait of measurement do not have the similar probability of providing the same (observed) evaluation.

However, the response style, that is the individual tendency to respond to a survey question in certain ways regardless of the measured content, is only one reason of the DIF occurrence (Wetzel et al., 2013). In these cases, self-evaluations are not directly comparable, and relying on them when assessing subjective matters may be misleading. Briefly, the output of self-assessments can be seen as the sum of a real, but unobserved, evaluation and the DIF resulting from the different use of response categories: this type of DIF has to be removed in order to properly analyse individual ratings. This fact may occur when comparing people from different countries, but also when respondents are from the same country and have similar socio-economic conditions.

King et al., 2004 developed an alternative approach to deal with the DIF problem due to the response tendencies, then generalised by King and Wand, 2007: the experiences of some fictitious individuals are described in the questionnaire, the so-called anchoring vignettes, and respondents are asked to evaluate such situations by means of the same proposed categories for the self-assessment. In so doing, researchers have a reference —an anchor— to properly adjust self-evaluations. This method considers the individual heterogeneity by identifying the difference in the use of the response scale between respondents. Indeed, response category DIF appears in the proposed solution as threshold variation. Applications of this approach may be found in a growing number of papers and in different cross-sectional studies, from health status (Bago d'Uva et al., 2008) to job satisfaction (Kristensen & Johansson, 2008), from life satisfaction (Angelini et al., 2014) to marketing (Paccagnella, 2011). On the contrary, the literature on longitudinal anchoring vignettes is still limited (Angelini et al., 2011; Paccagnella, 2021).

2.1 Anchoring vignettes

Vignettes have a long history in investigating social phenomena 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 & Becker, 1978). In the anchoring vignettes introduced by King et al., 2004, the same scenario (without any differences) is proposed to all respondents, but different questions are addressed and respondents are then asked to evaluate these scenarios, as well as the own status in the same domain investigated by the anchoring vignettes. Even though there are some similarities, it is important to underline that anchoring vignettes and vignettes in factorial surveys are different issues. Individual variations in responses may be caused by the use of different cut-points between response categories. Vignettes provide an anchor scale, that adjusts individual self-evaluations: after correction, self-assessments may be compared across countries or socio-economic groups, because all subjective evaluations are reported to a common response category DIF-corrected scale. It is worth noting that the aim of the researcher is not designing DIF-free vignettes, rather writing vignette questions that are characterised by the same type of DIF as the self-ratings.

Two fundamental assumptions are needed for the validity of the anchoring vignette approach: *response consistency* and *vignette equivalence*. According to response consistency, for each individual response tendencies are assumed to be the same in self-assessments as well as in anchoring vignette evaluation. This implies that the same response category DIF is applied in answering to both the self-evaluation question and the anchoring vignette, allowing to correct the self-evaluation for interpersonal differences by using the vignettes as anchors. If the thresholds applied by each respondent changed between questions, it would be no longer possible to use the evaluation given to the anchoring vignettes as a standard. The vignette equivalence assumption states that all respondents perceive in the same way the underlying actual level of the variable described in any anchoring vignette. In other words, all interviewees agree on the real (unobserved) level for each anchoring vignette and place it at the same location on the latent scale. Therefore, the perception of each anchoring vignette does not depend on the individual characteristics: respondents apply their own DIF in choosing response categories, even if all understand the additional question in the same way. The vignette equivalence is required to obtain the measurement to be used as the anchor: it is possible to measure the thresholds of each respondent because the differences in the vignette evaluation are only caused by the response category DIF. This assumption would be violated if different respondents understood the anchoring vignette in different ways.

Figure 1 shows how anchoring vignettes can work. For instance, the self-reported health is the domain under investigation and the self-assessment question may ask: "In general, how would you rate your overall health?".

Some examples of anchoring vignette might be:

- "Tom has diabetes, and controls it by managing his diet. In general, how would you rate Tom's overall health?"
- "Karen has been diagnosed with high blood pressure. Her blood pressure goes up quickly if she feels under stress. Karen does not exercise much and is overweight. In general, how would you rate Karen's overall health?"

For all questions, the available answering categories are "(1) Very good", "(2) Good", "(3) Fair", "(4) Bad", and "(5) Very bad".

Two different individuals (to address these questions) have two different response scales: individual 1 turns her unobserved level (Y^*) of the domain of interest into the category "Fair", while individual 2 turns her latent level in the category "Good". According to the self-ratings, respondent 2 evaluates herself healthier than respondent 1, while the actual values lead to the opposite conclusion. At the same time, the perceived level of each anchoring vignette (Z_1^* and Z_2^*) is the same among the people, but the reported level can vary because of the response category DIF. Therefore, the (observed) answers to the anchoring vignettes (Z_1 and Z_2) provide the necessary information to measure individual DIF and then compute individual adjusted assessments, by the estimation of the cut-points for each respondent. Indeed, self-rating of individual 1 is in between of Z_1^* and Z_2^* values, whereas individual 2 evaluates herself unhealthier than Z_2^*. In other words, the aim of this approach is not defining a common threshold to be used to rescale all self-evaluations, but rescaling each self-assessment according to the own response cut-points.

The validity of the anchoring vignette assumptions has been criticised in the literature. Paccagnella, 2013 provides a review on these topics, while Grol-Prokopczyk, 2014 discusses some concrete recommendations to design anchoring vignettes that maximise vignette validity. However, anchoring vignette assumptions can be so far tested only using some additional assumptions or restrictions, particularly on the specification of the models used to analyse vignette data (van Soest et al., 2011; Greene et al., 2021).

