Title:

Noblesse oblige
Reflections on methodological pitfalls of the Eurobarometer

Authors:

Prof. Dr. Jelke Bethlehem
Leiden University | Institute of Political Science
Wassenaarseweg 52, 2333 AK Leiden, Netherlands
j.g.bethlehem@fsw.leidenuniv.nl

Prof. Dr. Joop van Holsteijn
Leiden University | Institute of Political Science
Wassenaarseweg 52, 2333 AK Leiden, Netherlands
holsteyn@fsw.leidenuniv.nl

Noblesse oblige
Reflections on methodological pitfalls of the Eurobarometer

1. Introduction[1]

The Eurobarometer (EB) is an established data collection institution. The multifaceted system of large-scale surveys has been commissioned by the European Commission (EC) since the early 1970s and covers over half a century of public opinion on the European community. EB surveys are conducted in all member states of the European Union (EU), in some cases even at the time when countries only were aspiring members. Its objective has always been to assess public awareness, knowledge and evaluations of and support for the institutional structure and the activities and ambitions of the EU (see also e.g., Signorelli, 2012; see also Anderson & Hecht, 2018).

There are three main types of EB surveys. The most important one is the *Standard Eurobarometer* (StEB), i.e. a longitudinal survey that primarily attempts to gauge trends in citizens' opinions about European integration, institutions, and policies. In July-August 2020 the fieldwork of StEB 93 took place; its first results were published in October 2020. The second type is the *Special Eurobarometer* (SpEB) that has been measuring opinions about a diversity of themes or policy areas. Questions are asked at one particular moment in time on a specific topic, e.g. on 'Making our food fit for the future – Citizens' expectations' (fieldwork August-September 2020, publication date October 2020).[2] The third and final main type is the *Flash Eurobarometer* (FEB). This type of EB uses a short questionnaire to quickly measure opinions about a specific topic and is sometimes restricted to a particular sub-population. For example, in July 2020 the report based on Flash Eurobarometer 487 regarding the 'Introduction of the euro in the Member States that have not yet adopted the common currency' was published.[3]

The EB serves multiple purposes and audiences. On the occasion of the 35th birthday of the EB then European Commissioner Margot Wallström stated: "Since 1973, the European Commission has been monitoring the evolution of public opinion in the Member States, thus helping the preparation of texts, decision-making and the evaluation of its work. (…) Over more than three decades, Eurobarometer polls have given us a solid indication of the concerns, needs and opinions of European citizens". Claus Sørensen, head of the Directorate-General of Communication of the EU, qualified the EB as "a unique source of knowledge and a remarkable tool for policy advice" (quotes from European Commission, 2008: no page numbers). Sørensen did not exaggerate: EB data have fueled public debate, were input for administration and policy decisions in EU member states and at the European level, and empirically founded numerous scholarly analyses and publications, helping to obtain a better understanding of the EU (Schmitt, 2003). The impact on the public at large is hard to establish, but the relevancy for the scholarly community and the debate on the EU is obvious. For example, *European Union Politics*, one of the leading journals in this field, published in its first ten years of existence (2000-2009) in 38 editions a total of 168 original papers: 28 papers (17 percent) were completely or partly based on EB data. From a public opinion approach, the importance of the EB is indicated by the fact that out of the 724 research articles published (until July 2020) in the *International Journal of Public Opinion Research*, 96 or about 13 percent refer to the Eurobarometer. And even the more US oriented *Public Opinion Quarterly* since 2000 contains at least once a year a paper that is based on data from one of the EB surveys. So understandably, when at the turn of the century Gabel and colleagues lamented the lack of "scientific maturity in the key area of data accumulation and

[1] This paper is partly based on [authors].

[2] More information on all Eurobarometer studies and research reports are available on http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm.

[3] There are also 'qualitative' EB studies on motivations, feelings and reactions of selected social groups towards a particular topic; this paper does not address such studies.

integration" in comparative EU research, they referred to the "value and prominence of the Eurobarometer" as the exception and shining example (Gabel *et al.*, 2002: 482, 485). In a competition for the most frequently used data set for scientific analyses the EB arguably has only a single contender: the American National Election Studies (e.g., Schmitt, 2003, 245).

Indeed, scholars by default seem to opt for the EB when looking for data on EU-related projects. On the rare occasion that they use different data, they justify not using EB data. "To test our hypotheses while adequately controlling for other influences, we cannot rely on the most readily available data (the Eurobarometer) because a number of the key concepts are either not included in the barometer or poorly operationalized. (…) We therefore collected new data (…)" (de Vreese et al., 2008: 516). However, this self-evident use of EB data causes researchers to pay insufficient attention to data quality; and a simple reference to its default use - "Like other students of EU public opinion, we draw our data from the Eurobarometer series" (Nelsen et al., 2001: 195) – should be considered unsatisfactory in terms of argumentation for and reflection on the EB and EB data quality.

Furthermore, the EB impact is not restricted to scholarly use, but is extended over national politics, policies and public opinion at large. For example, in 2015 Dutch MP Pia Dijkstra proposed a law on organ donation in which she explicitly referred to the 2009 StEB showing that 64 percent of Dutch citizens was willing to donate their organs after death – a finding supportive of her bill. In the European Parliament the EB is used as well, for instance in April 2020 as part of a COVID19 related question to the EC, wherein the consideration was: 'As shown in a special Eurobarometer from 2019, citizens expect the EU to ensure access to affordable energy and drastically reduce the number of families facing energy poverty.'[4] Finally, on numerous occasions EB data or references are used in media reports. In February 2020 the Flemish *Het Laatste Nieuws* reported on the low trust of Belgian citizens in their political institutions; according to *The Guardian* in October 2019 "Two-thirds of British people see overseas aid as 'a major priority'"; and in May 2020 Virginijus Sinkevičius, EU Commissioner for the Environment, Oceans and Fisheries, referred in the Dutch *NRC Handelsblad* to public support for his policy plans: "Look at the Eurobarometer: European citizens demand a better environmental protection, ask for action against climate change".[5]

