

Measuring the effect of questionnaire design on rating responses with the class of CUB models

Stefania Capecchi¹ · Romina Gambacorta² · Rosaria Simone¹ ·
Domenico Piccolo¹
¹University of Naples Federico II
²Bank of Italy

Responses to questionnaire items can be influenced by various factors including sample design, interview mode and/or how questions are phrased. To analyse these aspects, this paper draws on the Bank of Italy's surveys of households and firms, which employ different survey modes or questions with different phrasings, response options, or graphical features for sub-samples of respondents. We exploit the potential of CUB (Combination of Uniform Discrete and shifted Binomial random variables) modelling for the analysis of ordinal data. CUB models are able to capture and identify the different components of the cognitive process behind the responses and to study how these are related to the relevant covariates (such as respondents' characteristics). The results show that in general, although diverse survey modes and a different phrasing or graphical representation of questions may yield somewhat different findings in terms of uncertainty, responses to relevant questions such as those on reported satisfaction or expectations did not produce pronounced differences in data reliability.

Keywords: survey mode; CUB models; data quality; questionnaire design

1 Introduction

Data from households and firms' sample surveys represent the primary data collection method in microeconomic research and a relevant source of information for policymakers, increasingly used to assist economic policy choices. For this reason, the accuracy of survey data assumes a growing importance, and a large part of the literature investigated the presence of any feature capable of distorting the information collected (Batini et al. 2009, Bound et al. 2001, Braunsberger et al. 2007, DeCastellarnau 2018).

Answers to survey questionnaires may be influenced by many factors, leading in some cases to the collection of biased information. These effects may be sometimes extensive, especially when sensitive or complex information is collected, or, in other cases, related to specific segments of the population or subjects. Therefore, to optimize survey questions, it is necessary to implement effective measurement instruments, which are able not only to gauge the la-

tent trait under evaluation but also to adjust for the impact of the design features on empirical evidence and resulting interpretation.

To date, the literature has mainly focused on the influence of survey mode, questions' wording, and response format on measurement errors and non-response patterns (Bowling 2005; Dillman et al. 2009; Jäckle et al. 2015; Jäckle et al. 2010; Roberts et al. 2014; Vannieuwenhuyze et al. 2010) without clarifying the effects of these elements on the subjective cognitive process underlying the formation of the response.

The goal of this paper is to analyze how response choices may be influenced by three specific factors: mode of interview, visual formulation of the question (in the case of a self-administered questionnaire) and the presence or absence of the "don't know" option. These issues are particularly relevant when dealing with data collected with mixed-mode surveys or questionnaires administered in different countries or periods, using diverse survey modes, questions' wording or visual representations Vannieuwenhuyze 2013.

Our research goals will be addressed by resorting to the class of CUB models (D'Elia and Piccolo 2005; Piccolo

Corresponding author: Rosaria Simone, University of Naples Federico II, Naples, Italy (Email: rosaria.simone@unina.it)

2003), specified as a mixture of a (discrete) Uniform and a (shifted) Binomial distribution.

These models allow disentangling the respondent's inner *feeling* towards the item (which represents the intrinsic awareness of the respondent and that can be interpreted as the agreement towards the object) while accounting for factors contributing to the fuzziness of the response choice, including both subjective elements (personal indecision, lack of knowledge, interest or comprehension of the question, time dedicated to the choice, response styles, among others) and concomitant circumstances (presence of the interviewer and resulting social desirability bias, survey mode, type of response scales and categories, for instance). All these factors may bias the correct assessment of the genuine latent feeling expressed through the ordinal score, and thus they are assumed to concur in forming the *uncertainty* component of the rating choice under the selected CUB rationale.

Furthermore, CUB models enable in a simple way to study the effect of specific covariates on the aforementioned components, to investigate how such components are related to subjective characteristics and, as this paper would confirm, to modalities of the questions (included as *object's covariates*).

The empirical analysis draws from data collected by the Bank of Italy on households and firms, and in particular from the Survey on Households Income and Wealth (SHIW), the Survey on Italian Households (SHIW-I), the web survey on Italian households (WEBIT), and the Business Outlook Survey of Industrial and Service Firms (BOSIF). These data are particularly challenging for our analysis as they have collected with different survey modes and different formulations of the questions, including alternative graphical representation in a self-administered questionnaire and the inclusion or the exclusion of the "don't know" option for sub-samples. Notice that data are the results of representative samples obtained by nation-wide survey agencies with well-trained professional interviewers.

The paper is organized as follows. Section 2 provides an overview of the relevant literature on the factors affecting the cognitive response patterns to survey questions with a focus on the modality of responses. Section 3 introduces CUB models and the specific extensions needed to address our research questions. Section 4 briefly describes the Bank of Italy surveys on households and firms used in the empirical analysis, and discusses the results obtained. Section 5 concludes.

2 A review of the effect of questionnaire design

Due to the increasing use of survey data in supporting both economic research and policy decision, a large share of the literature has tried to identify all the causes of possible distortions in the collection of sample data to warn users of their potential effects on data accuracy (for an extensive review, see DeCastellarnau 2018). In this section, we report research on the three main sources of bias that may be influenced by decisions taken by the survey agency and related to the choices of (i) the *mode of questionnaire administration*, (ii) the *visualization of the categories of responses* and (iii) the *presence of don't know option*. These factors may affect respondents' behaviour in different ways, also in an interrelated manner.

A. Survey mode may induce differences in the way respondents are contacted or recruited, and thus affect their response rate. Furthermore, the vehicle of delivering the questionnaire may affect respondents' ability to focus and provide accurate answers.

The main methods of mode collections are:

- Paper-and-pencil interviewing (PAPI): the questionnaire is face-to-face administered by an interviewer using a traditional paper questionnaire.
- Computer-assisted personal interviewing (CAPI): the questionnaire is face-to-face administered by an interviewer using an electronic device (PC, tablet, mobile), which manages the questionnaire through a specifically designed program/application.
- Computer-assisted telephone interviewing (CATI): the interview is administered on the telephone. Interviewers insert responses directly into the computer where the questionnaire is managed by a specifically designed program.
- Computer-assisted web interviewing (CAWI): the questionnaire is self-administrated by the respondent using a web questionnaire specifically programmed.

Among these methods, PAPI is gradually falling into disuse as the other methods, based on the aid of a computer program, allow reducing considerably the errors associated with data entry, as the software can customize the flow of the questionnaire based on the answers provided and perform pre-established consistency or range checks. Although CAPI can be considered the survey mode which ensures the greatest data accuracy, it also embeds the greatest costs, due to the necessity for interviewers to physically reach respondents, while CATI and CAWI questionnaires can be administered remotely.

In particular, the CAWI mode has gained popularity in recent years due to its low cost, the high potentiality to reach

a global audience, the possibility to assist in the administration of the questionnaire with pictures, audio and video clips, links to different web pages, etc. Moreover, with regard to the specific topic of the research study, web-based surveys seem to be more effective especially when dealing with sensitive themes. Likely reasons may be related to the anonymity of the process leading to an increased level of willingness to answer truthfully, as well as to a lower feeling of stigmatisation Rhodes et al. 2003. This evidence seems to be more marked in the case of young people and young adults using mobile devices in answering online self-administered questionnaires Regmi et al. 2016. Using this survey mode, respondents may also choose to respond at their own convenience, thus, in principle, increasing their ability to focus on the topics covered Couper 2013.

The main distinction among survey modes is related to the presence of the interviewer. As a matter of fact, the interviewer can introduce distortions in the answers induced by his specific behaviour, such as directly or indirectly suggesting answers (interviewer effect), or simply because of his presence, leading the respondent to provide answers more in line with behaviours or opinions considered as convenient or adequate (social desirability bias). The effect of the interviewer may vary between face-to-face and telephone interviews and it can be linked to the specific socio-demographic characteristics of the interviewers with respect to those of the respondents Davis et al. 2009. As a consequence, the absence of the interviewer in the CAWI mode allows for the social desirability bias to be reduced, making this method more appropriate for sensitive questions (Kreuter et al. 2008; Montagni et al. 2019; Tourangeau and Yan 2007). More generally, respondents to CAWI surveys show a lower social desirability bias and tend to be more honest about their inner feelings due to the absence of the interviewer (Chang and Krosnick 2010; Sarracino et al. 2017; Tourangeau et al. 2020).

Empirical evidence also indicates that, in the self-administered mode, answers are more accurate and respondents show less *satisficing* behaviours Krosnick 1991, as they have more time to check for relevant information before answering the questions and can choose when to fill the questionnaire, in order to be more focused and less distracted Fricker et al. 2005. While self-administered surveys have the advantage of making respondents feel freer to express their opinion, they may, on the other hand, encounter problems, mainly related to the rate of participation and misunderstanding of the questions. The presence of a professional interviewer is, in fact, able to favour the active involvement of the respondent, clarify questions in case of doubts, motivate and avoid respondents from dropping the interview, and support the completion of the entire questionnaire.

B. Visual features of survey questions may affect respondents' choices in several ways (Maloshonok and Terentev 2016; Tourangeau et al. 2004), and the effects are mostly consistent even when varying the mode of survey administration (De Leeuw et al. 2011; Melani et al. 2008). The response order itself (positively/negatively oriented) may affect the selected category in several ways Keusch and Yang 2018. In general, in multiple choice questions, respondents are more prone to choose the first options available (primacy effect), especially when the options are positively ordered Tourangeau et al. 2013, for less educated respondents or for those who fill out the questionnaire most quickly Malhotra 2008. There is also evidence that some interviewees fill the first category that (approximately) matches his/her basic orientation towards the item. The use of different response formats such as radio buttons, slider scales, Likert scales, and text boxes, which requires a higher effort for respondents, may also influence the responses provided and the respondent's involvement. Slider user interface and text answer box may increase item non-response and dropout rate (Couper et al. 2001; Funke 2016). In scales, full labelling is in general preferred to labelling endpoints only De Leeuw et al. 2011 as the former adds further information, in principle reducing respondents' indecision. Furthermore, some response formats, such as the use of a grid in web surveys, may induce respondents to speed and straight-line their responses Chan and Conrad 2014.