2.2 Question-order effects

In surveys, the order in which questions are presented to the respondents may affect the answers in a more or less systematic way, because earlier items can change respondent interpretation of next questions (Moore, 2002; Schuman & Presser, 1996; Sudman et al., 1996; Tourangeau et al., 2000). Priming effects may be particularly striking analysing self-assessment questions (Garbarski et al., 2015; Golik, 2018; Lee & Schwarz, 2014; McClendon & O'Brien, 1988; Schwarz, 1999; Strack et al., 1988; Tourangeau et al., 1991). Preceding questions may also affect how respondents apply the response scales provided to them (Podsakoff et al., 2003; Sudman et al., 1996) and this is also strictly connected with the extreme response style (ERS) behaviour, that is the tendency of mainly selecting outer response categories regardless of the own opinion (Morren et al., 2011). In marketing, the analysis of the effects in positioning self-reported questions on a questionnaire has distant origins (Bradburn & Mason, 1964) and since then, it has been the topic of a growing amount of contributions measuring good or service quality (Auh et al., 2003; Bickart, 1992; DeMoranville & Bienstock, 2003; Schreier et al., 2018).

However, it is not usually easy to provide definitive recommendations on the most appropriate placing of the items, because many contextual factors have to be accounted for, as well as the main purposes of the analysis. The cognitive processes, involved when an individual formulates the answer to a question, are complex and not yet fully understood (Sudman et al., 1996), even if a general agreement on the series of processes that respondents go through answering questions can be found in the literature (Bradburn, 2004; Tourangeau et

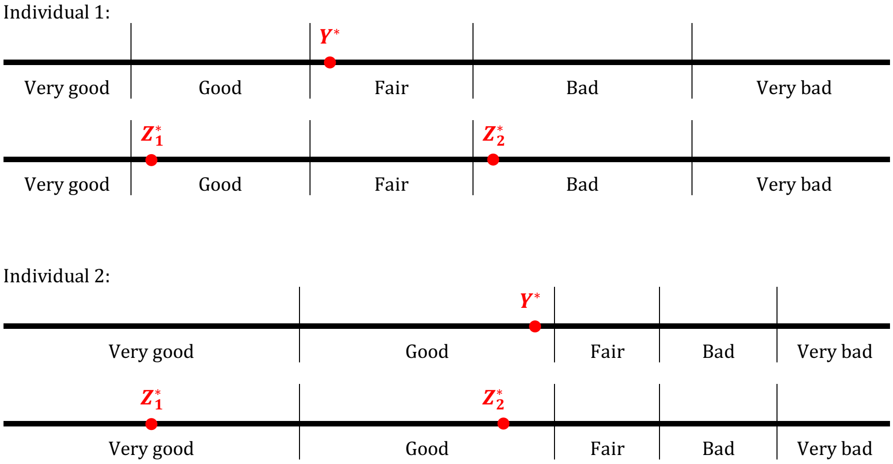

Figure 1

Example of reported heterogeneity between two individuals, comparing self-ratings and the evaluation of two anchoring vignettes. Y^* refers to the latent self-evaluation and Z_i^* refers to the latent value of anchoring vignette i.

al., 2000): i) comprehending the meaning of the question; ii) retrieving relevant information; iii) formulating an answer; iv) formatting and editing the answer.

When introduced in the literature, the anchoring vignettes were thought of as addressed just after the self-reported question. However, Hopkins and King, 2010 supported an intentional use of priming of the anchoring vignettes (i.e. the set of these questions immediately prior to 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”. Indeed, “when multiple judgments are made by a respondent using the same scale, respondents use their initial ratings to anchor the scale and thereby influence the scaling of their subsequent judgments” (Podsakoff et al., 2003). In other words, it is more likely that a respondent understands the idea in the same way as intended by the researcher when anchoring vignettes are heard just before answering the self-assessment question (Janiszewski & Wyer Jr, 2014).

In a survey experiment based on a small sample of German students, Hoffmann, 2013 did not confirm the beneficial effects of reversing the vignette 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 questions’ order may lead to different effects, for instance according to the context to which the anchoring vignette methodology is adopted or to question wording (Grol-Prokopczyk, 2018). Reasonably, priming effects may be stronger when people have low familiarity with the research topic under investigation by the questionnaire (Stefkovics & Kmetty, 2022). Hence, Buckley, 2008 claimed a complete randomisation of the order of all questions, that is anchoring vignettes and self-assessment questions together. Auspurg and Jäckle, 2017 showed 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).

Priming effects in longitudinal anchoring vignettes was investigated by Paccagnella, 2021, showing that the question order produces some priming effects only at the first wave[1], but not over time, when the time span between data collections was short (i.e. less than six months). In this context, panel conditioning can play a key role when individuals evaluate themselves, because the scenarios described in some questions might be remembered, regardless of their position inside the questionnaire.

2.3 Panel conditioning

Panel conditioning is a phenomenon that may affect the quality of survey data collected longitudinally. It refers to the fact that a respondent, who has already taken part in a survey previously, may provide a different answer to a ques-

[1] The order of the questions affects both the unobserved level of the analysed customer satisfaction (who read the anchoring vignettes before self-rating are more satisfied than people who read later the scenarios, ceteris paribus) and the individual response scales.

tion than the response would have given if she were taking part in the survey for the first time (Struminskaya & Bosnjak, 2021). The term *conditioning* is adopted because the current answers may be conditioned by previous participations in the survey. Indeed, this may help respondents in retrieving the knowledge and the information they need to provide their judgements in the current survey (Das et al., 2011; Kroh et al., 2016; Sun et al., 2019). Individual reporting or knowledge may be changed by the repeated participation in the panel survey (Bach & Eckman, 2019; Warren & Halpern-Manners, 2012). A comprehensive theoretical framework of this phenomenon was introduced by Bergmann and Barth, 2018.