All in all, the Eurobarometer is an authoritative and influential instrument for policy decisions, public debate and research on issues related to the EU and European integration. But here is the catch: the EB can only properly fulfil this function and be honored for it if it produces data of high quality. Noblesse oblige! However, there has been serious concern whether the EB in fact provides the 'solid indication of the concerns, needs and opinions of European citizens' that Wallström claimed, and about EB data quality on general (see for similar critical reflections e.g., Nissen, 2014; Höpner & Jurczyk, 2012, 2015; see Bläser, 2013 for a rebuttal).[6] Specifically, EB findings seem to be positively biased: "The Eurobarometer never predicts bad weather" (*de Volkskrant*, 19 January 2005). Is it possible that the EC, as the commissioner and funder of this expensive data collection project, has an interest in its outcomes and consciously or unconsciously pushes the EB in a particular direction? Has the EB become a policy and communication instrument more than a neutral instrument to gauge public

[4] https://www.europarl.europa.eu/doceo/document/E-9-2020-002331_EN.html

[5] See respectively https://www.hln.be/nieuws/buitenland/eurobarometer-vertrouwen-van-de-belgen-in-politieke-instellingen-ligt-laag~adf774c0; https://www.theguardian.com/global-development/2019/oct/23/two-thirds-of-british-people-see-overseas-aid-as-a-major-priority-eurobarometer; and https://www.nrc.nl/nieuws/2020/05/20/europese-burgers-vragen-om-betere-natuurbescherming-a4000326. Of course, many more examples of EB impact on the policy process and public opinion are available, both for EU member states and the EU.

[6] On occasion reference is made to such critical reflections, but only in general terms and while neglecting these critical reflections (see e.g., Brenner, 2016).

opinion (see e.g., Nissen, 2014), or even worse: an instrument for strategic manipulation of public opinion on EU matters (see e.g., Höpner and Jurczyk (2015: 5)?

The key question in this paper is whether or not and from a survey methodology perspective the EB is a sound measurement instrument. We assess various key aspects of questionnaire design, sample selection, nonresponse, and the possibility of longitudinal analyses. Our main conclusion is that the EB leaves much to be desired; there is ample room for improvement and transparency in particular. In the meantime, scholars using EB data should be aware of these methodological pitfalls. Admittedly, this may not be a revolutionary reflection or a truly original call for all EB users, but as long as the EB continues to dominate the field, our call is warranted and methodological warnings are in order. The EB is an impressive data institution, but the "continuously growing body of data is not an indicator of data quality per se" (Nissen, 2014: 716).

2. Methodological reflections on the Eurobarometer

The Code of Professional Ethics and Practices (November 2015) of the American Association for Public Opinion Research recommends: "Good professional practice imposes the obligation upon all public opinion and survey researchers to disclose sufficient information about how the research was conducted to allow for independent review and verification of research claims" (AAPOR, 2015: 4; see also e.g., Bethlehem, 2018: 233). For the EB, however, it is very hard to obtain insight into methodological choices and characteristics: the EB is a *black box* (Marcus, 2009). Information presented in general terms often only refers to: the fieldwork period; the number of completed interviews (per country); a sketch of the sampling design; the mode of data collection; the weighting adjustment procedure; and the margins of error. This documentation is relevant - but highly insufficient for establishing data quality. In our diagnosis we assess several aspects in order to present an informed idea of EB data quality. Section 2.1 is on sample selection and data collection. Section 2.2 is on nonresponse and weighting adjustment techniques. Section 2.3 indicates problems with the questionnaires and section 2.4 describes issues with the longitudinal nature of the EB. Section 3 contains concluding remarks.

2.1. Sample selection and data collection

It is a truism: in order to validly generalize outcomes of a sample-based survey to a target population, e.g., the population of the EU or one of the member states, a probability sample should be selected. Selection probabilities must be known and all units (most often: individual citizens) in the target population must have non-zero selection probabilities. If these conditions are met, unbiased estimates of population characteristics can be computed (for sampling theory see e.g., Kish, 1965; Cochran, 1977; Fink, 1995b; Bethlehem, 2009). Already in 1952 Horvitz and Thompson have shown that a random sample with each unit in the population having a non-zero probability of selection and with all probabilities known, results in unbiased estimators of population parameters.

The target population consists of the populations of the participating countries of 15 years and older.[7] However, it is not clear which people belong to the population. Moreover, Nissen (2014: 717) shows that the definition of the target population of EU citizens is not identical over the years. Until StEB 41 (1993) the population in a member state consisted of all citizens of that state, but from this date onwards the target population was re-defined as citizens who lived in any member state and who were citizens of any member state. Also, in some member states particular ethnic minorities were not

[7] The StEB is not a single survey but a group of surveys, i.e. a combination of separate surveys in participating countries. To be able to create a single dataset and compare the data for these countries, this assumes an identical or at least equivalent methodological approach as regards sampling and sample size, questionnaire design and question wording, fieldwork and modes of data collection, and weighting adjustment procedures in all participating countries. These assumptions are not met.

included in the population. All this makes it difficult (strictly speaking impossible) to draw a pure probability sample as well as to compare outcomes over time and over different EU member states.

Second, the EB documentation is too limited on how samples are selected in separate countries, making it hard to determine sample quality. Nevertheless, based on the publicly available information and some extra information solicited from the EB team, we have been able to take a closer look at the sampling design for the StEB, in particular for the Dutch case.