C. Finally, the presence or absence of a "neutral" response category Kankaras and Capecchi 2025, as well as that of the "don't know/no answer" option (DK hereafter) can influence the information provided, and the effect on the quality of the response Velez and Ashworth 2007. For some Authors, respondents may yield no opinion or don't know answers if they are not able to understand the question Dillman 2002. In this perspective, the selection of "don't know" could be also regarded as a proxy for the respondent's lack of knowledge of the answer. According to other interpretations, this option, if "neutral", may be treated as a midpoint response on an ordinal scale Coombs and Coombs 1976 or merely as missing data. Furthermore, if, on the one hand, the presence of the "don't know" option could induce the respondent to adopt a *satisficing behaviour* Simon 1957, thus avoiding providing an adequate answer in the attempt to reduce its burden, especially in the case of a complex question, on the other hand, forcing the respondent to provide an answer when he couldn't, may introduce another source of errors. The pros and cons of using this option have been extensively discussed in the literature and the most widely reached conclusion is that the "don't know" choice should be used only for more complex topics and in general avoided in other cases Krosnick et al. 2002. The optimal approach can only be obtained through the help of

the interviewer that can be trained to maximize the benefit of the inclusion of this option. In particular, the interviewer should try to get the answer to the question, proposing the “no answer” option only when he/she has the impression that the respondent wouldn’t be able to provide useful information.

Summing up the aforementioned literature only found marginal differences between answers provided using alternative survey modes and question designs (Melani et al. 2008; Sarracino et al. 2017). Some evidence, for instance, emerged that individuals’ characteristics, such as age or education, may affect respondents’ cognitive skills (Chang and Krosnick 2010; Malhotra 2008). With respect to this state of the art, the application of CUB models hereafter proposed can provide added value to disentangle the effects of survey features on respondents’ inner feelings and, in particular, on the fuzziness of the rating choices when faced with different questionnaire designs. Thus, the approach can be useful in determining which survey characteristics affect the measurement of the actual signal the most, by generating a greater uncertainty around the signal.

3 The methodological setting

Ordinal evaluations are a very popular tool to survey questions since they allow us to assess the extent by which agreement, belief, satisfaction, etc. hold for respondents. In this regard, scholars involved in psychological, medical, social sciences and marketing disciplines have different approaches and introduced varied methodologies to tackle with responses which are substantially qualitative in nature.

3.1 Statistical models for rating data in brief

The literature on statistical methods and regression for ordinal data analysis follows two main research threads.

The apparently most consolidated ones, also for historical contingencies, adheres to the general construct of latent variables: these include generalized linear models, hierarchical models for longitudinal and nested responses, as well as structural equation modelling (see Agresti 2010; McCullagh and Nelder 1989; Skrondal and Rabe-Hesketh 2004; Tutz 2012 for general references). In this setting, cumulative link models McCullagh 1980 have gained over time the role of a benchmark statistical setting for regression of ordinal data, lending themselves to a variety of applications, and both computational and methodological developments. The paradigm is grounded on a regression construct over the latent continuous trait underlying a given ordinal response, with variants to account for category-specific effects of covariates, or dispersion and hierarchical

structures (see Tutz for a collection of recent results). With respect to psychometric studies, there has been a significant focus on Item Response Theory Hambleton 1991 to provide a unified framework to jointly measure person and item-specific effects of responses to multiple survey items, typically addressing a unique trait. The most popular and versatile specifications include the graded response model Samejima 1969 and the partial credit model Masters 1982, mostly used in the educational setting. Being specified on multiple ordinal responses, item response models are also a fruitful setting to investigate response styles, occurring when consistent response patterns are observed, independently of the actual content of the questions and of individual perceptions and opinions Van Vaerenbergh and Thomas 2012.

The second research thread involves directly the discrete support to parameterize relevant features of the ordinal distribution: two of the most versatile instances are provided by the Binomial and the Discretized Beta models (see (Scian-dra et al. 2024; Ursino and Gasparini 2018) and the references therein), as well as finite mixtures of discrete distributions to deal with multimodal responses (Simone 2022; Sur et al. 2015). Mixture modeling on the discrete support can be also exploited to operationalize a relevant amount of psychological literature that casts the paradigm that the response choice can be assumed to be a combination of perceptual aspects of the selection and of the uncertainty surrounding the choice due to non-contingent aspects, as the response support, the time dedicated to the answer and so on Tourangeau et al. 2000. According to this paradigm, an increasing number of statistical structures have been successfully introduced leading to the so-called *class of CUB MODELS*, a general mixture representation for rating data (Piccolo and Simone 2019a; Piccolo and Simone 2019b) which includes several parsimonious and flexible models able to assess both *feeling* and *uncertainty* of ordinal evaluations, where those components may eventually be functions of subjects’ and objects’ covariates. This setting forms the reference model for the forthcoming investigation: Sect. 3.2 is dedicated to dwell in its statistical background and frame it with respect to the research goals of the paper.

3.2 The reference model

Let us assume that a sample R_1, R_2, \dots, R_n consists of the ordinal responses of n interviewees, collected along with some subjects’ characterizations (socio-demographic, cultural and economic variables, for instance). For a given number of ordinal categories $m > 3$, a CUB model for the response R_i is defined as a Combination of a (shifted) Binomial distribution for *feeling* and a (discrete) Uniform dis-

tribution for *uncertainty*. Formally, $R_i \sim \text{CUB}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ if its probability distribution is specified by:

$$\Pr(R_i = r | \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y}_i, \mathbf{w}_i) = \pi_i b_r(\xi_i) + (1 - \pi_i) \frac{1}{m}, \quad (1)$$

$$r = 1, 2, \dots, m$$

for $i = 1, \dots, n$ and:

$$\text{logit}(1 - \pi_i) = -\beta_0 - \sum_{j=1}^p \beta_j y_{ij}; \quad \text{logit}(1 - \xi_i) = -\gamma_0 - \sum_{l=1}^q \gamma_l w_{il}. \quad (2)$$

The estimable parameters are $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)'$, whereas $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$, $\mathbf{w}_i = (w_{i1}, \dots, w_{iq})$ are row vectors of the covariates values for the i -th respondent. In (1), the probability mass of a (shifted) Binomial distribution at category r is denoted by: $b_r(\xi_i) = \binom{m-1}{r-1} \xi_i^{m-r} (1 - \xi_i)^{r-1}$, $r = 1, \dots, m$.

Among possible alternatives, the selection of such distributions obeys the criteria of parsimony and consistency: the Binomial accounts for the combinatorial alternatives faced by respondents when ordinal ratings have to be singled out, whereas the Uniform is the least informative distribution among the discrete ones with finite support. In addition, this choice for the uncertainty component yields model parsimony, since no estimable parameter is involved, and it bears an evocative interpretation of intrinsic fuzziness of the response, due to both subjective and contingent factors (see, also, Golia 2015 for a discussion on further interpretation of uncertainty parameters¹). The selected parametrization implies that $1 - \pi_i$ is the weight attached to the uncertainty distribution, and $1 - \xi_i$ is a measure of feeling towards the item in a positively oriented scale. Although several statistical implications are discussed in Piccolo and Simone 2019a and related references, a fundamental issue should be emphasized with regard to CUB models in order to corroborate its application finalized to our research goals: CUB specification implies that *the response of each individual* is a mixture between the expression of the substantial feeling (driven by $1 - \xi_i$), and an individual propensity to be influenced by contingent factors when marking a score to evaluate the item (conveyed by $1 - \pi_i$), letting the feeling assessment be more or less fuzzy accordingly. In other words, CUB prescribes that π_i quantifies the probability that an observed score is generated solely from the background overall feeling. Finally, among several admissible alternatives, the logistic link: $\text{logit}(z) = \log\left(\frac{z}{1-z}\right)$, for any $z \in (0, 1)$, is motivated by easiness and robustness reasons (Iannario et al. 2016; Iannario et al. 2017).

With respect to the classical approach to ordinal regression models (see (Agresti 2010; Tutz 2012), among others), a remarkable advantage of the class of CUB models is that subjects' covariates are an important qualification of model (1), yet they are not compulsory. Indeed, assuming a unique probability model for the respondents, a CUB model can be estimated for the ordinal response R , in which case $\pi = \pi_i$ determines the relative weight of the feeling component, or dually, its complement to 1 quantifies the heterogeneity of the distribution, whereas the probability parameter $1 - \xi = 1 - \xi_i$ summarizes the overall feeling gauged from the ordered evaluations. Although a full interpretation of parameters is related to the context of analysis (appreciation, evaluation, fear, worry, etc.), generally the feeling measure is simply associated to the location of the distribution (in terms of *modal value*), whereas uncertainty relates to heterogeneity in the responses (in terms of Gini heterogeneity index, see Capecchi and Iannario 2016). Beyond allowing a twofold interpretation of the response distribution, evidence is found that properly accounting for CUB uncertainty in preference models enhances also prediction performance Simone 2023.

Different survey items or groups of responses can be compared by plotting estimated feeling and uncertainty coefficients as points over the parameter space $(0, 1) \times [0, 1]$. This visual representation, firstly proposed in (D'Elia and Piccolo 2005; Piccolo and D'Elia 2008), offers distinctive insights of the CUB modelling approach since the possible effects of covariates (hereafter, the different modes of presentation of the questionnaire) may be immediately checked in size, direction and significance. This circumstance should not be underestimated since, in many instances, ordinal regression models may identify different significant covariates for different item, letting less straightforward the comparisons among respondents and items.

Further refinement of CUB models can be obtained to take into account the presence of an *inflated* category (Cordeas et al. 2009; Iannario 2012). A *shelter* category $s \in 1, \dots, m$ is a modality of the support of R_i which receives an upward bias of preference with respect to the expected response pattern. This effect can be accommodated in CUB model by specifying a degenerate distribution $D_r^{(s)} = I(R = s)$, with probability mass concentrated at $r = s$. Thus, the inflated CUB model becomes:

$$\Pr(R_i = r | \boldsymbol{\theta}) = \delta_i D_r^{(s)} + (1 - \delta_i) \left[\pi_i b_r(\xi_i) + (1 - \pi_i) \frac{1}{m} \right], \quad (3)$$

¹ More complex structures for explaining the uncertainty component have been proposed: for the sake of simplicity, these extensions will not be considered in this paper. For a broad discussion, see: (Piccolo and Simone 2019a, p. 391–393).

for $r = 1, 2, \dots, m$ and $i = 1, 2, \dots, n$. Here, $\delta_i \in [0, 1]$ so that $\text{logit}(\delta_i) = \mathbf{x}_i \boldsymbol{\nu}$ if relevant covariates \mathbf{x}_i are possibly modifying the shelter effect².