According to their features, longitudinal anchoring vignettes refer to the form of panel conditioning called "changes-in-reporting" in the classification provided by Bach, 2021, because respondents can acquire a better understanding of the meaning of these questions repeating the interviews.

3 Data and Methods

3.1 Questionnaire

The data analysed in this paper were collected by means of a questionnaire investigating online banking services in Italy and the corresponding customer satisfaction. The questionnaire was carried out by a research team from the Department of Statistical Sciences at the University of Padua. The same questionnaire was administrated in May and September 2015.

The whole questionnaire is made up of 23 questions divided in three sections, that collect information on different features of the online banking customer experience[2]. The first section is screening, with the goal of collecting some information on the bank account allowing online operations, held by the respondent. A key question in this section asks for the types of services experienced in individual online operations. Then, the questionnaire focuses on the satisfaction related to some online banking operations and is divided in two sections, having both the same structure: people who navigate in the online bank account, for checking the account balance or movements, answer the first five questions; individuals who carry out operations, such as paying taxes, stamp duties, utilities, 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.

The focus of this paper is on browsing the main bank account only. The self-assessment question asks: *How satisfied are you with the easiness of the website navigation of your main bank account?* The available answering categories are: 1.Very Satisfied; 2.Satisfied; 3.Neither satisfied, nor dissatisfied; 4.Dissatisfied; 5.Very Dissatisfied.

Two anchoring vignettes are then proposed and both have to be evaluated by the respondents:

1. "Carl is an employee and has had an online bank account for three years. Every day he looks at the movements in his account, in order to check the presence of possible irregular movements. Carl 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. The list is loaded in a few seconds and Carl usually needs less than one minute to complete his control procedure."

2. "Marine is a housewife who checks the list of her family expenses with the credit card every three days, more or less. 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 centre in order to solve the problem. With the help of the operator, she is able to find the list of movements she is looking for."

After each description, respondents have to rate using the same response categories adopted for self-evaluation: "How satisfied is [Carl/Marine] with the easiness of the website navigation of [his/her] main bank account?"

Before asking any self-assessment question or anchoring vignettes evaluation, the sample is randomly divided in two groups that differ just in the question order. Respondents in Group A read the self-assessment question before the anchoring vignettes. This entails a more instinctive answer because respondents are likely to be affected by the mood of the moment and recent problems might be more relevant than past issues (Bower, 1981). People belonging to Group B provide their opinion on the anchoring vignettes before evaluating their personal experiences. The scenarios described in the anchoring vignettes make respondents ready for the self-assessment question and this may cause a more 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 events represented by the anchoring vignette (Podsakoff et al., 2003), as well as activate thoughts on the phenomenon of interest able to improve the interpretation of the question (Bradburn, 2004).

3.2 Sample

Data collection was carried out by a CAWI survey in two periods: May 2015 and September 2015. The project planned the collection of 1000 interviews at each wave, half

[2] The questionnaire is Italian. In this section we just report the translations of the most important questions for the analysis in this paper

of them as re-interviews in the second wave[3]. Only one member per household had to be interviewed. All datasets are available in a public data repository (Bassi et al., 2020).

Since no register or list of the population of interest exists in Italy, our sample was selected with a non-random procedure from a web panel that reproduces the profile of the Italian population holding a bank account (about 40,000 individuals), made available by the research institute that collected the data. Households were first screened in order to select people owning at least one bank account which allowed also online operations; then, they were interviewed until reaching the target number of interviews at each wave. In the final datasets, 1017 household members completed the questionnaire in May 2015, whereas 1051 individuals took part in the second wave: 538 already had answered the questionnaire in the first wave (longitudinal component), while the remaining 513 interviewees completed the whole questionnaire for the first time just in September 2015 (refresher component).

Our work analyses data only from the second wave of the survey.

3.3 Variables

The variable of interest is the individual satisfaction of the website navigation easiness of the own bank account, defined as an ordinal variable from 1 (the highest satisfaction) to 5 (the lowest satisfaction).

The set of individual covariates comprises demographic information (gender, age, living alone), socio-economic status (education, occupational status), area of residence and the experience of any problems browsing or managing the account. The survey did not collect detailed information on the respondents, such as incomes, wealth, cognitive abilities, social connectedness and so on.

To study in depth priming effects and panel conditioning, we also created four new dummy variables combining Group belonging (self-rating before or after the anchoring vignettes) and sample component (longitudinal or refresher) of wave 2. In the second wave, the self-evaluation question, the anchoring vignettes and the Group assignment of each respondent of the longitudinal component were exactly the same as wave 1 of the survey. For the wave 2 refresher sample, Groups A and B were randomly created as in the first wave.

3.4 Statistical solution

The parametric solution proposed by King et al., 2004 to exploit the anchoring vignettes data is called “chopit” (Compound Hierarchical Ordinal Probit) model, also labelled as “hopit”. It can be seen as a generalisation of the ordered probit model, as it basically consists of a joint estimation of some ordered probit models. As in ordered probit models, a latent variable is observed through an ordinal response variable, defined through some cut-points. While these thresholds are not allowed to vary across respondents in the standard ordered probit solution, in the chopit specification the anchoring vignette information is used to model the response category DIF through variations in the thresholds, which are functions of some individual characteristics. After the identification of response scales for each respondent, the self-assessment answers may be easily corrected.