As said, ideally a simple random sample is drawn for the EU at large or in terms of procedure identical simple random samples for separate participating countries, but in practice such handbook wisdom is hard to follow. A more complex design has been used. First, stratification is applied in each country via the geographical EUROSTAT NUTS II classification. In the Netherlands stratification is by means of provinces; there are 12 strata corresponding to 12 provinces. Subsequently a separate sample is selected in each stratum, an (unbiased) estimate is computed per stratum, and finally all stratum estimates are combined into an (unbiased) estimate for the country as a whole.

Selecting a sample of units from a stratum is done in three phases:

(1) selecting sampling points: a sampling point is a cluster of postal codes; in the Netherlands sampling points correspond to municipalities. The sampling points are selected with probabilities proportional to size, i.e., the number of people.

(2) selecting addresses from selected sampling points: clusters of addresses are selected in each point. In many countries, addresses are chosen systematically within sampling points using a standard random route procedure: a starting address is selected at random and this starting address is the first address in a cluster, while the remainder of the cluster is selected as every *k*-th address obtained by means of a standard random route procedure. Note that the combined probability of selecting an element in the overall sample is obtained by multiplying the probability of selecting the sampling point by the probability of selecting an address in this point (an unequal probability sample). However, often the sample sizes of selecting addresses in sampling points are taken equal and in that case all addresses have the same selection probabilities (a so-called self-weighting sample). The latter is a convenient approach: it is easier to compute unbiased estimates (no weighting required) and the workload is more evenly spread over sampling points (municipalities). In the Netherlands addresses are selected in a slightly different way. In each selected municipality a postal address file is used that contains all addresses of private houses; a sample of addresses in obtained via a simple random sample from this file.

(3) selecting a person: finally, one person is selected at each selected address. To do this, an external company attempts to find a telephone number for each address; this company should be able to link telephone numbers to selected addresses. The telephone numbers are used to make contact. Once contact is established, a random person is selected at the address with the ‘first birthday method’: the interviewer asks the respondent to identify the household member (in the target population) whose birthday is next. This person is included in the sample.

This design for the StEB sample raises several issues. One issue is the use of the random route for selecting addresses. According to the available documentation standard random route is applied in most countries, but it is not clear what this exactly means. The procedure suggests first drawing a random starting address, but this requires a complete list of addresses for all selected sampling points; it is far from obvious that such lists are available. Moreover, after selecting their first address, interviewers have to follow a route through the neighbourhood and visit every *k*-th address, leaving ample room for interviewer’ discretion, or arbitrariness. How free are they in deciding where to go? What to do if there is no contact? Do they return later, and if so how many times? What if selected citizens refuse to participate?[8] Such challenges make it very surprising that in the end the realized sample size in almost every country and StEB surveys is about the required sample size of 1,000 persons, suggesting a ‘hunt for the last respondent’ (Stoop, 2005) until the target is reached. Moreover, this procedure essentially constitutes a form of substitution: nonresponding addresses and individuals are replaced by responding addresses and individuals, resulting in the selection of ‘low hanging fruit’ is’. Consequently the realized sample cannot be considered a valid representation of the target population and estimates will be biased; substitution should be discouraged (e.g., Kohler, 2007; Vehovar, 1999).

Second, the StEB sample is a cluster sample: geographical areas are selected and subsequently addresses and people within these areas. However, people in such clusters resemble each other more than people in different clusters, constituting a cluster effect (e.g., Fink, 1995b:16). Consequently, samples are ‘inefficient’ and compared to simple random sampling more respondents are required to obtain a specified level of precision. The impact of cluster effects can of course be reduced by selecting more clusters and less respondents per cluster, but this would make data collection more expensive. Unfortunately, the StEB lacks detailed information about the number of respondents per sampling point, making it impossible to estimate cluster effects. Even worse: the specification in StEB reports suggests that the samples are simple random samples - this is incorrect and potential harmful *fake* information.

Thirdly, in the final sampling phase a single person is sampled from each selected address. This means that selection probabilities depend on the number of people living at their address belonging to the target population: selection probabilities are highest for single person households and decrease with the increase of the number of people in the household. Estimators of population characteristics should be corrected for this bias; however, it is not clear whether such a correction takes place.

Finally, the documentation states that for the StEB face-to-face interviewing (CAPI) is applied for data collection *where possible*, but in how many countries is this the case? Moreover, there is no information on the alternative mode(s) of data collection if CAPI is impossible. It is an established fact that modes of data collection have impact: a respondent gives a different answer to the same question in a different data collection mode (e.g., Jans, 2008). Consequently, observed differences between EU member states may be partly real or partly results from mode effects.

The issues that we addressed so far were illustrated with the StEB, but other EB versions have similar problems. For instance, telephone interviewing is applied for the FEB, but there is limited information on sampling. The GESIS website states: ‘For each Flash Eurobarometer new and independent samples are drawn by random selection. Representativity is stated for the respective universe. Selection details (e.g. RDD, regions quotas, ...) for the total population (15+) and the youth surveys are not published.’[9] At best, this is obscure and incomplete.

Additional information about the sampling approach can be found with extra effort, however, for example in the technical specification of FEB 454 (EU report published in August 2017); fieldwork was conducted by TNS Opinion & Social. Here a form of Random Digit Dialing (RDD) likely was applied, with a set of seed numbers as the starting point of the selection procedure. Seed numbers are telephone numbers collected from respondents of previous probability surveys. Subsequently, a sample of new numbers was generated by replacing the final two digits of a seed number by two random digits, and the sample was stratified by region to obtain a good distribution over the country. TNS uses both landline and mobile telephone numbers – but how to create a stratification of mobile

[8] There is more attention to this issue in the section on nonresponse.
Note that the Dutch StEB approach (see main text) produces real random samples and avoids these problems of random route.

[9] https://www.gesis.org/en/eurobarometer-data-service/survey-series/flash-eb/sampling-fieldwork

numbers, as these numbers often do not contain geographical information? Also, some people have both a landline and a mobile telephone: their selection probability is twice as large compared to people with one type of telephone. It is not clear if and how the sample is corrected for these biases.