Then, the special case in which the mixing weight of Binomial distribution collapses to zero leads to a CUSH model, that is a Combination of a discrete Uniform and a Shelter effect Capecchi and Piccolo 2017. It is defined by:

$$\Pr(R_i = r | \boldsymbol{\theta}) = \delta_i D_r^{(s)} + (1 - \delta_i) \frac{1}{m}, \quad \delta_i \in [0, 1], \quad (4)$$

where $s \in \{1, \dots, m\}$ is the known location of the shelter effect.

Thus, exploiting the CUB paradigm, the objective of this paper will be pursued with the following approach: *ad-hoc* dummy variables will be included within the specification of CUB regression models with both subjects and objects' covariates on the grouped responses to:

- test and compare the effect of the mode of interview and/or the visual layout for the rating questions;
- check the effect of the presence/absence of DK option on the response support.

If necessary, CUB extensions will be considered under the same research scheme to encompass the presence of a shelter effect.

From the methodological point of view, it is worth to emphasize that CUB methods currently lack of a complete multivariate extension that would render possible its application and validation to consolidated scales in psychometric and educational studies. Some extensions to multivariate ordered ratings have been proposed in Ip and Wu 2024 using copula-based approaches assuming CUB margins, in Colombi and Giordano 2016 by assuming a mixture of multivariate uniform and Sarmanov distributions with CUB margins, and in Simone et al. 2020 in the setting of random effect models to account for subjective heterogeneity in response attitude: the latter proposal included a subject-specific random intercept in the regression link involving the uncertainty parameter π_i to account for the individual propensity to a meditated response choice following the feeling model.

3.3 Modelling the effects of survey modes and questionnaire features

CUB models with covariates have been broadly applied to identify response profiles in terms of subjects' characteristics (Capecchi and Piccolo 2017; Corduas et al. 2009; Fin et al. 2017) or objects' features (Capecchi et al. 2016; Piccolo and D'Elia 2008). With respect to the state of the art on the topics, we propose to resort to CUB models to identify if there is any significant difference in response features (feeling, uncertainty, possible shelter effects) among independent groups of responses corresponding to different survey modes or questionnaire features. This circumstance applies, for instance, if a given ordered evaluation is collected on independent groups of respondents (yet homogeneous with respect to relevant covariates) via different survey modes (CATI, CAPI, CAWI), various visual layouts (vertical, horizontal, etc.), or with different scales (with or without a don't know option, with different numbers of categories, labelled or not labelled categories, etc.). This task can be accomplished with the definition of suitable dummy variables identifying the different groups of responses, to be then included as covariates in a CUB model specification.

Then, the proposed method is analogous to the introduction of objects' covariates in the CUB statistical framework (Capecchi et al. 2016; Piccolo and D'Elia 2008), with the important difference that the requirement of independence of the response groups is strictly respected under our setting: thus, all the inferential results are enforced.

The general CUB model specification with covariates could be usefully applied also to determine if independent groups of respondents to the same ordinal evaluations express different feelings and/or uncertainty. Assume, for instance, that two independent and homogeneous samples:

$$\mathbf{R}^{(1)} = \left(R_1^{(1)}, \dots, R_{n_1}^{(1)} \right)'; \quad \mathbf{R}^{(2)} = \left(R_1^{(2)}, \dots, R_{n_2}^{(2)} \right)';$$

of ordinal evaluations are collected for the same survey item over a scale with m categories, but the survey has been administered in two different ways to the two samples (say, different survey modes or questionnaire features). Then, the two samples can be merged to derive a unique sample $\mathbf{R} = (\mathbf{R}^{(1)}, \mathbf{R}^{(2)})$ of $n_1 + n_2$ observations for the given survey item. Therefore, a dummy variable D_i can be defined to flag the two samples according to the way the survey has been administered, namely:

$$D_i = \begin{cases} 0, & \text{for } i = 1, \dots, n_1; \\ 1, & \text{for } i = n_1 + 1, \dots, n_1 + n_2. \end{cases}$$

² In the formal specification of model (3), s is known: as a matter of fact, it seems consistent to admit that a refuge category is well qualified in terms of location over the support or for its peculiar wording. If this assumption cannot be maintained, a sequential testing procedure for varying $s \in \{1, 2, \dots, m\}$ should be considered, with possible corrections for multiple testing (Benjamin and Hochberg 1995).

Specifying a CUB model with D_i explaining the possible effect on feeling and uncertainty

$$\text{logit}(1 - \pi_i) = -\beta_0 - \beta_1 D_i ; \quad \text{logit}(1 - \xi_i) = -\gamma_0 - \gamma_1 D_i, \quad (5)$$

or also on the shelter category, when present, with the extra specification of:

$$\text{logit}(\delta_i) = \nu_0 + \nu_1 D_i, \quad (6)$$

for $i = 1, \dots, n_1, n_1 + 1, \dots, n_1 + n_2$, could reveal if the chosen survey feature entails any difference in either uncertainty or feeling components of the rating response process or modifies the refugee attitude.

To this aim, it is sufficient to test the significance of β_1 and/or γ_1 according to classical likelihood-based inference. As to interpretation, the positive (resp. negative) sign of these parameters implies that the survey feature identified by $D_i = 1$ (namely, the one corresponding to $\mathbf{R}^{(2)}$), decreases (resp. increases) the corresponding uncertainty $1 - \pi_i$ and/or feeling $1 - \xi_i$, respectively. Eventually, by including more respondents' covariates in the model specification (5), possible interaction effects could be further tested. For instance, to check if covariate X has an effect on either feeling (γ_2) or uncertainty (β_2) of the response, as well as if there is any interaction effect of X with the survey feature identified by D_i , the following model (7) can be fitted to the observations drawn from the grouped sample \mathbf{R} :

$$\begin{aligned} \text{logit}(1 - \pi_i) &= -\beta_0 - \beta_1 D_i - \beta_2 X_i - \beta_3 D_i X_i ; \\ \text{logit}(1 - \xi_i) &= -\gamma_0 - \gamma_1 D_i - \gamma_2 X_i - \gamma_3 D_i X_i. \end{aligned} \quad (7)$$

Then, $-\beta_3$ and $-\gamma_3$ measure the contribution that the combination of covariate value X_i and survey mode D_i induces on uncertainty $1 - \pi_i$ and feeling $1 - \xi_i$ (on the logit scale), respectively, with respect to the linear effect of both covariate and survey mode/design feature.

Analogously, to analyse the direct effect of covariate X on the shelter category or its interaction with the survey feature identified by D_i , we use the following model:

$$\text{logit}(\delta_i) = \nu_0 + \nu_1 X_i + \nu_2 D_i + \nu_3 D_i X_i, \quad (8)$$

so that ν_3 measures the contribution that the combination of covariate value X_i and survey mode D_i induces on the probability to shelter δ_i with respect to the linear effect of both covariate and survey mode/design feature (on the logit scale).

For CUB models, parameters can be effectively estimated by maximum likelihood (ML) methods using available software. A devoted library for the R environment is available on the official CRAN repository (Iannario et al. 2024). Dedicated libraries for CUB models are available

also for Gretl (Simone et al. 2019), and for STATA (Cerulli 2020; Cerulli et al. 2022). An accelerated version of the EM algorithm and the corresponding implementation of best-subset variable selection is available within the R library (Simone 2020), also on CRAN. To address the research goals of the paper, the CUB estimation procedure has been adapted to account for sample weights: the code is available as online supplementary material.

4 Empirical evidence from the bank of Italy's surveys

To check for the usefulness of the proposed approach three datasets will be considered: the first case study concerns measurements of the perceived value of future inheritance collected within the 2016 questionnaire of the Survey on Households Income and Wealth (SHIW) in order to verify the possible *modifications induced by the presence/absence of a DK option* in the response support (see Sect. 4.1). Then, two case studies based on the WEBIT/SHIW-I survey on households (Sect. 4.2) and the BOSISF survey on enterprises (Sect. 4.3) will be discussed to investigate *the effects of different survey modes* in the cognitive response process. The WEBIT survey will be used also to test for *the effects of different visual representations* in self-administered questionnaires.

4.1 Estimating the DK effect on SHIW dataset

To analyse the effect of the introduction of the “don't know-no answer” option we gather from the dataset which stems from the 2016 edition of the Survey of Households Income and Wealth (SHIW). The survey, conducted periodically by the Bank of Italy since 1962, collects information about the economic conditions of Italian households both with respect to real and financial assets held and to their sources of income together with a complete set of information about the socio-demographic characteristics of each of the family members (Baffigi et al. 2016; Bank of Italy. 2018). Data are collected from professional interviewers specifically trained using the CAPI method.

The SHIW adopts a two-stage stratified sampling design. Provided weights adjust for unequal selection probability and non-response, account for the correlation in the panel component and are post-stratified to external information about the socio-demographic characteristics of the reference population³. In order to deal with the complex survey design, without sharing respondents' characteristics

³ See Faiella and Gambacorta (2007) for detailed information about the weighting process in the SHIW and the resulting efficiency of weighted sampling estimates.

Table 1

Distribution of response options for expected value or real assets future inheritance

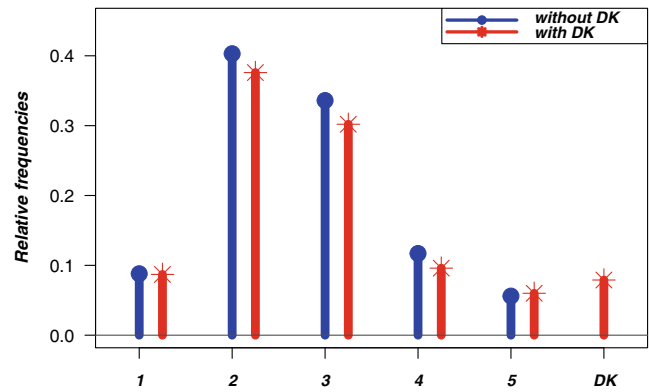
Categories	Absolute frequencies		Relative frequencies		
	Without DK	With DK	Without DK	With DK	$f^{(reduced)}$
(1) up to 50,000 euros	60	55	0.088	0.087	0.094
(2) from 50,000 to 150,000 euros	273	239	0.403	0.376	0.408
(3) from 150,000 to 300,000 euros	228	192	0.336	0.302	0.328
(4) from 300,000 to 500,000 euros	79	61	0.117	0.096	0.104
(5) over 500,000 euros	38	38	0.056	0.060	0.066
(6) Don't know (DK)	–	50	–	0.079	–
Total	678	635	1.000	1.000	1.000
Laakso and Taagepera index	–	–	0.584	0.572	0.589

protected by privacy (such as the stratum they belong to), replication weights are disseminated with data.