The model adopted in this paper is the extension of the chopit model provided by Kapteyn et al., 2007. As every chopit specification, the model can be divided into two parts with a similar structure: the self-assessment component and the vignette component. Let X_i's be observed covariates, β the coefficients' vector and ε_i an independent and identically distributed random effect, independent of the set of exogenous variables. For model identification, the vector β does not include the constant and the unit variance of the ε error term is required. The noise ε_i includes reporting error and/or unobserved heterogeneity:

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

Respondent i is asked to turn her continuous perceived level Y_i^* into a reported category k ($k = 1, \dots, K$) by means of this criterion:

$$Y_i = k \quad \text{if } \tau_i^{k-1} < Y_i^* < \tau_i^k \tag{2}$$

where $-\infty = \tau_i^0 < \tau_i^1 < \dots < \tau_i^K = \infty$. Thresholds vary across units as functions of some exogenous variable V_i (which may overlap X_i) and a vector of parameter γ:

$$\tau_i^1 = \gamma^1 V_i + \eta_i$$
$$\tau_i^k = \tau_i^{k-1} + \exp(\gamma^k V_i) \quad k = 2, \dots, K-1 \tag{3}$$

where $\eta_i \sim N(0, \sigma_\eta^2)$, assumed to be independent of both V_i and all other error terms in the model; 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. Moreover, this solution models the thresholds with both a set of observed individual features and an unobserved individual heterogeneity term η_i, entailing that different anchoring vignettes' assessments might be correlated with each other. When σ_η^2 is null, the model is equal to the original chopit solution; van Soest and Voňková, 2014 showed that such extended approach is able to substantially reduce some misspecification problems of the original chopit specification.

Since model (1)–(3) is not identified, a vignette component is added to increase the information content. Each respondent i is characterised by one equation for each anchoring vignette answer. Let $\theta_j (j = 1, \dots, J)$ denote the actual level for the hypothetical person described in vignette

[3] Sample size at each wave was established according to project funding

j. According to the vignette equivalence assumption, it is perceived in the same way by all respondents, therefore:

$$Z^*_{ij} = \theta_j + u_{ij} \quad u_{ij} \sim N(0, \sigma^2_u) \tag{4}$$

where the error term u_{ij} is independent of ε_i and all individual covariates (X_i and V_i). Its variance is assumed to be the same across respondents and anchoring vignettes; however, it is possible to let σ^2_u vary over vignettes and their estimates can be seen as an indicator of how well each anchoring 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:

$$Z_{ij} = k \quad \text{if } \tau^{k-1}_i < Z^*_{ij} < \tau^k_i \tag{5}$$

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 the parameter θ_j does not vary across respondents, so differences in evaluating the anchoring vignettes are only a function of DIF. As a consequence of both assumptions, the anchoring vignettes allow to identify the type of response category DIF for each person and, consequently, to estimate individual thresholds. Then, with this information, adjusting the self-assessment is easy, as estimating β parameters.

Model (1)–(5) 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 a univariate ordered probit with varying thresholds, while the likelihood function for the vignette component is a J-variate one. Chopit models are estimated by using the *gllamm* (generalized linear latent and mixed models) procedure of the Stata software (Rabe-Hesketh et al., 2004). Estimation programs are available in the replication materials .

4 Results

4.1 Descriptive evidence

As described in Section 3, the analysis uses data only from the second wave of the survey and the total sample size is equal to 1051 respondents, of which 538 pertain to the longitudinal component, while the remaining 513 compose the refresher one. The proportion of Group A respondents is equal to 49% (50% in the longitudinal component only).

In the overall sample there is a slight majority of male respondents (55%) and the average age is close to 44 years: the oldest respondent is 85 years old, while the youngest is aged 18 years (respondents mainly belong to the classes 25–34 and 35–44 years). More than half of the respondents have a medium level of education, while only about 11% of them report a lower level of education. Most respondents (about 49%) are employed, while the proportion of self-employment is about 16%. Almost 30% of the customers experienced some forms of problems browsing or managing the account. Table A1 of the Online Appendix shows the main characteristics of the dataset according to the sample component and the question order. There are limited differences across these groups, the largest regards the proportion of respondents who experienced some problems browsing or managing the account in the refresher component and, for the longitudinal component, the percentages of people living in the North of Italy.

In general, respondents are satisfied with the service of their main bank account (Table 1): the distribution is right skewed and the categories “Very satisfied” and “Satisfied” include more than 90% of the sample. Overall, respondents of the refresher component seem more satisfied than those of the longitudinal one.

Interestingly, both in the refresher and in the longitudinal component, people who answered the self-evaluation *after reading the anchoring vignettes* seem to be more satisfied than those who rated themselves *before reading* them. In other words, respondents could find themselves more rewarded for their own condition after reading about other possible situations. The descriptive evidence also suggests that answering the anchoring vignettes after the self-rating seems to lead to less frequent use of the extreme positive category in favour of some intermediate positive answers.

Among longitudinal respondents, the distributions of the answers according to the order of the anchoring vignettes are much more similar each other than the distributions we may observe in the refresher component.

The analysis of the anchoring vignettes answers, reported in Table 2, shows how much Carl and Marine are perceived as satisfied with their online banking account. More than 90% of the interviewees assessed Carl’s condition as “Very satisfied” or “Satisfied”. On the other hand, respondents evaluated Marine’s scenario more negatively than Carl’s scenario, which is reasonable since she experienced a problem navigating through her bank account, while Carl’s vignette just described a standard situation. However, the Marine distribution of the longitudinal component is shifted toward unsatisfactory evaluation with respect to her distribution in the refresher component (the mode is “Satisfied” among the refresher component, but “Dissatisfied” among respondents who already knew the anchoring vignettes).