There is another caveat with increasing relevancy as regards telephone samples. Several countries have, under various names and in different forms, *Do Not Call Registers*, i.e., lists of telephone numbers whose users have indicated that they do not wish to receive sales and marketing calls. Examples are the *Do Not Call Registry* in the US, the *Telephone Preference Service* in the UK, the *Opt-Out Public Register* in Italy, and the *Bel-me-niet-Register* in the Netherlands.[10] A survey interview may not be a sale attempt, but the difference is not always obvious and one wonders whether it is effective to call to these numbers, as many people may not see the difference between various types of calls and callers. Unwanted calls raise anger and negatively impact on response and data quality. However, not using the do-not-call numbers results in a specific part of the target population being excluded from the survey, causing outcomes to be biased.[11]

All in all, drawing a random sample of telephone numbers is a major, probably insurmountable problem - and a recipe for major failures as the UK Polling Disaster of 2015 (see e.g., Sturgis et al., 2016; Bethlehem, 2018: 172). All telephone polls conducted during the final campaign days of the UK general election of 7 May 2015 were ‘wrong’, i.e., in predicting the election result. Research by the British Polling Council showed that the polls were flawed mainly because of low response rates: often less than 20 percent and in urban areas less than 10 percent. According to Martin Boon, director of ICM Research, about 30,000 call attempts had to be made to realize 2,000 interviews…[12]

2.2. Nonresponse and weighting adjustment

There are many factors affecting the quality of survey data as population estimates and unit nonresponse is one of the most infamous threats: people in the sample do not participate because of non-contact, refusal, the inability to cooperate, or any other reason. Unit nonresponse likely leads to biased estimates (e.g., Bethlehem, 2018: 131).

Nonresponse has two main consequences: (1) the net, realized sample is smaller than the initial or gross sample, leading to larger margins of error; (2) the net sample is selective due to relevant differences between respondents and nonrespondents. The latter problem is the most notorious one: it directly affects the validity of outcomes via biased estimates. This makes the response rate a key data quality indicator: “(…) current rules of thumb of good survey practice dictate striving for a high response rate as an indicator of the quality of all survey estimates” (Groves, 2006: 670). Consequently, response rates should always be reported. However, response rates and types of nonresponse are scarcely mentioned in EB reports. The term ‘nonresponse’ does not appear in Standard and Flash reports - as if there is no nonresponse; the contrary is true, although it is hard to estimate response rates.

[10] In 2013, the Dutch Do Not Call Register contained 8 million telephone numbers on a Dutch population consisting of 17 million people. See https://www.bel-me-niet.nl/nieuws/berichten/8-miljoen-inschrijvingen-bel-me-niet-register. The MOA, the Dutch association for market research, developed a Do Not Call Register specifically for market research. People in this register will not be contacted for polls and surveys. The number of people in this MOA register is about 110,000 (May 2019; personal communication).

[11] The EU General Data Protection Regulation (GDPR, as of May 2018) further complicates the use of telephones for collecting data, stipulating that organizations must have the explicit consent to be contacted before they can call. See https://www.contactspace.com/blog/do-not-call-register.

[12] See http://www.bbc.com/news/uk-politics-33228669

As said, sampling for the StEB is a multi-stage process, with the subsequent selection of sampling points, addresses, and respondents. In the first stage nonresponse is less likely since this essentially is an administrative activity, but nonresponse can and does occur in subsequent stages. Interviewers are instructed to contact addresses by telephone, select one person at random per address, and make an appointment for a face-to-face interview. However, the address information does not include a telephone number; this number has to be linked from another source, e.g., the telephone directory. Unfortunately, telephone directories do not contain all telephone numbers. A substantial part is missing: some people do not have a telephone, other numbers are secret and not registered. Moreover, telephone directories usually contain no mobile numbers: people with only a mobile phone will not be sampled. As an example, Beukenhorst (2012) describes the situation in the Netherlands where for only between 65 and 70 percent of the people a telephone number can be found – and it is very likely that this estimation is outdated and far too optimistic for the 2020s. In the most optimistic estimate, about sixty percent of the people can be contacted.[13] For selected telephone numbers a contact attempt is made, but often without any success. According to the EB team response in this so-called pre-screening phase of sampling is 'in single digits, i.e. below 10%'![14] This is a major source of bias.

The 'single digits' group that remains are persons with which an appointment can be made for a face-to-face interview. And since they agreed to participate, one would expect a high response, but substantial nonresponse occurs: in the Netherlands at this stage response is 70 to 75 percent.[15] Combining nonresponse in the three phases one can expect the average response probabilities to be about $0.60 \times 0.10 \times 0.75 = 0.045$, i.e., an overall response rate less than 5 percent. Needless to say that this is a major threat of the representativity of the realized sample.

Many countries apply a random route procedure to select addresses for the StEB. In that case nonresponse occurs if people are not at home (non-contact), do not want to participate (refusal), or do not speak the language (unable). It is not clear how often this happens and, more important, what action is undertaken in such cases of unit nonresponse. This makes is even more intriguing that the realized sample size in each country is almost always equal to the intended or target sample size, i.e., 1,000 (in most countries). Most likely sampling was continued until this sample size was reached. But since nonresponse is not documented, it is unclear whether and how response rates differ from country to country – a net sample of 1,000 respondents out of a gross sample of 2,000, 5,000, 10,000 or more?

It would be informative to have more detailed response information, in particular since response rates do differ for countries. In the European Social Survey (ESS), that similar to the EB is conducted in various European countries, response rates (ESS third wave; 2006-2007) are relatively high due to major (and expensive!) efforts to reach the target response rate of 70 percent, but vary from 46 (France) to 73 percent (Slovakia) (Stoop *et al.*, 2010: 93) and, moreover, show divergent country trends over time (Beullens *et al.*, 2018). At a lower level of overall response, similar differences must be expected for the EB, and differential response rates may increase nonresponse bias and complicate comparative analyses in public opinion: to what extent are differences caused by substantive differences between countries or by different response rates?