Among several questions listed in the questionnaire submitted in the 2016 survey, the main interest of this paper concern the item related to the interviewee's opinion about the global monetary value of the parent's house, on December 31, 2016. More specifically in the 2016 questionnaire, a section was dedicated to inspecting the value of future inheritance, asking first for the number of dwellings owned by parents not living in the households and then for an estimate of their value. Due to the potential difficulty on the part of the respondents to provide an answer to this question, amounts have been expressed using ordinal categories. Furthermore, the “don't know-no answer” option was randomly inserted for half of the sample (leading to two formulations of the question, D50a and D50b), allowing us to test whether only those who were not aware of the phenomenon or even those who adopted a satisficing behaviour, made use of it. In fact, to limit this latter conduct, interviewers were trained not to explicitly read this option even when available.

Both the formulations of the question are expressed with the same wording and $m=5$ ordinal categories are offered for an orderly evaluation; the only difference consists in the absence (Question D50a) and presence (Question D50b), respectively, of a sixth possibility of response, denoted as “I don't know/I don't remember”, which will be simply denoted by DK.

Can you give me even a rough estimate of the total value of these properties on 31/12/2016? Choose one of the ranges listed below:

**Fig. 1**

Frequency distribution for responses without (blue, points) and with (red, stars) DK option

- up to 50,000 euros ... 1
- from 50,000 to 150,000 euros ... 2
- from 150,000 to 300,000 euros ... 3
- from 300,000 to 500,000 euros ... 4
- over 500,000 euros ... 5
- Don't know ... 6

The statistical problem is to measure the significance of the effect of option DK on the expressed evaluations and if this situation is different with respect to definite clusters (with respect to gender, age, family composition, marital status, geographical area, income, financial education, etc.). Thus, the sample has been randomly split into two groups—statistically equivalent with respect to the main democio-economic variables—which consist, respectively, of 678 interviewees who received a questionnaire with Question D50a (without DK) and 635 interviewees who received a questionnaire with Question D50b (with DK).

Table 1 and Figure 1 show the different frequency distributions among the groups. The normalized Laakso and Taagepera index⁴ confirms a substantially equal heterogeneity for the distributions of the two sub-samples. Furthermore, the DK option was selected approximately only by 8% of respondents,⁵ therefore the distributions on the ordered support are quite similar, except for categories 2, 3, 4

⁴ Laakso and Taagepera index is a monotonic transformation of the Gini heterogeneity index and it has the advantage to better discriminate in the comparison of frequency distributions. More details in Gapecchi and Piccolo (2017).

⁵ In Manisera and Zuccolotto (2014), an extension to CUB model is introduced to measure the impact of the presence of a DK option on the heterogeneity of the distribution (see also Fin et al. (2017) for more applications). For responses to D50b, the circumstance that only 50 interviewees have selected the DK option limits the possibility to test this approach.

Table 2

Parameter estimates and corresponding standard errors (in parentheses) for Model (5) and Model (6) (with shelter at second or third category)

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\nu}_0$	$\hat{\nu}_1$
Model (5)	2.101 (0.526)	-0.820 (0.654)	0.417 (0.062)	-0.033 (0.107)	-	-
Model (6, she(2))	1.900 (0.527)	-0.807 (0.721)	0.350 (0.078)	-0.059 (0.129)	-2.377 (0.584)	0.046 (0.806)
Model (6, she(3))	1.763 (0.456)	-0.781 (0.530)	0.469 (0.077)	-0.015 (0.136)	-2.791 (0.826)	0.156 (1.305)

whose relative frequencies reduce in the presence of *DK* (see Fig. 1). However, after removing the *DK* responses, the (relative) frequency distribution of the observed rating variable, denoted as $f^{(reduced)}$ for convenience and reported as the last column in Table 1, appears more similar to the distribution of responses to Question D50a.⁶

Thus, according to the exploratory evidence, the presence/absence of a *DK* option has no relevant impact on the rating distribution. To verify this statement with suitable statistical models introduced in paragraph Sect. 3.3, Table 2 reports the estimation results for the model (5). Standard errors of parameters are obtained via replication weights using JRR (Jackknife Repeated Replication). Results indicate that the presence of a ‘*don’t know*’ response option does not significantly modify either the uncertainty or the actual feeling of the observed scores.

Since a certain amount of inflation in frequency at the second and at the third response options is observed with respect to CUB fit, a CUB model with shelter at each of these categories was tested to check if the presence of the “*don’t know*” option in the response scale significantly modifies the refuge attitude.

Results for model (6) are reported in Table 2 as well, showing that no significant effect is found.

Thus, it may be safely inferred that the presence of a *DK* option with the rating scale does not modify the way the rating options are perceived and used.

Model (7) can be considered to check if there is any covariate effect and interaction with the presence of a *DK* option in the rating scale. Accordingly, Table 3 reports the Wald statistics for the test of significance of the corresponding effects. Aside from identifying effects provided

⁶ This evidence has been numerically confirmed by computing the normalized *dissimilarity index*, to compare the relative frequency distributions of D50a and D50b also considering the reduced distribution.

Table 3

Wald statistics for parameters of model (7)

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Has children	2.579	-0.700	-0.569	0.194
Gender*	2.796	-1.427	-0.056	0.773
Income	2.219	6.304*	-0.396	-29.632*
Northern Italy	2.216	-1.760	-1.904	1.661
Central Italy	3.412	0.089	3.760*	-3.074*
University Degree	4.676	0.254	0.232	-0.154
Home ownership	2.950	0.323	0.932	-1.144
Other real estate	0.580	-0.437	-0.502	0.452
	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
Has children	4.534	-0.906	-0.766	0.873
Gender	5.009	-0.207	-1.174	0.131
Income	7.531	-0.568	-5.084*	-0.581
Northern Italy	7.859	-0.640	-1.118	0.419
Central Italy	6.521	0.244	-3.149*	-1.643
University Degree	9.781	0.012	-5.120*	-1.446
Home ownership	3.881	-0.832	-1.456	0.699
Other real estate	6.273	-0.107	0.022	-0.500

* Male = 0, Female = 1.

*p < 0.05

by increasing income, having a university degree and living in central Italy, all of which act by increasing the feeling $1 - \xi_i$ of the respondents (i.e., they report on average higher values for parents’ dwellings), interaction effects between subjects’ characteristics and presence of *DK* option that are relevant for our analysis are found only with respect to the uncertainty component. Specifically, *ceteris paribus* for a fixed level of income, heterogeneity is lower if the *DK* option is present (D50b) than if there is no *DK* option (D50a). In addition, for D50a there is no significant income effect on the heterogeneity of the distribution, whereas heterogeneity of D50b increases with income. Responses of individuals from Central Italy are significantly less heterogeneous than those of the rest of the respondents, to a greater extent for respondents to D50a than for respondents to D50b. It is worth stressing that these results are only revealed by the

Table 4

Distribution of response options for the subjective economic condition between CAPI and CAWI surveys

Categories	Absolute frequencies		Relative frequencies	
	CAWI	CAPI	CAWI	CAPI
(1) with great difficulty	115	359	0.136	0.181
(2) with difficulty	90	277	0.106	0.140
(3) with some difficulty	207	522	0.245	0.264
(4) fairly easily	237	576	0.280	0.291
(5) easily	122	190	0.144	0.096
(6) very easily	75	54	0.089	0.028
Total	846	1978	1.000	1.000
Laakso-Taagepera Index	–	–	0.816	0.723

use of the specific extension of CUB models derived for this analysis.

4.2 Modelling the effects of CAWI and CAPI survey modes and visual representations on WEBIT/SHIW-I dataset

To investigate the effect of the survey mode for households, we gather information from the Web Survey on Italian Households (WEBIT) and the Intermediate Survey in Italian Households (SHIW-I) administered using the CAWI and the CAPI modes, respectively. The two surveys have been conducted in parallel, between two editions of the SHIW, with a shorter questionnaire containing mainly qualitative items. The WEBIT has been managed in collaboration by the Bank of Italy and ISTAT (the Italian National Institute of Statistics) to investigate the use of web surveys in collecting data on households income and wealth on a probabilistic sample of about 1000 individuals (Barcaroli et al. 2019). At the same time, the SHIW-I was carried out by the Bank of Italy using the traditional CAPI mode on a sample of about 2000 households selected from those who had participated in the 2014 edition of the SHIW. To make the two surveys as much comparable as possible, participants to both surveys were drawn from the population of the same municipalities, with a similar two stage sample design and the questionnaires were designed to contain common questions and the same information about respondents socio-demographic characteristics and their economic conditions that could be used as covariates in the analysis (Gambacorta et al. 2018).

To compare answers using different survey modes, we refer to the question regarding the subjective perception of

Table 5

Horizontal response options to question: “Is your household’s income sufficient to see you through to the end of the month ...?”

1	2	3	4	5	6
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
With great difficulty					Very easily

the economic condition of the household. The question was present in both questionnaires adopting the same wording as follows:

Is your household’s income sufficient to see you through to the end of the month ... ?

- with great difficulty ... 1
- with difficulty ... 2
- with some difficulty ... 3
- fairly easily ... 4
- easily ... 5
- very easily ... 6

Table 4 presents the corresponding frequency distributions for the CAPI and the CAWI surveys. The normalized Laakso and Taagepera index indicates a larger heterogeneity within the CAWI results than within the CAPI results.

The same question can be used also to investigate how *visual features* of survey questions may influence respondents’ choices. Indeed, in the WEBIT survey, this question was asked using two different visual presentations on random sub-samples of respondents. In particular, response options were organized in a traditional *vertical list of categories*, as reported above, for half of the sample (vert-traditional) and with *horizontal radio buttons* for the remaining part (horiz-radio) (Table 5).

Table 6 summarizes the frequency distributions for the vertical and the horizontal option layouts. In this comparison, it turns out that an important difference between heterogeneity measured by Laakso and Taagepera index is obtained when comparing horizontal-radio versus vertical-traditional layouts.