4.2 Chopit model: the role of individual characteristics

The extent of individual relationships, *priming effects*, and *panel conditioning* on the reported customer satisfaction are investigated by the estimation of the Kapteyn et al., 2007 ver-

Table 1

Distribution of the self-evaluations, by sample component and question order (%)

		Refresher component		Longitudinal component	
Self-assessment	Overall	Self-rating before AV	Self-rating after AV	Self-rating before AV	Self-rating after AV
Very Satisfied	35	35	41	31	34
Satisfied	56	54	54	56	59
Neither satisfied, nor dissatisfied	7	11	5	9	6
Dissatisfied	1	0	1	4	1
Very Dissatisfied	1	1	0	1	0
Total	100	100	100	100	100

Table 2

Distribution of the evaluation of the two anchoring vignettes, by sample component (%)

	Carl scenario		Marine scenario	
Assessment	Refresher component	Longitudinal component	Refresher component	Longitudinal component
Very Satisfied	45	40	9	7
Satisfied	46	51	31	26
Neither satisfied, nor dissatisfied	7	7	23	19
Dissatisfied	1	2	28	35
Very Dissatisfied	0	0	9	12
Total	100	100	100	100

sion of the chopit model, whose results are reported in Table 3.

Many variables in the thresholds are significantly different from zero, meaning that reporting styles vary according to some individual features. Wald tests to further check if the parameter estimates (except the intercept) are jointly different from zero are performed and reported in Table 4: we cannot reject the null hypothesis at 1% of significance level for three out of four thresholds. The estimated parameters of all thresholds are also jointly significant at 1% level (χ^2_{48} = 104.78, p = 0.00).

It is worth noting that some individual characteristics significantly affect only the perceived level of self-reported satisfaction, such as occupational status, others influence only thresholds, such as education, while some features are statistically significant both in the self-assessment and in the threshold equations (such as gender, age and reporting problems); area of residence is never statistically significant.

Individual characteristics affect the response tendencies in different ways and in different response categories. For instance, the first cut-off coefficient for female respondents is estimated positively. Hence, there is a tendency to move the first threshold to the right, compared to the other categories, thus making the “Very satisfied” category larger. Therefore, women are more likely to rank themselves in this category than men, other things being equal. Respondents with higher levels of education or people aged 50 or more tend to significantly move the first threshold to the left, narrowing the “Very satisfied” category, and the second threshold to the right, widening the “Satisfied” category and increasing the probability of providing this response. Reporting problems browsing or managing the bank account relocates the third cut-off to the right, hence widening the “Neither satisfied nor dissatisfied” category, ceteris paribus.

The perceived level of self-reported satisfaction is significantly related to a large set of individual features. The most important are gender, age, employment status, and reporting a problem: other things equal, women are less satisfied than men, respondents who had a problem in their experience are less satisfied than those who never experienced such issue and workers are more satisfied than people belonging to any of the other job classes (non-workers).

Looking at the vignette equation parameters, both estimates are significantly different from zero: the evaluation given to the Marina experience takes a higher value, hence a lower satisfaction, than Carlo’s scenario. This result

Table 3

Estimates of the chopit model

Variable	Self-assessment equation	Threshold equation				Vignette equation
		γ^1	γ^2	γ^3	γ^4	
Gender (female)	0.386***	0.338***	−0.037	0.104	−0.039	-
Age class: 50 or over years old	−0.475***	−0.357***	0.116**	−0.047	−0.002	-
High Education	−0.140	−0.376***	0.152***	−0.268**	0.281***	-
Living as a single	−0.121	0.255	−0.172**	0.218	0.055	-
Living in the North of Italy	−0.077	−0.026	0.005	−0.162	0.113	-
Living in the Central of Italy	0.047	0.096	−0.104	−0.140	−0.057	-
Being employee	−0.317**	−0.071	−0.077	0.083	0.015	-
Being self-employed	−0.461***	−0.157	−0.022	−0.031	−0.087	-
Having problems browsing/managing account	0.545***	−0.181	0.026	0.215**	0.098	-
Self-rating after AV-refresher component	−0.378**	−0.097	0.027	−0.060	−0.158	-
Self-rating before AV-longitudinal component	−0.105	−0.208	−0.008	−0.169	−0.100	-
Self-rating after AV-longitudinal component	−0.329**	−0.304**	0.093	−0.349**	0.036	-
Constant	-	−0.498***	0.750***	0.059	0.321*	-
θ_1 (Carl)	-	-	-	-	-	−0.609***
θ_2 (Marine)	-	-	-	-	-	1.843***
Variance	1.000	0.517				2.020

Log-likelihood = −3501.32, AV = Anchoring Vignettes
* $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

Table 4

Results from Wald test after the chopit model estimation

Significance test	Self-assessment equation	Threshold equation			
		γ^1	γ^2	γ^3	γ^4
All variables (except the constant)	-	41.95***	30.07***	25.78**	14.61
Refresher component: self-rating after vs before AV	5.83**	0.45	0.18	0.22	1.14
Longitudinal component: self-rating after vs before AV	2.22	0.46	2.90*	1.70	1.21
Self-rating before AV: longitudinal vs refresher component	0.47	2.08	0.02	1.62	0.50
Self-rating after AV: longitudinal vs refresher component	0.10	2.07	1.18	4.21**	2.03

AV = Anchoring Vignettes
* $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

strengthens the descriptive evidence.