There are serious problems with nonresponse in the FEB as well. In general, response rates are extremely low in telephone surveys (see e.g., Rivers, 2007). According to the Pew Research Center (2012; see also e.g., Dutwin & Lavrakas, 2016) response rates of telephone surveys in the US were

[13] According to estimates of September 2020, 37 percent of Dutch citizens only have a mobile phone. See https://telecomnieuwsnet.wordpress.com/2020/09/16/rapport-vaste-telefonie-in-nederland-blijft-maar-dalen/

[14] Personal communication (e-mail) with the authors, 29 March 2019.

[15] Personal communication (e-mail) with the EB team and the authors, 28 January 2019. Later communication even reveals a response of 80 percent, without substantiation; we are very skeptical as regards the correctness of this response percentage.

around 9 percent in 2012. *YouGov* notes that response rates for telephone surveys in the UK have been declining to below 10 percent.[16] In general, in many countries there is a trend of decreasing response rates and nonresponse has become a serious global problem (see e.g., Bethlehem *et al*., 2011).

Nonresponse affects the representativity of surveys: estimators of population characteristics will be biased. To correct for this bias a weighting adjustment procedure should be carried out (e.g., Kalton & Floes-Cervantes, 2003; Bethlehem *et al*., 2011: chapter 8).[17] To be able to compute the adjustment weights, weight variables are required that must have been measured in the survey *and* of which the population distribution is available. Moreover, weighting can only be effective if the weight variables satisfy two conditions: they must be 1) (strongly) correlated with the response behavior of the people in the sample; and 2) (strongly) correlated with the target variables of the survey. Obviously, it is difficult to find such effective weight variables. Population distributions of relevant variables may be unknown and potential weight variables do not always satisfy both conditions. Standard available weight variables are demographic variables like gender, age and region, but if these variables are not strongly correlated with response behavior and target variables, weighting with these variables is ineffective. Indeed, the weight variables for the StEB are gender by age, region, and degree of urbanization, but it is very unlikely that correlations of this limited number of socio-demographic variables with response behavior and target variables are strong. Consequently, in the StEB adjustment weighting will not be effective to correct for nonresponse bias.[18]

2.3. The questionnaire

Opinions and attitudes can be measured in a valid and reliable way if questions are asked in a correct way (on questionnaire design, see e.g., Bradburn *et al*., 2004; Fink, 1995a; Fowler, 1995; Holyk, 2008; Schuman & Presser, 1981). For one thing, questions should be worded in a neutral, objective way and unintended leading questions should be avoided.[19] The specific wording of questions and response alternatives is already an issue in earlier critical assessments of the EB (e.g., Höpner & Jurczyk, 2012, 2015; Klein, 2012; Nissen, 2014). Höpner and Jurczyk (2012, 2015) show leading and otherwise technically incorrect questions that seem to follow a particular pattern: all violations of question design guidelines “systematically steer responses in a pro-European, integration-friendly direction. In fact, we did not find a single example in which the violations steered responses inversely” (Höpner & Jurczyk, 2015:18).

We present some extra examples to stress the point that EB questions leave much to desire. For instance, StEB 44.1 (1995) contains a question on the extension of the EU (see Example 1).[20] First, the question opens with “Some say…”, implying that it may be a good idea to extend the EU with Central and East European countries; the position “Other people say…” to balance the question, is not presented. Second, interviewers are instructed to read out three response alternatives - all three options favor extension. The fourth, less EU positive alternative is not explicitly offered and only recorded in case the respondent spontaneously insists on expressing this (op)position. Obviously, this question steers respondents in favor of extension.

[16] YouGov https://yougov.co.uk/about/panel-methodology/

[17] In essence: every respondent is assigned a weight with overrepresented respondents getting a weight smaller than 1 and those in underrepresented groups getting a weight larger than 1.

[18] The same goes for the weighting adjustment procedures applied to the FEB.

[19] A fictitious but convincing example of the impact of leading questions is given in an episode of the British sitcom ‘Yes Prime Minister’, where two versions of a questionnaire are used to come to contrary positions by the same respondent. See https://www.youtube.com/watch?v=G0ZZJXw4MTA .

[20] The example was taken from the British version of the questionnaire, but all versions of EB questionnaires contain these flaws: they are based on ‘mother questionnaires’.

Example 1. A leading question.
(source: Standard Eurobarometer 44.1, 1995).

Some say countries of Central and Eastern Europe, such as the Czech Republic, Hungary, Poland and Slovakia, should become member states of the European Union. What is your opinion on this? Should they become members …

□ In less than 5 years
□ In the next 5 to 10 years
□ In over 10 years time

□ I don't think these countries should become members of the European Union (SPONTANEOUS)

□ Don't know

Another dubious question is frequently included in EB surveys and is one of the key EB questions. The question gauges whether people think that the EU membership of their country is advantageous or not (see Example 2). One of the guidelines of question design is that response alternatives should be balanced: "A lot of wording effects research, as well as practical advice, focuses on *balance* – on trying to pose questions as objectively as possible" (Holleman, 2000: 4; italics in original). Negative response alternatives should mirror positive ones. This is not the case for the EB question on EU advantageousness: the second response option does not mirror the first, since it does not suggest that the EU can be *dis*advantageous.[21] Balancing this question would make a real difference, as Höpner and Jurczyk (2012) show with an experiment in which the German version of the question was compared to a version containing three substantive response alternatives. According to the EB question 49 percent stated that Germany benefited from its EU membership. In the experimental version only 22 percent had this opinion and 46 percent opted for the middle or neutral alternative (on middle positions, see e.g. Schuman & Presser, 1981: 161-178).