To test the effect of both survey mode and visual representation on the feeling and the uncertainty components, we modify the model (5) to include two dummy variables as covariates: (1) the CAPI dummy variable to identify the survey mode ($CAPI_i = 1$ for CAPI respondents, and $CAPI_i = 0$ otherwise) and the Horiz dummy variable used to identify the different visual presentations of the question ($Horiz_i = 0$ for the vertical list of response options and $Horiz_i = 1$ for the horizontal sequence of response options). Thus, the reference survey mode is the CAWI-vertical combination, against which the effect of CAPI mode

Table 6

Distribution of response options for subjective economic condition by options visualization feature

Categories	Absolute frequencies		Relative frequencies	
	Vert-trad	Horiz-radio	Vert-trad	Horiz-radio
Midrule				
(1) with great difficulty	51	64	0.122	0.150
(2) with difficulty	45	45	0.107	0.105
(3) with some difficulty	115	92	0.274	0.215
(4) fairly easily	135	102	0.322	0.239
(5) easily	56	66	0.134	0.155
(6) very easily	17	58	0.041	0.136
Total	419	427	1.000	1.000
Laakso-Taagepera Index	–	–	0.690	0.915

and horizontal layout will be separately tested under the model:

$$\begin{aligned}\text{logit}(1 - \pi_i) &= -\beta_0 - \beta_1 \text{CAPI}_i - \beta_2 \text{Horiz}_i; \\ \text{logit}(1 - \xi_i) &= -\gamma_0 - \gamma_1 \text{CAPI}_i - \gamma_2 \text{Horiz}_i.\end{aligned}$$

No significant effect has been found of survey mode and visual representation for the feeling component. With respect to the uncertainty component, no significant difference is found between CAPI respondents and CAWI respondents, while we observe a significant difference between vertical and horizontal layouts for CAWI mode (see Eqs. 9 and 10)⁷.

$$\text{logit}(1 - \hat{\pi}_i) = -1.229 + 0.254\text{CAPI}_i + 3.254\text{Horiz}_i \quad (9)$$

(0.635) (0.727) (0.771)

$$1 - \hat{\xi} = 0.426 \quad (10)$$

(0.014)

In particular, results show that uncertainty rises when the horizontal layout is considered, thus leading to less homogeneous response patterns and, subsequently, higher fuzziness around the actual response signal. These results are in line with what is generally found in the literature, i.e. that the use of radio buttons increases uncertainty such as the la-

bellung of only extreme categories with respect to providing labels for all items.

Next, we investigate if there is any interaction between vertical and horizontal layouts with relevant socio-demographic covariate X (gender: male=0, female=1, presence of children in the household, having a university degree, age: young if under 35 years and elderly if aged more than 64-, households main residence ownership, number of household components). Specifically, the following CUB specification with covariates effects only for the uncertainty component has been tested:

$$\text{logit}(1 - \pi_i) = -\beta_0 - \beta_1 \text{Horiz}_i - \beta_2 X_i - \beta_3 \text{Horiz}_i \times X_i \quad (11)$$

Results are reported in Table 7:

It turns out that responses given by homeowners are more homogeneous than responses given by people not owning their homes. The general conclusion that CAWI responses given on a horizontal layout are more heterogeneous than responses collected via CAPI or CAWI responses on the vertical layout is not modified if one controls for covariates.

Next, we performed a model selection within the class of CUB models for each response group separately to test possible different shelter effects. Table 8 reports some indicators of fitting performances for competing models. Accordingly, Fig. 2 shows that a richer CUB model specification is needed to account also for the inflation in frequency at the first category for all the groups.

Table 9 reports estimated parameters and standard errors: when focusing on responses on the vertical scale, one notices that the CUB with shelter at $c=1$ reduces to a Binomial with shelter. In particular, the weight of the Binomial in the mixture is similar in the two groups, whereas the feeling measure $1 - \xi$ is higher for CAPI than it is for CAWI. Although the difference is not statistically significant, this circumstance could be due to a social desirability bias manifest for CAPI respondents along the whole scale.

As a matter of fact, between CAPI and CAWI vertical responses we observe a larger tendency in CAWI mode to choose options that identify relevant economic difficulties, possibly due to the presence of a social desirability bias, which reduces the choice of these categories in favour of those who identify a better economic situation when the interviewer is present. Furthermore, while in the vertical layout, there is a larger tendency to choose more often the category “fairly easily” (a circumstance that can be identified as a central tendency bias as this option can be seen close to a neutral category, not identifying economic distress neither an extreme affluence), there is a larger tendency to choose the extreme categories when the horizontal layout is adopted. These are the only labelled ones, and thus they can be more clearly identified by the respondents, whereas neutral responses between the third and fourth categories

⁷ Maximum likelihood estimation has been performed via a weighted version of the EM algorithm: estimates of standard errors for parameters were obtained via a Jackknife Repeated Replication method, exploiting available replication weights.

Table 7*Estimates of parameters and standard errors for Model (11)*

	X						
	Gender	Has children	Degree	Young	Elderly	Homeowner	Household size
$\hat{\beta}_0$	1.013 (0.25)	1.018 (0.291)	1.018 (0.524)	0.916 (0.258)	0.805 (0.368)	-0.545 (0.862)	1.29 (0.467)
$\hat{\beta}_1$	-1.999 (0.635)	-1.981 (0.633)	-1.284 (0.523)	-1.848 (0.507)	-2.318 (1.174)	-8.912 (8.237)	4.61 (10.465)
$\hat{\beta}_2$	-0.019 (0.585)	-0.025 (0.413)	0.04 (1.116)	2.386 (3.955)	0.559 (0.503)	1.925 (0.842)	-0.12 (0.136)
$\hat{\beta}_3$	-0.605 (4.309)	-0.763 (4.65)	-13.664 (7.892)	-7.349 (14.206)	0.087 (1.829)	7.048 (8.197)	-5.775 (9.337)

Table 8*Fitting results for competing models for different sub-groups of responses (best performances highlighted in bold fonts)*

	CAPI (SHIW-I)		CAWI-vert (WEBIT)		CAWI-horiz (WEBIT)	
	Loglik	BIC	Loglik	BIC	Loglik	BIC
CUB	-3148.101	6311.381	-581.228	1174.532	-681.710	1375.534
CUB + she(1)	-3018.891	6060.551	-556.097	1130.308	-674.446	1367.062
CUB + she(3)	-3148.255	6319.279	-581.315	1180.744	-679.107	1376.385
CUB + she(4)	-3078.720	6180.209	-566.933	1151.979	-681.710	1381.591

Table 9*Estimated parameters (and standard errors) of CUB and CUB with shelter(1), given survey mode (see Table 8)*

	CUB		CUB+she(1)		
	$\hat{\pi}$	$\hat{\xi}$	Binomial weight	Uniform weight	$\hat{\xi}$
CAPI	0.721 (0.063)	0.568 (0.018)	0.856 (0.020)	0.000 (0.000)	0.509 (0.014)
CAWI-vert	0.810 (0.103)	0.598 (0.022)	0.856 (0.037)	0.000 (0.000)	0.534 (0.015)
CAWI-horiz	0.219 (0.073)	0.594 (0.159)	0.353 (0.092)	0.548 (0.114)	0.519 (0.050)

cannot be clearly distinguished due to the absence of complete labelling of all categories.

To test for these hypotheses (social desirability bias and central tendency bias) we report the z-test for the comparison of two independent proportions⁸.

To verify the presence of a social desirability bias we compare responses concerning the two lowest categories (conveying actual economic difficulties) obtained using the

CAPI mode with those coming from the CAWI, considering in the latter survey first only responses to questions using the vertical representation (Table 10) and then all answer (Table 11). Both the results confirm evidence of social desirability bias at significance level $\alpha = 0.05$.

Similarly, Table 12 reports relevant information on the z-test to compare the frequency of the central categories (third and fourth, conveying pseudo-neutral evaluation) between vertical response mode (with all labelled categories) and the horizontal mode (with radio-buttons, and labels only for the extreme categories). Significant evidence is found for a strong tendency to central categories characterizing

⁸ Notice that the application of this inferential procedure—computing the proportion of lowest scores on the basis of the weighted frequency distribution and with respect to the observed sample sizes for the two groups—is a stretch in view of the sampling design.

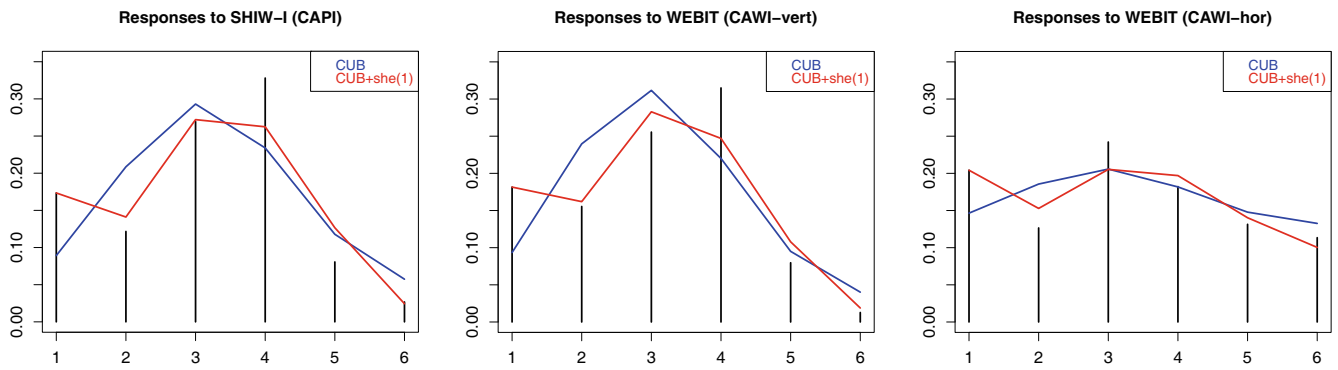


Fig. 2

Graphical check of the goodness of fit of candidate models for (weighted) observed distributions, given survey mode

Table 10

Testing for social desirability bias for CAPI versus CAWI respondents on the basis of the observed weighted distribution for responses collected on vertical scales

	$\sum_{r=1}^2 f_r^{(w)}$
CAPI	0.295
CAWI-vert	0.337
p-value for one-sided z-test	0.045

Table 11

Testing for social desirability bias for CAPI versus CAWI respondents on the basis of the observed weighted distribution

	$\sum_{r=1}^2 f_r^{(w)}$
CAPI	0.295
CAWI	0.334
p-value for one-sided z-test	0.021

Table 12

Testing for central tendency bias for Vertical versus Horizontal response layout on the basis of the observed weighted distribution

	$\sum_{r=3}^4 f_r^{(w)}$	$\sum_{r=3}^4 f_r$
Vertical	0.593	0.562
Horizontal	0.424	0.454
p-value for one-sided z-test	$1 \cdot 10^{-11}$	$1 \cdot 10^{-5}$

responses collected on vertical layout, due to the complete labelling of categories.