4.3 Chopit model: the role of the question order

The order of the anchoring vignettes plays an important role in explaining the reported level of satisfaction with the online banking service.

Among people who read the anchoring vignettes for the first time, that is the refresher component, respondents who rated themselves after the vignette questions were more satisfied than those who rated themselves before them, *ceteris paribus*. A similar behaviour appears also for the longitudinal component, but the difference between "after and before" is not statistically significant (as reported in Table 4).

On the other hand, longitudinal respondents who evaluated themselves after the anchoring vignettes tend to shift the first cut-off to the left, meaning that they are less likely to rank themselves as "Very Satisfied" than respondents who did not know the anchoring vignettes, other things being equal.

4.4 Sensitivity analyses

Some sensitivity analyses were performed to check the strength of our results.

First, we estimated Model (1)–(5) using the original version of King et al., 2004, that is without the unobserved individual heterogeneity term in the threshold equations. Results are reported in Table A2 and A3 of the Online Appendix. Cut-offs may be accurately computed with a large number of individual characteristics specified in the threshold equations. The Kapteyn et al., 2007 version of the chopit model can be helpful when this range of covariates is limited (as in our case) or, more generally, in case of omitted variables. Estimates of the two models are similar, but the point estimates in the original model specification are smaller in magnitude than the ones in our study. The two sets of statistically significant variables are almost identical, even if in some cases (particularly in the self-assessment equation) with a significance level lower in the King and colleagues version than in the Kapteyn and colleagues model. The likelihood-ratio test supports the Kapteyn et al., 2007 version of the model.

Second, to check our findings on the priming effects due to the order of the anchoring vignette questions, we re-estimated the same model over two other samples of the same datasets: the refresher component in the second wave of the survey only and all refresher respondents from both waves. Estimated parameters related to the question order are shown in Table 5 (results from Table 3 are also reported for comparison); they are presented in terms of point estimates and confidence intervals at 95% of level. First of all, among all samples, findings are very similar and lead to the same conclusions: the order of the questions significantly influences the actual satisfaction, but not, or very weakly, the threshold responses. Moreover, when people experienced the anchoring vignettes for the first time, it does not matter if this happened during the first or the second wave.

5 Discussion

The estimates of the chopit model in Section 4 allow explaining how the response category DIF works when an individual provides his/her self-evaluation.

Analysing the satisfaction of the website navigation easiness of the own bank account, we may argue that the introduction of anchoring vignettes induces some priming effects, that occur both in the perceived level of satisfaction and in the response tendencies (regardless of the individual characteristics). However, these question order effects move in two opposite directions that, in some cases, might lead to some forms of compensation: in the end, the final result might hide different individual behaviours.

The perceived level of satisfaction of the individuals who answered the anchoring vignettes before self-reporting is higher than the one reported by those who first assessed themselves, *ceteris paribus*: this difference is statistically significant for the refresher component, but not for those respondents who knew the anchoring vignettes from the previous wave, that is those pertaining to the longitudinal component (Table 4, second column). Anchoring vignettes may prime everybody to a more thoughtful answer, because respondents compare their own situation and problems with the hypothetical situations which are presented. Respondents might find themselves more satisfied than expected, checking over some examples dealing with problems they have never experienced. Reading the scenarios described in the anchoring vignettes before self-evaluating causes deeper reasoning in the respondent's mind, which leads to movements on the perceived level of satisfaction.

Results from Table 4 (second column) also show that, once we control for the question order, the perceived level of satisfaction of the refresher individuals is not higher than the level of the longitudinal component.

The order of the questions does not affect the response tendencies in the refresher component, while it does in the longitudinal one only when the anchoring vignettes are asked before self-evaluations: their cut-offs move in a way that they are less likely to rank themselves in the first satisfied category than non-panel individuals, other things being equal, while differences with respect to panel respondents who read the vignette questions after self-ratings are not statistically significant, apart from a very limited effect in the second threshold (Table 4, from column three to six).

Table 4 also highlights the important role done by panel conditioning. Differently from the refresher component, priming effects due to the question order are not observed within the longitudinal component and this could be due by the fact that all of these longitudinal respondents have already experienced the anchoring vignettes during the previous wave of the survey (about four months earlier). Our conclusions on panel conditioning taking priming effects into account corroborate the results reported by Paccagnella, 2021, who estimated an extension of the chopit model (the so-called longitudinal chopit model) analysing only the longitudinal component of the dataset described in Section 3.2. He found that, among all parameters associated to the position of self-rating after the anchoring vignettes in the September wave, only the estimate in the second threshold equation was statistically significant at 10% level. The same conclusions are reached in our model looking at the Wald tests reported in third row of Table 4, even though the two models can be only partially compared.

According to all of these findings, differences between longitudinal and refresher components are larger when the anchoring vignettes are firstly addressed. For instance, we may define the profile of a male from the South of Italy younger than 50 years old, who does not live alone and is not employed, low educated and without experiencing any prob-

Table 5

Point estimates and confidence intervals at the 95% level of the parameters related to the question order of the refresher component in the chopit model, estimated over three different samples