Example 2. An unbalanced set of response alternatives I.
(source: Standard Eurobarometer 73.4, 2010)

Taking everything into account, would you say that the UK has on balance benefited or not from being a member of the European Union?

□ Benefited
□ Not benefited
□ Don't know

[21] Arguably, three response alternatives would be in order to create balance: 'Benefited', 'Did not benefit and did not suffer', and 'Suffered'.

Example 3. An unbalanced set of response alternatives II.
(source: Standard Eurobarometer surveys 47.2 (1997) to 55.1, 2001)

Which of the following statements best describe(s) what the European Union means to you personally?

□ A way to create a better future for young people
□ A European government
□ The ability to go wherever I want in Europe
□ Guaranteed lasting peace in Europe
□ A means of improving the economic situation in Europe
□ A way to create jobs
□ A way to protect the rights of citizens
□ A lot of bureaucracy, a waste of time and money
□ Just a dream, a Utopian idea
□ The risk of losing our cultural diversity
□ Others
□ Don't know

Another example of a question with unbalanced response options is presented in Example 3.[22] First, there are more positive (7) than negative response alternatives (3), increasing the probability of a positive answer. A second issue is that the question may suffer from a primacy effect. Respondents have to select a response from a long list presented on a show card, but are often not motivated to read the full list: they stop after having read a few options, causing a 'preference' for options early in or high on the list, on the impression that "up means good" (Tourangeau *et al*., 2013). This primacy effect is a manifestation of *satisficing*. Respondents do not make the effort to give the correct, best answer, but opt for an acceptable answer that needs a minimal effort to pick (e.g., Krosnick & Alwin, 1997; Krosnick, 2000). All in all, the question on the relevancy of the EU contains a double imbalance, with more positive than negative options and all positive options in the first, highest part of the list. Both aspects increase the probability of selecting a positive option on what the EU means to the respondent personally.

Another problem that is neglected too often in EB studies relates to the inclusion of many and diverse questions in a questionnaire, assuming that people have an opinion on *all* topics. However, citizens may not be familiar with the topic or for various reasons do not have an opinion about it before they are confronted with the question (e.g., Zaller, 1992). Nevertheless, they most often answer the question by picking a 'real' option, because they do not want to admit they don't know the answer (e.g., van de Maat, 2019).

EB surveys contain questions on complex topics, such as the EU navigation system Galileo (see Example 4). First, respondents are asked whether the EU should set up its own system or rely on three different systems (American, Russian or Chinese). What if the respondent thinks that the EU should rely on the American system but not on the Russian and/or Chinese one? Likely the framing of the question partly resulted in 80 percent favoring the EU to develop its own system. The follow-up question refers to the Galileo project, without any explanation. And while it is highly unlikely that ordinary EU citizens have ever heard of the system, a substantial minority of 40 percent state that they know about the project – too good to be true. The essence of the Galileo project is subsequently revealed: it is a positioning system (what?) that the EU has started to develop. A majority of 63 percent is in favor of securing the necessary funds for it; it is intriguing that a majority wants to spend

[22] The question was included in StEB 47.2 (1997) to 55.1 (2001).

400 billion euro, while a minority knows about the system. If the answers of these questions are cross-classified, it turns out that 59 percent want to spend a lot of money on something they do not know anything about, i.e., not prior to the question on spending. This implies that not all respondents have given a ‘true’ opinion but they were pushed in this direction by question wording and response alternatives. Moreover, the third question suggests that the Galileo Project is not expensive; the costs are equivalent to the costs for (a lousy…) 400 km of highway. This all constitutes the formation of a positive opinion - or non-opinion - on the spot.

Example 4. A question on a complex topic.
(source: Flash Eurobarometer 211, 2007)

According to your opinion, should Europe set up its own navigation system, or should Europe rely on American, Russian or Chinese systems?	
□ The EU should set up its independent system.	80%
□ There is no need for an independent system.	12%
□ Don’t know.	8%
Have you already heard about the European Galileo project?	
□ Yes.	40%
□ No.	59%
□ Don’t know.	1%
Galileo is the name of the positioning system that the European Union has started to develop seven years ago. Currently, it seems that in order to complete the Galileo system additional public funding is necessary (about 2.4 billion €, which is the cost of about 400 km motorway). What do you prefer:	
□ The EU should secure the necessary funds in order to complete Galileo as soon as possible .	63%
□ The EU should not secure extra funds, even if it means that the project will be significantly delayed, or even that it fails.	23% 14%
□ Don’t know.	
Total	100%
Sample size	25,664

2.3. The longitudinal nature of the Eurobarometer

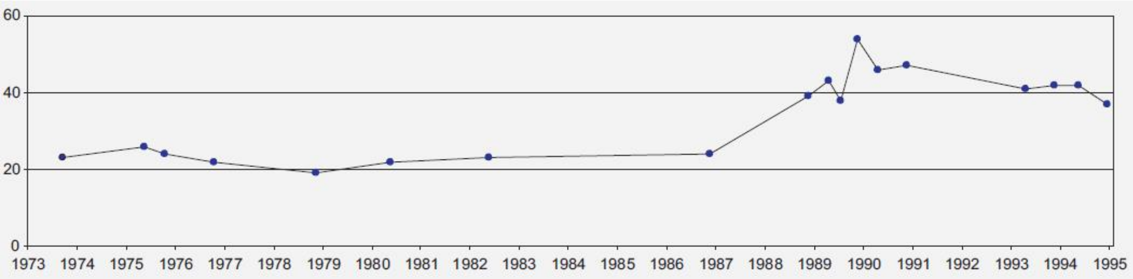
On the occasion of its 40th anniversary the European Commission (2014) stated that the “Eurobarometer has been surveying the views of Europeans since 1973 and gives a unique insight into how opinions and attitudes have changed over time”. Indeed, the series of EB surveys is impressive, but critical reflection is in order on its longitudinal nature, if only due to the ambiguity of the term. Longitudinal research basically comes in two flavors: (1) trend/cohort analysis, in which a new sample is selected for each wave; (2) panel analysis, in which a sample is selected only once and the same people are interviewed in each wave (e.g., Babbie, 2007: 102-106). For the analysis of individual level change of opinions, the panel approach is necessary; only if the sample remains the same, observed differences at the individual level can be considered ‘real differences’ (e.g., Hansen, 2008). The StEB, however, is not such a panel design. A new, fresh sample is drawn for every survey, making it impossible to measure individual opinion change over time. Observed changes in subsequent samples can be attributed to changes in the composition of samples; only changes larger than the margins of error of the samples can be interpreted as real changes, and even such changes can only be detected at the macro- and not at the micro-level.