4.3 Comparing the effect of CATI and CAWI survey mode on firms data with the BOSISF dataset

The Business Outlook Survey of Industrial and Service Firms (BOSISF hereafter) is annually conducted by the Bank of Italy since 1993 to collect qualitative information on firms’ performance and on the main economic variables (Bank of Italy 2017b). The survey is conducted on about 4500 firms (3000 firms in industry with 20 and more workers, 1000 firms in non-financial private services and 500 construction firms with 10 and more workers). Firms are contacted by e-mail and can decide either to fill in the questionnaire on the web (CAWI) or to provide the information by telephone (CATI).⁹ Interviews by phone are administered by officials of the Bank of Italy’s local branches, specially trained to conduct business surveys (Bank of Italy 2017a).

To study the effect of survey mode on businesses’ answer elicitation mechanism we consider two questions. The first question collect the “realization rate of investment”, i.e. how much of the investment expenditure that was planned in the previous year has been actually realized in the current year (Q_1). The second question collects the “expected investment growth rate”, that is the change in investment expenditures expected in the future year with respect to the current one (Q_2). Particularly, we refer to the 2020 edition of the survey since more answer options were provided for these questions due to the larger volatility of investment associated with the economic crisis resulting from the Covid-19 pandemic.

⁹ In few cases interviews are conducted with a personal visit. These only represent about 1 per cent of the sample in the used 2020 wave and have been excluded from this study.

Table 13

Distribution of response options for realized investment variation (Q_1) in CATI and CAWI surveys

Categories	Absolute frequencies		Relative frequencies	
	CATI	CAWI	CATI	CAWI
(1) Lower by more than -50%	173	227	0.111	0.086
(2) Lower by between -50% and -25%	111	264	0.071	0.100
(3) Lower by between -25% and -10%	139	259	0.090	0.099
(4) Lower by between -10% and -3%	102	200	0.066	0.076
(5) Stable between -3% and +3%	819	1297	0.527	0.494
(6) Higher by between +3.1 and 10%	82	138	0.053	0.053
(7) Higher by between +10.1 and 50%	68	90	0.044	0.034
(8) Higher by more than +50%	31	31	0.020	0.012
(9) Do not know, do not wish to answer	28	122	0.018	0.046
Total	1553	2628	1.000	1.000
Laakso-Taagepera Index	-	-	0.274	0.316

Namely, the questions are the following:

Q₁: Compared to the level planned at the end of 2019, nominal expenditure on (tangible and intangible) fixed investment in the current year will be:

Q₂: How does planned nominal expenditure on fixed investment in 2021 compared with that in 2020? with the same response options:

- Lower by more than -50% ... 1
- Lower by between -50% and -25% ... 2
- Lower by between -25% and -10% ... 3
- Lower by between -10% and -3% ... 4
- Stable between -3% and +3% ... 5
- Higher by between +3.1 and 10% ... 6
- Higher by between +10.1 and 50% ... 7
- Higher by more than +50% ... 8
- Do not know, do not wish to answer ... 9

Tables 13 and 14 summarize the frequency distributions respectively of the variation in realised (Q_1) and expected (Q_2) investment for CATI and CAWI respondents.

It should be noted that the intrinsic uncertainty within the answers to the two questions is by construction dif-

Table 14

Distribution of response options for expected investment variation (Q_2) in CATI and CAWI surveys

Categories	Absolute frequencies		Relative frequencies	
	CATI	CAWI	CATI	CAWI
(1) Lower by more than -50%	76	94	0.049	0.036
(2) Lower by between -50% and -25%	33	105	0.021	0.040
(3) Lower by between -25% and -10%	68	131	0.044	0.050
(4) Lower by between -10% and -3%	68	99	0.044	0.038
(5) Stable between -3% and +3%	663	1201	0.427	0.457
(6) Higher by between +3.1 and 10%	221	367	0.142	0.139
(7) Higher by between +10.1 and 50%	162	205	0.104	0.078
(8) Higher by more than +50%	69	49	0.045	0.019
(9) Do not know, do not wish to answer	193	377	0.124	0.143
Total	1553	2628	1.000	1.000
Laakso-Taagepera Index	==	==	0.401	0.352

ferent since the first question concerns an observable item while the second requires the formulation of an expectation. Therefore, considering both items, we can test how respondents' choice is realized with different survey modes in dissimilar uncertainty frameworks.

Before dwelling into a model-based analysis, it is worth pursuing a preliminary investigation of the target rating variables Q_1 and Q_2 : from Tables 13 and 14, it can be inferred that, overall, the number of "don't know" responses is higher when evaluations refer to the future (Q_2) with respect to evaluations referring to the past (Q_1) (the association is highly significant according to the X^2 test).

With respect to survey mode and according to a standard X^2 test, it follows that *don't know* options occur more frequently for respondents answering via CAWI than for those answering via CATI only for Q_1 : there is no significant association between the occurrence of "don't know" responses and survey mode (CATI-CAWI) for Q_2 . This first result might indicate that there is a tendency to adopt a satisficing behaviour when providing information concerning variation between planned and realized investments when the interview takes place via the web.

In order to provide a unified summary of these results, we fitted a logistic regression on the indicator variable D_i

Table 15

Estimation results for CUSH model for Q_1 and Q_2 in terms of CAWI covariate

	ν_0	ν_1	BIC
Q_1	0.019 (0.106)	-0.121 (0.140)	10942.01
Q_2	-0.281 (0.115)	0.281 (0.147)	11061.26

reporting if the response is “don’t know” ($D_i = 1$) or observed $D_i = 0$, after merging Q_1 and Q_2 evaluations and thus assuming that they are conditionally independent given the chosen explanatory variables, namely a dummy variable CAWI to flag the survey modality (CAWI = 1 for CAWI respondents, CAWI = 0 for CATI respondents), and a dummy variable X_i created to identify past ($X_i = 0$, namely Q_1) from future evaluations ($X_i = 1$, namely Q_2). It follows that, overall, both the modality CAWI of the questionnaire and the fact that the question requires an assessment about the future (rather than about the past) contribute to significantly increase the probability of observing a “don’t know” response. This result confirms the findings that satisficing behaviours that may lead to the DK choice are better contrasted by the interviewer with CATI mode with respect to a self-administered mode and that expectations are generally more subject to uncertainty than realized outcomes.

After omitting “don’t know” responses from the analysis, the Spearman correlation coefficient for the observed ratings for Q_1 and Q_2 amounts to $\rho = -0.0381$, indicating poor dependence between the two ratings, with a slightly negative direction: it follows then that for increasing observed variation in the current year for nominal expenditure, there is a slight tendency to expect a lower variation in 2021 than that observed in 2020. This result is in line with the fact that the latent continuous variables to which the two qualitative outcomes refer to contain the realized investment in the current year, respectively in the numerator for the realized rate and in the denominator for the expected rate.

With reference to the observed frequency distributions reported in Tables 13 and 14, it is seen that response distributions to Q_1 and Q_2 present both a strong frequency inflation in category 5 (conveying overall stability). This result, referred to as central tendency bias (Pimentel 2019), reflects the tendency to choose the neutral category in a Likert scale, when available: this circumstance is particularly relevant in this case. Thus, CUB models cannot be expected to be sufficiently adequate to fit the data, even after including a possible shelter effect: indeed, the shelter parameter δ is not significant. This circumstance may be due to the fact that the modal value coincides with the category where the inflation is observed and to the large heterogeneity of the distribution. This conclusion continues to hold even after

controlling for selected covariates. For this reason, in the following discussion we focus on fitting results obtained with the CUSH model (4) (Capecchi and Piccolo 2017), which allows investigating more carefully the variables related to the central tendency bias (even if the scale is not balanced with respect to the centre), which assumes extreme importance in this circumstance¹⁰.

First, consider the CUSH model:

$$\text{logit}(\delta_i) = \nu_0 + \nu_1 \text{CAWI}_i \quad (12)$$

for both Q_1 and Q_2 , where CAWI is a dummy factor identifying CAWI respondents ($\text{CAWI}_i = 1$) against CATI respondents ($\text{CAWI}_i = 0$). Results are reported in Table 15.¹¹

Significant differences in central tendency bias between CAWI and CATI respondents are found only for Q_2 evaluation. In particular, the positive sign of estimated ν_1 for Q_2 is in line with the finding that the interviewer may help to reduce satisficing behaviours, in this case, related to the tendency to choose the neutral category due to the difficulty to provide information with respect to a future event.

When stratifying responses across sectors of activity, there is a significant CAWI effect for inflation in category $c=5$ for chemicals, rubbers and plastics industry in Q_1 and for retail trade and food industries in Q_2 . For the sake of completeness, Table 16 reports estimates of parameters and standard errors (in parentheses) for Model (12) fitted to Q_1 and Q_2 responses provided by enterprises in each economic sector to check for differences in inflation at $c=5$.

For both Q_1 and Q_2 , no significant effect modifying central tendency bias are found due to covariates and their interactions with survey mode: this statement follows from the estimation of model (8) for each covariate X (geographical region, dimensional class and export quota).

Next, with focus on responses to Q_2 only, the following model ($BIC = 10592$) can be estimated to explain if frequency inflation at $c=5$ can be interpreted in terms of negative variation declared for Q_1 ($I(Q_1 \leq 4)$) or positive variation declared for Q_1 ($I(Q_1 \geq 6)$):

$$\begin{aligned} \text{logit}(\delta_i) = & 0.418 + 0.420\text{CAWI}_i - 1.910 I(Q_{1i} \leq 4) \\ & - 1.550 I(Q_{1i} \geq 6) \end{aligned} \quad (13)$$

(0.143) (0.162) (0.192)
(0.266)

¹⁰ CUB estimation results are omitted for this case study since this model has poor fitting performance with respect to the proposed CUSH model.

¹¹ For model-based inference within the class of CUB models, estimates of standard errors for model parameters have been obtained via Taylor linearization method given the sampling design and the availability of strata allocation.

Table 16

Estimates of parameters and standard errors for CUSH, Model (12) fitted to Q_1 and Q_2 responses for each economic sector.