	All wave 2 data (results from Table 3)			Wave 2 data of the refresher component			Wave 1 & 2 refresher component		
		95% C.I.			95% C.I.			95% C.I.	
Variable	Coef.	Lower	Upper	Coef.	Lower	Upper	Coef.	Lower	Upper
Self-assessment equation									
Self-rating after AV	−0.378	−0.685	−0.071	−0.391	−0.712	−0.070	−0.249	−0.431	−0.066
Wave 2	-	-	-	-	-	-	−0.068	−0.261	0.125
Cut-off 1 equation									
Self-rating after AV	−0.097	−0.380	0.187	−0.120	−0.413	0.173	−0.126	−0.299	0.047
Wave 2	-	-	-	-	-	-	−0.058	−0.242	0.125
Cut-off 2 equation									
Self-rating after AV	0.027	−0.097	0.150	0.031	−0.093	0.154	0.099	0.027	0.173
Wave 2	-	-	-	-	-	-	0.022	−0.055	0.099
Cut-off 3 equation									
Self-rating after AV	−0.060	−0.314	0.194	−0.070	−0.324	0.184	−0.066	−0.214	0.083
Wave 2	-	-	-	-	-	-	0.025	−0.134	0.184
Cut-off 4 equation									
Self-rating after AV	−0.158	−0.449	0.132	−0.172	−0.476	0.131	−0.149	−0.314	0.017
Wave 2	-	-	-	-	-	-	0.003	−0.176	0.181
N		1051			513			1524	

AV = Anchoring Vignettes

lems browsing the account. Based on the model estimates of Table 3, when the self-assessment is provided before the evaluation of the anchoring vignettes, we may estimate a probability of reporting "Very satisfied" with the service of his main bank account equal to 27% in the longitudinal case and 31% for the refresher one; when self-rating is asked after the vignettes' judgement, these percentages rise to 32% and 41% for the longitudinal and the refresher component respectively.

To some extent, reading the anchoring vignettes after self-reporting seems to imply a lesser use of the positive answer categories with respect to the opposite order. It could be interesting to investigate the reasons of such finding. A possible explanation may involve the ERS behaviour, as a result of an interaction between individual features and analysed item. As well-underlined by Morren et al., 2011, the extreme response style is "a characteristic of the respondent (a trait) indicating whether he or she tends to answer more extremely than other respondents in the investigated population. The degree to which this tendency actually appears in a particular rating scale depends on item characteristics such as response format, item content, location in the questionnaire, and so forth. Thus, some questions are more likely to elicit extreme response style than others". Differently from other standard approaches (for instance, the ordinal logit modelling), the chopit solution allows to define and estimate individual thresholds and evaluate the contribution of these respondent traits, as well as of the item characteristics (such as, in particular, the *location in the questionnaire* of the question) in forming the own response scale. Moreover, Klar et al., 2022 showed a strong and positive relationship between ERS scores and the "Openness to Experience" factor, which is a global personality trait in the Five Factor Model. Among some other characteristics, open people are receptive to new experiences and a deeper examination of their own thoughts, feelings, and values (McCrae, 1993).

In the end, the widths of the 95% confidence intervals reported in Table 5 suggest that it could be interesting to investigate these results also in terms of meaningful effects to highlight the substantive significance of the model estimates (Bernardi et al., 2017; Rainey, 2014). However, defining meaningful effect sizes among anchoring vignettes is complicated, because the question order plays a role both in the perceived level of satisfaction and in the response cut-points (and for the refresher component these effects move in opposite directions), but our results may provide an interesting starting point for future studies on this issue.

5.1 Limitations

The main limitation of our work lies in the nature of the analysed sample. By construction, we cannot have a probabilistic sample of our population. However, the criteria used to select the households and the similarities in the group characteristics and in many results comparing the samples of respondents in different waves (Paccagnella et al., 2018) support our belief in the good quality of our dataset.

The set of individual characteristics available for the analysis is not very large, but, as better explained below, we are investigating a homogeneous sample of Italian respondents.

Another critical issue is that in our studies we assume the validity of the vignette assumptions, i.e., response consistency and vignette equivalence, since we do not have additional information to be used for applying the solutions proposed in the literature for testing them. However, we do not compare respondents living in different countries, rather we consider a specific target population: individuals who reside in just one country and have a bank account that allows online operations. Our respondents are homogeneous according to many socio-economic conditions.

In the end, our results cannot be conclusive for a specific presentation order (anchoring vignettes before or after self-ratings), because our analyses provide evidence of priming effects due to the question order, but not which presentation order is more beneficial working with anchoring vignettes.

6 Conclusions

Exploiting the richness of information underlying the anchoring vignette data and the flexibility of the parametric solution to analyse them, we highlight how response category DIF works when an individual evaluates herself, providing evidence of priming effects and panel conditioning of the anchoring vignettes' questions measuring customer satisfaction towards the use of a service.

Even if positioning self-evaluations after the anchoring vignettes in a questionnaire cannot remove the DIF in the self-assessed item, this may be nevertheless helpful for respondents, because they enhance familiarity with the topic under investigation by the questionnaire. In turn, this improves the measurement of self-evaluation in practice, which is a very complicated task due to the multidimensional and unobservable structure of this phenomenon.

Acknowledgements

The authors are grateful to the Editor and two anonymous referees for their helpful comments and suggestions. They also thank Francesca Bassi and Chiara Basei for their comments on an earlier version of the manuscript. The authors gratefully acknowledge the support of the University of Padua project SID2016 ("Advances in Multilevel and Longitudinal Modelling").

References

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

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (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.

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

Auh, S., Salisbury, L. C., & Johnson, M. D. (2003). Order effects in customer satisfaction modelling. *Journal of Marketing Management*, *19*(3-4), 379–400.

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

Bach, R. L. (2021). A methodological framework for the analysis of panel conditioning effects. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement error in longitudinal data* (pp. 19–41). Oxford University Press.

Bach, R. L., & Eckman, S. (2019). Participating in a panel survey changes respondents' labour market behaviour. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *182*(1), 263–281.