Another complication for longitudinal comparative research is that the overall target population of the EB has changed over time. The EU as we know it started with six countries and counts 27 member states as of January 2020, after the formal completion of the Brexit process. Some countries are in the EB series from the start, other countries have data on a limited number of years. Researchers should be careful to have the same set of countries in every year of the period they study. Germany is another complication: before reunification in 1990, West Germany was the member state, while after the reunification this was Germany.

Whatever the form of longitudinal analysis, a crucial condition is that questions compared over time are identical. This is not so for EBs. The EC publication on the occasion of the 35th anniversary of the EB contains a telling example (see Example 5a). The graph suggests increasing interest in European affairs and the text supports this impression: “During the first waves of the Eurobarometer, from 1973 to 1986, approximately a quarter of European citizens declared that they were very interested in Community affairs. Interest levels increased gradually after 1986. In November 1988, the proportion of respondents who expressed an interest in European affairs exceeded 30% for the first time. In November 1989, it reached the historic level of 54%” (European Commission, 2008: 13). However, it is not clearly mentioned that question wording changed several times (see Example 5b).[23]

It is a truism that different question wordings result in different responses. There are numerous examples of even minor changes in wording leading to substantive differences in responses. Changing question wording consequently puts comparability at stake: it is not clear whether observed differences over time are ‘real’ differences or artefacts of question change. “Changes in the distribution of findings from one survey to the next may result from instrument changes” (Schmitt, 2003: 246).

Example 5a. European citizen’s interest in European affairs.
(source: European Commission (2008)



[23] Admittedly, the publication refers to this change - in a footnote in very small font size.

Example 5b. Changing question wording.
(source: various Standard EB surveys)

Year	Question text
1975	The press, newspapers, radio, television, often mention the European Community - the Common Market - as being a factor in the future of Britain and the other countries of Europe. Are you personally very interested, a little interested, or not at all interested in the problems of the European Community?
1976	Are you personally very interested, a little interested, or not at all interested in the problems of the European Community (The Common Market)?
1982	Newspapers, radio and TV often present news and commentaries about the European Community (also called the Common Market). Are you personally very interested, a little interested, or not at all interested in the problems of the European Community?
1993	To what extent would you say you are interested in European politics? That is to say matters related to the European Community: a great deal, to some extent not much or not at all?
2006	Would you say that you are very interested, fairly interested, not very interested or not at all interested in European affairs.

All in all, researchers creating time series and analyzing and presenting longitudinal trends on the basis of EB data should be warned. If there have been any change in question wording over time and if they detect changes, they should check for its consequences or at least refer to the possibility that observed and reported changes may not be substantive changes. Understandably the European Commission - and many scholars and policymakers - love the data generated by the EB and the longitudinal nature of the EB in particular, but we know that love is blind (for methodological issues).

3. Concluding remarks

The Eurobarometer is an esteemed institution, notwithstanding critical methodological remarks that can be made and of which the diverse community of users of EB data sets should be fully aware of. Obviously, similar remarks are likely in order with respect to other large scale surveys and public opinion polls, but the EB is arguably the dominant instrument in and outside Europe for collecting data and gauging public opinion on European affairs, the EU and European integration, and potential disintegration after the unprecedented Brexit process. *Noblesse oblige!*

More than anything, our reflection on several methodological aspects of this important data collection instrument or institution implies opportunities for improvement, with respect to data quality and questionnaire design and, more importantly, as regards its documentation and transparency (see also Schmitt, 2003: 248). The first step to be taken is to improve the methodological account of the design and practical conduct of the studies. Better documentation is needed on the way in which the surveys are set up and carried out, in order to allow scholars and others interested in and working with the huge body of collected data to assess data quality.

Simple tables and fancy figures are what the general public most likely sees as core products of the Eurobarometer. These are the result of a number of consecutive and related decisions in the research process. Such decisions and choices are necessary and inevitable and will not always be set in accordance to the ‘wisdom’ of methodological handbooks, but it is crucial to inform the users about these decisions in a comprehensive and transparent way. Only then it is possible to establish whether the outcomes are real and meaningful and if any development in public opinion on European affairs is primarily the result by changes in the design of the surveys, or whether they reflect substantive public opinion in Europe. This real and robust insight in citizens' opinions and attitudes on 'the European project' has always been the *raison d'être* of the Eurobarometer. In times when the European project seems to be threatened both from within and outside more than ever before, it is imperative to know what the citizens of this economic and political community themselves know, think and prefer. The Eurobarometer can be invaluable in this respect – if designed, conducted and presented in a proper way.

References

AAPOR (2015). *AAPOR Code of Ethics*. http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx

Anderson, C.J. & J.D. Hecht (2018). The preference for Europe: Public opinion about European integration since 1952, *European Union Politics*, 19/4, 617-638.

Babbie, E. (2007). *The Practice of Social Research*. Belmont: Thomson Higher Education. 11th ed.

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. New York: John Wiley & Sons.

Bethlehem, J. (2018). *Understanding Public Opinion Polls*. Boca Raton: CRC Press.