Economic Sector	$Q_1: \hat{\nu}_0$	$Q_1: \hat{\nu}_1$	$Q_2: \hat{\nu}_0$	$Q_2: \hat{\nu}_1$
Food beverages and tobacco	-0.024 (0.241)	0.347 (0.319)	-0.681 (0.262)	0.775* (0.335)
Textiles, clothing, leather, footwear	-0.163 (0.381)	-0.501 (0.517)	-0.473 (0.398)	-0.371 (0.540)
Chemicals, rubber, plastics	0.343 (0.304)	-1.028* (0.402)	-0.080 (0.319)	-0.375 (0.407)
Basic metals	-0.223 (0.451)	0.519 (0.574)	-1.093 (0.488)	1.055 (0.613)
Engineering	0.052 (0.185)	-0.188 (0.234)	-0.160 (0.188)	0.416 (0.234)
Other manufacturing	-0.145 (0.351)	0.243 (0.442)	-0.180 (0.352)	-0.047 (0.449)
Energy and mining	-1.087 (0.442)	0.303 (0.585)	-0.484 (0.424)	0.520 (0.556)
Retail trade	0.080 (0.228)	0.169 (0.293)	-0.319 (0.239)	0.705* (0.301)
Hotels and restaurants	-0.701 (0.712)	-0.143 (0.978)	-4.459 (7.231)	3.529 (7.268)
Transport, storage and communication	0.470 (0.272)	-0.392 (0.350)	0.055 (0.281)	-0.173 (0.359)
Other services	0.182 (0.374)	-0.198 (0.461)	0.730 (0.383)	-0.443 (0.468)

*p < 0.05

All the effects are significant: then, it follows that inflation at $c=5$ for Q_2 increases for CAWI respondents, also when accounting to responses provided to the question about current investment Q_1 . Notice also that inflation at $c=5$ for Q_2 decreases for negative or positive variations reported for Q_1 evaluations (mostly, for negative variations). This means that people reporting either $Q_1 \leq 4$ or $Q_1 \geq 6$

Table 17

Estimates of parameters and standard errors for model (14) for ratings on Q_2 .

Economic Sector	$\hat{\nu}_0$	$\hat{\nu}_1$	$\hat{\nu}_2$
Food beverages and tobacco	-1.646 (0.799)	0.802* (0.356)	0.205 (0.138)
Textiles, clothing, leather, footwear	-1.984 (0.761)	-0.329 (0.559)	0.379* (0.163)
Chemicals, rubber, plastics	-0.715 (0.799)	-0.386 (0.454)	0.151 (0.168)
Basic metals	-3.532 (5.426)	1.122 (1.618)	0.534 (0.94)
Engineering	-1.637 (0.448)	0.511* (0.247)	0.333* (0.086)
Other manufacturing	-2.377 (0.951)	0.010 (0.528)	0.517* (0.186)
Energy and mining	-1.831 (1.375)	0.638 (0.737)	0.267 (0.277)
Retail trade	-1.04 (0.558)	0.723 (0.32)	0.161 (0.110)
Hotels and restaurants	-23.253 (2803139.009)	17.093 (2803130.832)	1.406 (2.437)
Transport, storage and communication	-2.297 (0.973)	-0.048 (0.414)	0.509* (0.191)
Other services	-0.959 (1.037)	-0.399 (0.585)	0.395 (0.206)

*p < 0.05

have lower probability to inflate frequency 5 for Q_2 , and thus expect unstable variation for Q_2 .¹²

Finally, to investigate differences across economic sectors in the joint effect of Q_1 ratings and survey mode on Q_2 evaluations we estimate, for each sector, the following model:

$$\text{logit}(\delta_i) = \nu_0 + \nu_1 \text{CAWI}_i + \nu_2 Q_{1i}. \quad (14)$$

Results are reported in Table 17 and indicate that inflation in the category conveying stability $c=5$ is significantly

¹² Further specifications of the model including the stability for Q_1 (namely, $Q_1 = 5$) among covariates confirm the persistence of shelter effect at $c=5$: i.e. those who reported the stability for Q_1 are more likely to shelter at $c=5$ also for Q_2 , regardless of the survey mode.

higher for CAWI respondents for sectors of food, beverages and tobacco, retail trade and engineering (in decreasing order of $\hat{\nu}_1$). Ratings on Q_1 , instead, significantly affect central tendency bias for other manufacturing industries, transport, storage and communication, textiles, clothing, leather and footwear, engineering and transport, storage and communication (in decreasing order of $\hat{\nu}_2$)¹³.

5 Conclusions

In order to contribute to the literature that investigates the presence of possible sources of distortion in micro-data from sample surveys, this work investigated many of the factors related to survey mode and questionnaire design that can potentially influence the response choice. In particular, using data from official surveys on both households and firms, we focus on the effects of different survey modes, visual representation of survey questions and the presence/absence of the don't know (DK) option.

The novelty of the approach here introduced with respect to the state of the art on CUB models applications consists in the possibility to test for a possible effect of the different survey modes or questionnaire features (DK option, visual features) on *both* the feeling and uncertainty of the responses in a more straightforward way than classical methods, as scale-location cumulative link models.

Although referring to specific cases, the results show that, with respect to the feeling component, neither the presence of the "Don't know" option, nor the survey mode or the graphical representation appear to significantly modify the way the rating options are perceived and used. On the other hand, we found evidence of the effects of these features on uncertainty and on shelter choices. In particular:

- **Survey mode:** Our study provided evidence that CAWI collection mode leads to an increase in propensity to CUB uncertainty in responses provided by firms regarding expected investment. For the same question, we also find that firms choose more often the options related to a neutral position or to the absence of knowledge when the CAWI mode was used. This result may be due to the fact that, in the absence of an interviewer, respondents may adopt more easily satisficing behaviour to reduce their effort, especially when more complex questions are concerned. On the other hand, when household surveys are concerned, results show that the social desirability

bias is reduced when using CAWI mode with respect to CAPI.

- **Don't know option:** This option is more frequently used when questions about expectations are concerned and sometimes interacts with variables regarding the size of the phenomenon. In the case of reported estimates of the value of parents' dwellings, the weight of CUB uncertainty was higher when the DK option was present for households with a higher level of income or leaving in the Centre of Italy, who were also reporting on average higher evaluations for the item.
- **Visual representation:** Comparing a horizontal layout, where labelling is provided only for extreme classes, with a classical vertical representation, where all options are labelled, we observe an increase in the heterogeneity of the responses in the former layout, and an increase in central tendency bias in the latter.

These results are in line with what is suggested by the literature, providing evidence that CUB models are a robust tool to interpret and compare data collected with different survey techniques and questionnaire design. However, the peculiar nature of these results suggests considering these findings as non-exhaustive and continuing to carry out appropriate tests whenever different techniques are used in sub-samples in order to exclude any possible source of distortion in the results relating to the examined population.

Acknowledgements Partially Supported by grant SI-WCWB from University of Naples Federico II (FRA 2022), DR n 3429, 07/09/2023 (CUP: E65F22000050001). The authors wish to thank Lucia Modugno and Andrea Neri for helpful comments. The views in this paper are those of the authors only and do not necessarily reflect those of the Bank of Italy.

Data Availability Statement Data used for the case studies in Sect. 4.1 and Sect. 4.2 are available as online supplementary material.

The microdata used in Sect. 4.3 are not publicly shareable due to confidentiality constraints. Access to data from the Business Outlook Survey of Industrial and Service Firms (SONDTEL) is only possible through the Bank of Italy's Research Data Center, which provides secure access to authorized researchers through the REX remote processing system or the laboratory on Bank of Italy's premises. Detailed information on how to apply for access is available on the Bank of Italy's official website: Banca d'Italia—Microdata of Industrial and Service Firms. Additionally, some of the data used in the paper, specifically regarding the used survey mode (e.g., CATI vs CAWI) are only available to staff affiliated at the Bank of Italy who are directly responsible for conducting the survey. Therefore these data cannot be gathered through the Research Data Center.

¹³ Notice that the variance-covariance matrix for the estimation of CUSH model (14) is ill-conditioned for the sector "Hotels and restaurants" due to small sample sizes overall and also within certain strata, given the survey mode: as a consequence, standard errors are divergent indicating that the model could be unstable for this economic sector.

References

- Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd edn.). Wiley.

- Baffigi, A., Cannari, L., & D'Alessio, G. (2016). *Cinquanta anni di indagini sui bilanci delle famiglie italiane: Storia, metodi, prospettive (fifty years of household income and wealth surveys: history, methods and future prospects)*. Bank of Italy Occasional Paper, (368).
- Bank of Italy (2017a). *Business outlook survey of industrial and service firms. Methods and Sources: Methodological notes*, November, https://www.bancaditalia.it/pubblicazioni/metodi-e-fonti-note/metodi-note-2017/en-metodologia_sondaggio_impr_industr_serv.pdf?language_id=1.
- Bank of Italy (2017b). *Survey of industrial and service firms. Methods and Sources: Methodological notes*, July, https://www.bancaditalia.it/pubblicazioni/metodi-e-fonti-note/metodi-note-2017/en_survey_methodology_invid.pdf?language_id=1.
- Bank of Italy. (2018). *The survey on household income and wealth. Methods and Sources: Methodological notes* March, https://www.bancaditalia.it/pubblicazioni/metodi-e-fonti-note/metodi-note-2018/MOP_IBF_en.pdf?language_id=1.
- Barcaroli, G., Gambacorta, R., Conte, L.L., Murgia, M., Neri, A., & Zanichelli, F. (2019). *L'indagine sperimentale web sulle famiglie italiane: Una valutazione della tecnica cawi per rilevare informazioni sul reddito e la ricchezza*. ISTAT Metodi letture Statistiche. <https://www.istat.it/it/archivio/228589>
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* <https://doi.org/10.1145/1541880.1541883>.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57, 289–300.
- Bound, J., Brown, C., & Mathiowetz, N. (2001). Chapter 59—measurement error in survey data. In J. Heckman & E. Leamer (Eds.), Elsevier. [https://doi.org/10.1016/S1573-4412\(01\)05012-7](https://doi.org/10.1016/S1573-4412(01)05012-7).
- Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27(3), 281–291. <https://doi.org/10.1093/pubmed/fdi031>.
- Braunsberger, K., Wybenga, H., & Gates, R. (2007). A comparison of reliability between telephone and web-based surveys. *Journal of Business Research*, 60(7), 758–764.
- Capecchi, S., & Iannario, M. (2016). Gini heterogeneity index for detecting uncertainty in ordinal data surveys. *METRON*, 74, 223–232.
- Capecchi, S., & Piccolo, D. (2017). Dealing with heterogeneity in ordinal responses. *Quality & Quantity*, 51, 2375–2393.
- Capecchi, S., Endrizzi, I., Gasperi, F., & Piccolo, D. (2016). A multi-product approach for detecting subjects' and objects' covariates in consumer preferences. *British Food Journal*, 118, 515–526.
- Cerulli, G. (2020). *CUB: Stata module to estimate ordinal outcome model estimated by a mixture of a uniform and a shifted binomial*. Statistical Software Components, S458727. Boston College Department of Economics.
- Cerulli, G., Simone, R., Di Iorio, F., Piccolo, D., & Baum, C. (2022). Fitting mixture models for feeling and uncertainty for rating data analysis. *The STATA Journal*, 22(1), 195–223.
- Chan, Z., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2), 127–135.
- Chang, L., & Krosnick, J. (2010). Comparing oral interviewing with self-administered computerized questionnaires. An experiment. *Public Opinion Quarterly*, 74(1), 154–167. <https://doi.org/10.1093/poq/nfp090>.
- Colombi, R., & Giordano, S. (2016). A class of mixture models for multidimensional ordinal data. *Statistical Modelling*, 16, 322–340.
- Coombs, C., & Coombs, L. (1976). “Don't know” item ambiguity or respondent uncertainty? *Public Opinion Quarterly*, 40, 495–514.
- Corduas, M., Iannario, M., & Piccolo, D. (2009). A class of statistical models for evaluating services and performances. In M. Bini, P. Monari, D. Piccolo & L. Salmaso (Eds.), *Statistical methods for the evaluation of educational services and quality of products, contribution to statistics* (pp. 99–117). Springer.
- Couper, M. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3), 145–156.
- Couper, M., Traugott, M., & Lamias, M. (2001). Web survey design and administration. *Public opinion quarterly*, 65(2), 230–253.
- Davis, R., Couper, M., Janz, N., Caldwell, C., & Resnicow, K. (2009). Interviewer effects in public health surveys. *Health Education Research*, 25(1), 14–26. <https://doi.org/10.1093/her/cyp046>.
- De Leeuw, E., Hox, J., & Scherpenzeel, A. (2011). Mode effect or question wording? measurement error in mixed mode surveys. In *Proceedings of the survey research methods section* (pp. 5959–5967). American Statistical Association.

- DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality. *Quality & Quantity*, 52, 1523–1559.
- D'Elia, A., & Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, 49, 917–934.
- Dillman, D. (2002). Survey nonresponse in design, data collection, and analysis. In *Survey nonresponse* (pp. 3–26).
- Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social science research*, 38(1), 1–18.
- Faiella, I., & Gambacorta, R. (2007). *The weighting process in the SHIW*. Bank of Italy Temi di Discussione (Working Paper), 636.
- Fin, F., Iannario, M., Simone, R., & Piccolo, D. (2017). The effect of uncertainty on the assessment of individual performance: empirical evidence from professional soccer. *Electronic Journal of Applied Statistical Analysis*, 10, 677–692.
- Fricker, S., Galesic, M., Tourangeau, R., & Ting, Y. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69(3), 370–392.
- Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, 34(2), 244–254.
- Gambacorta, R., Conte, M.L., Murgia, M., Neri, A., Rizzi, R., & Zanichelli, F. (2018). *Mind the mode: Lessons from a web survey on household finances*. Bank of Italy Occasional Paper, (437).
- Golia, S. (2015). On the interpretation of the uncertainty parameter in CUB models. *Electronic Journal of Applied Statistical Analysis*, 8, 312–328.
- Hambleton, R.K. (1991). *Fundamentals of item response theory*. Vol. 2. New York: SAGE.
- Iannario, M. (2012). Modelling shelter choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, 21, 1–22.
- Iannario, M., Monti, A.C., & Piccolo, D. (2016). Robustness issues in CUB models. *TEST*, 25(4), 731–750.
- Iannario, M., Monti, A., Piccolo, D., & Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics*, 11, 3407–3445.
- Iannario, M., Piccolo, D., & Simone, R. (2024). *CUB: A class of mixture models for ordinal data*. R package version 1.1.5.
- Ip, R., & Wu, K. (2024). A mixture distribution for modelling bivariate ordinal data. *Statistical Papers*, 65, 4453–4488.
- Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1), 3–20. <https://doi.org/10.1111/j.1751-5823.2010.00102.x>.
- Jäckle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: effects of a mixed mode design on item and unit non-response. *Survey Research Methods*, 9(1), 57–70.
- Kankaras, M., & Capecchi, S. (2025). Neither agree nor disagree: use and misuse of the neutral response category in Likert-type scales. *Metron*, 83, 111–140.
- Keusch, F., & Yang, T. (2018). Is satisficing responsible for response order effects in rating scale questions? *Survey Research Methods*, 12, 259–270.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. *Public opinion quarterly*, 72(5), 847–865.
- Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, (5), 213–236.
- Krosnick, J., Holbrook, A., Berent, M., Carson, R., Hanemann, M., Kopp, R., Mitchell, C., Presser, S., Ruud, P., Smith, V.K., Moody, W., Green, M., & Conaway, M. (2002). The impact of “no opinion” response options on data quality: non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66(3), 371–403. <https://doi.org/10.1086/341394>.
- Malhotra, M. (2008). Completion time and response order effects in web surveys. *Public opinion quarterly*, 72(5), 914–934.
- Maloshonok, N., & Terentev, E. (2016). The impact of visual design and response formats on data quality in a web survey of mooc students. *Computers in Human Behavior*, 62, 506–515.
- Manisera, M., & Zuccolotto, P. (2014). Modeling “don’t know” responses in rating scales. *Pattern Recognition Letters*, 45, 226–234.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42, 109–142.
- McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. Hall: Chapman.
- Melani, C., Dillman, D., & Smyth, J. (2008). The effects of mode and format on answers to scalar questions in telephone and web surveys. *Advances in telephone survey methodology*, 12, 250–275.
- Montagni, I., Cariou, T., Tzourio, C., & Gonzalez-Caballero, J. (2019). “I don’t know”, “I’m not sure”, “I don’t want to answer”: A latent class analysis explaining the informative value of nonresponse

- options in an online survey on youth health. *International Journal of Social Research Methodology*, 22(6), 651–667.
- Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, 5, 85–104.
- Piccolo, D., & D'Elia, A. (2008). A new approach for modelling consumers' preferences. *Food Quality & Preferences*, 19, 247–259.
- Piccolo, D., & Simone, R. (2019a). The class of CUB models: Statistical foundations, inferential issues and empirical evidence (with discussions and rejoinder). *Statistical Methods & Applications*, 28, 389–493.
- Piccolo, D., & Simone, R. (2019b). Rejoinder to the discussion of "The class of CUB models: statistical foundations, inferential issues and empirical evidence". *Statistical Methods & Applications*, 28, 477–493.
- Pimentel, J. (2019). Some biases in likert scaling usage and its correction. *International Journal of Science: Basic and Applied Research*, 45(1), 183–191.
- Regmi, P., Waithaka, E., Paudyal, A., Simkhada, P., & Tejligen, E. V. (2016). Guide to the design and application of online questionnaire surveys. *Nepal Journal of Epidemiology*, 6(4), 640–644.
- Rhodes, S., Bowie, D., & Hergenrather, K. (2003). Collecting behavioural data using the world wide web: considerations for researchers. *Journal of Epidemiology & Community Health*, 57, 68–73.
- Roberts, C., Vandenplas, C., & Ernst Stähli, M. (2014). Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs. *Survey Research Methods*, 8, 67–80.
- Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores*. Psychometric Monograph No. 17. Richmond: Psychometric Society.
- Sarracino, F., Riillo, C., & Mikucka, M. (2017). Comparability of web and telephone survey modes for the measurement of subjective well-being. *Survey Research Methods*, 11, 141–169. <https://doi.org/10.18148/srm/2017.v11i2.6740>.
- Sciandra, M., Fasola, S., Albano, A., & Plaia, A. (2024). Discrete beta and shifted beta-binomial models for rating and ranking data. *Environ Ecol Stat*, 31(5), 317–338.
- Simon, H. (1957). *Models of man*. Wiley.
- Simone, R. (2020). *FastCUB: Fast EM and best-subset selection for CUB models for rating data*. R package version 0.0.2..
- Simone, R. (2022). On finite mixtures of discretized beta model for ordered responses. *TEST*, 31, 828–855.
- Simone, R. (2023). Uncertainty diagnostics of binomial regression trees for ordered rating data. *Journal of Classification*, 40, 79–105.
- Simone, R., Di Iorio, F., & Lucchetti, R. (2019). CUB for GRETL. In F. Di Iorio & R. Lucchetti (Eds.), *GRETL 2019: Proceedings of the International Conference on GNU Regression, Econometrics and Time series Library* (pp. 147–166). feDOA University Press.
- Simone, R., Tutz, G., & Iannario, M. (2020). Subjective heterogeneity in response attitude for multivariate ordinal outcomes. *Econometrics and Statistics*, 14, 145–158.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modelling: multilevel, longitudinal, and structural equation models*. Chapman; Hall/CRC.
- Sur, P., Shmueli, G., Bose, S., & Dubey, P. (2015). Modelling bimodal discrete data using conway–maxwell–poisson mixture models. *Journal of Business and Economic Statistics*, 33, 352–365.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, 133(5), 859.
- Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
- Tourangeau, R., Couper, M., & Conrad, F. (2004). Spacing, position, and order: interpretive heuristics for visual features of survey questions. *Public opinion quarterly*, 68(3), 368–393.
- Tourangeau, R., Couper, M., & Conrad, F. (2013). Up means good": the effect of screen position on evaluative ratings in web surveys. *Public Opinion Quarterly*, 77(S1), 69–88. <https://doi.org/10.1093/poq/nfs063>.
- Tourangeau, R., Yan, T., & Sun, H. (2020). Who can you count on? understanding the determinants of reliability. *Journal of Survey Statistics and Methodology*, 8, 903–931.
- Tutz, G. Ordinal regression: a review and a taxonomy of models. *WIREs Computational Statistics*, 14(2), e1545.
- Tutz, G. (2012). *Regression for categorical data*. Cambridge University Press.
- Ursino, M., & Gasparini, M. (2018). A new parsimonious model for ordinal longitudinal data with application to subjective evaluation of a gastrointestinal disease. *Stat Methods Med Res*, 27(5), 1376–1393.
- Van Vaerenbergh, Y., & Thomas, T. (2012). Response styles in survey research: a literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25(2), 195–217.
- Vannieuwenhuyze, J. (2013). On the relative advantage of mixed-mode versus single-mode surveys. *Survey Research Methods*, 8, 31–42.

- Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74(5), 1027–1045. <https://doi.org/10.1093/poq/nfq059>.
- Velez, P., & Ashworth, S.D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods*, 12, 69–74.