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

Bassi, F., Guidolin, M., & Paccagnella, O. (2020). Online banking satisfaction in italy-obs project [Dataset]. https://doi.org/10.25430/researchdata.cab.unipd.it.00000356

Bergmann, M., & Barth, A. (2018). What was I thinking? A theoretical framework for analysing panel conditioning in attitudes and (response) behaviour. *International Journal of Social Research Methodology*, *21*(3), 333–345.

Bernardi, F., Chakhaia, L., & Leopold, L. (2017). Sing me a song with social significance: The (mis) use of statistical significance testing in European sociological research. *European Sociological Review*, *33*(1), 1–15.

Bickart, B. A. (1992). Question-order effects and brand evaluations: The moderating role of consumer knowledge. In N. Schwarz & S. Sudman (Eds.), *Context* *effects in social and psychological research* (pp. 63–79). Springer.

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*(2), 129–148.

Bradburn, N. M. (2004). Understanding the question-answer process. *Survey Methodology*, *30*(1), 5–15.

Bradburn, N. M., & Mason, W. M. (1964). The effect of question order on responses. *Journal of Marketing Research*, *1*(4), 57–61.

Buckley, J. (2008). *Survey context effects in anchoring vignettes*. http://polmeth.%20wustl.%20edu/media/Paper/surveyartifacts.%20pdf

Das, M., Toepoel, V., & van Soest, A. (2011). Nonparametric tests of panel conditioning and attrition bias in panel surveys. *Sociological Methods & Research*, *40*(1), 32–56.

DeMoranville, C. W., & Bienstock, C. C. (2003). Question order effects in measuring service quality. *International Journal of Research in Marketing*, *20*(3), 217–231.

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research*, *24*, 1443–1453.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science*, *12*(2), 133–157.

Golik, J. (2018). Testing question order effects of self-perception of risk propensity on simple lottery choices as measures of the actual risk propensity. *ASK. Research & Methods*, (27), 41–59.

Greene, W. H., Harris, M. N., Knott, R. J., & Rice, N. (2021). Specification and testing of hierarchical ordered response models with anchoring vignettes. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *184*(1), 31–64.

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

Hoffmann, S. (2013). *Essays on the measurement of economic concepts in surveys* (Doctoral dissertation). Ludwig-Maximilians-Universität (LMU). Müunchen.

Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum.

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

Janiszewski, C., & Wyer Jr, R. S. (2014). Content and process priming: A review. *Journal of Consumer Psychology*, *24*(1), 96–118.

Kapteyn, A., Smith, J. P., & van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, *97*(1), 461–473.

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

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

Klar, A., Costello, S. C., Sadusky, A., & Kraska, J. (2022). Personality, culture and extreme response style: A multilevel modelling analysis. *Journal of Research in Personality*, *101*, 104301.

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

Kroh, M., Winter, F., & Schupp, J. (2016). Using person-fit measures to assess the impact of panel conditioning on reliability. *Public Opinion Quarterly*, *80*(4), 914–942.

Lee, S., & Schwarz, N. (2014). Question context and priming meaning of health: effect on differences in self-rated health between Hispanics and non-Hispanic Whites. *American Journal of Public Health*, *104*(1), 179–185.

McClendon, M. J., & O'Brien, D. J. (1988). Question-order effects on the determinants of subjective well-being. *Public Opinion Quarterly*, *52*(3), 351–364.

McCrae, R. R. (1993). Openness to experience as a basic dimension of personality. *Imagination, Cognition and Personality*, *13*(1), 39–55.

Moore, D. W. (2002). Measuring new types of question-order effects: Additive and subtractive. *Public Opinion Quarterly*, *66*(1), 80–91.

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

Paccagnella, O. (2011). A new tool for measuring customer satisfaction: The anchoring vignette approach. *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*, *15*, 69–94.

Paccagnella, O. (2021). Self-evaluation, Differential Item Functioning, and longitudinal anchoring vignettes. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement error in longitudinal data* (pp. 289–309). Oxford University Press.

Paccagnella, O., Guidolin, M., & Basei, C. (2018). *Priming effects and customer satisfaction towards online banking services* [Working Paper Series n.1/2018].

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual* [Working Paper 160], Division of Biostatistics, University of California at Berkeley.

Rainey, C. (2014). Arguing for a negligible effect. *American Journal of Political Science*, *58*(4), 1083–1091.

Schreier, J. H., Biethahn, N., & Drewes, F. (2018). Question order effects in partial least squares path modelling: An empirical investigation. *Quality & Quantity*, *52*, 71–84.

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

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93.

Stefkovics, Á., & Kmetty, Z. (2022). A comparison of question order effects on item-by-item and grid formats: Visual layout matters. *Measurement Instruments for the Social Sciences*, *4*, 1–12.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*(5), 429–442.

Struminskaya, B., & Bosnjak, M. (2021). Panel conditioning: Types, causes, and empirical evidence of what we know so far. In P. Lynn (Ed.), *Advances in longitudinal survey methodology* (pp. 272–301). Wiley.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass.

Sun, H., Tourangeau, R., & Presser, S. (2019). Panel effects: Do the reports of panel respondents get better or worse over time? *Journal of Survey Statistics and Methodology*, *7*(4), 572–588.

Tourangeau, R., Rasinski, K. A., & Bradburn, N. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly*, *55*(2), 255–266.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(3), 575–595.

van Soest, A., & Voňková, H. (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.

Warren, J. R., & Halpern-Manners, A. (2012). Panel conditioning in longitudinal social science surveys. *Sociological Methods & Research*, *41*(4), 491–534.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, *34*(2), 69–81.