Bethlehem, J.G., F. Cobben & B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Beukenhorst, D. (2012), The Netherlands. in: Häder, S., Häder, M. & M. Kühne (eds.) *Telephone Surveys in Europe* (17-24). Heidelberg: Springer.

Beullens, K., G. Loosveldt, C. Vandenplas & I. Stoop (2018). Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?, *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=9673

Bläser, K.-A. (2013). Europa im Spiegel der öffentliche Meinung: Bilanz und Perspektiven des Eurobarometers nach 40 Jahren, *Leviathan*, 41/3, 351-357.

Bradburn, N., S. Sudman & B. Wansink (2004). *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass. Rev.ed.

Brenner, P.S. (2016). Research synthesis: Cross-national trends in religious service, *Public Opinion Quarterly*, 80/2, 563-583.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons. 3d ed.

Dutwin, D. & P. Lavrakas (2016). Trends in Telephone Outcomes, 2008 – 2015, *Survey Practice*, 9/3. doi: 10.29115/SP-2016-0017

European Commission (2008). *35 Years of Eurobarometer. European integration as seen by public opinion in the Member States of the European Union: 1973-2008*. http://ec.europa.eu/public_opinion/docs/35_years_en.pdf

European Commission (2014). *Effects of the economic and financial crisis on European public opinion*. 40 Years Eurobarometer. http://ec.europa.eu/public_opinion/topics/eb40years_en.pdf

European Union (2016). *Standard Eurobarometer 86 – Autumn 2016. Public opinion in the European Union: Report*. Brussels: European Commission, Directorate-General for Communication.

Fink, A. (1995a). *How to Ask Survey Questions*. Thousand Oaks: SAGE. The Survey Kit Vol. 2.

Fink, A. (1995b). *How to Sample in Surveys*. Thousand Oaks: SAGE. The Survey Kit Vol. 6.

Fowler, F.J., Jr. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks: SAGE. Applied Social Research Methods Series Vol. 38.

Gabel, M., S. Hix & G. Schneider (2002). Who is afraid of cumulative research? Improving data on EU politics, *European Union Politics*, 3/4, 481-500.

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys, *Public Opinion Quarterly*, 70/5, 646-675.

Hansen, J. (2008). Panel surveys, in: W. Donsbach, M.W. Traugott (eds.), *The SAGE Handbook of Public Opinion Research*. Los Angeles etc.: SAGE, 330-339.

Holleman, B. (2000). *The forbid/allow asymmetry: On the cognitive mechanisms underlying wording effects in surveys*. Amsterdam: Rodopi.

Holyk, G.G. (2008). Questionnaire Design, in: P.J. Lavrakas (ed.), *Encyclopedia of Survey Research Methods*. Thousand Oaks: SAGE, 657-659.

Höpner, M. & B. Jurczyk (2012). Kritik des Eurobarometers. Über die Verwischung der Grenze zwischen seriöser Demoskopie und interessengeleiter Propaganda, *Leviathan*, 40/3, 326-349.

Höpner, M. & B. Jurczyk (2015). *How the Eurobarometer Blurs the Line between Research and Propaganda*. Cologne: Max-Planck-Institut für Gesellschaftsforschung. MPIfG discussion paper 15/6.

Hornitz, D.G. & Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* 47/260, 663-685.

Jans, M. (2008). Mode Effects, in: P.J. Lavrakas (ed.), *Encyclopedia of Survey Research Methods*. Thousand Oaks: SAGE, 476-480.

Kalton, G. & I. Flores-Cervantes (2003). Weighting Methods, *Journal of Official Statistics*, 19/2, 81-97.

Kish, L. (1965), *Survey Sampling*. New York: John Wiley & Sons [in 1995 re-published in the Wiley Classics Library].

Klein, M. (2012). *Europäische Meinungsmacher - Wie man Umfrageforschung für seine Zwecke missbraucht*. http://www.sciencefiles.org/2012/03/06/europaische-meinungsmacher-wie-man-umfrageforschung-fur-seine-zwecke-missbraucht

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria, *Social Research Methods*, 1/2, 55-67.

Krosnick, J.A. (2000). The threat of satisficing in surveys: The shortcuts respondents take in answering questions, *Survey Methods Newsletter*, 20/1, 4-8.

Krosnick, J.A., D.F. Alwin (1987). An evaluation of a cognitive theory of response order effects in survey measurement, *Public Opinion Quarterly*, 51/2, 201-219.

Maat, J. van de (2019). *Public Opinion Without Opinion?!*. Leiden: Leiden University [Phd thesis].

Marcus, J. (2009). Der Einfluss von Erhebungsformen auf den Postmaterialismus-Index, *Methoden-Daten-Analysen*, 3/2, 137-166.

Nelsen, B.F., J.L. Guth & C.R. Fraser (2001). Does religion matter? Christianity and public support for the European Union, *European Union Politics*, 2/2, 191-217.

Nissen, S. (2014). The Eurobarometer and the process of European integration: Methodological foundations and weaknesses of the largest European survey, *Quality & Quantity*, 48/2, 713-727.

Pew Research Center (2012). *Assessing the Representativeness of Public Opinion Surveys*. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Rivers, D. (2007). *Sampling for Web Surveys*. Paper presented at the Joint Statistical Meetings, Section on Survey Research Methods. Salt Lake City, Utah.

Schmitt, H. (2003). The Eurobarometers: Their evolution, obvious merits, and ways to add value to them, *European Union Politics*, 4/2, 243-251.

Schuman, H., S. Presser (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego etc.: Academic Press.

Signorelli, S. (2012). *The EU and Public Opinions: A Love-Hate-Relationship?* Paris: Notre Europe - Jacques Delors Institute. Studies & Reports 93.

Stoop, I.A.L. (2005). *The Hunt for the Last Respondent: Nonresponse in sample surveys*. The Hague: The Netherlands Institute of Social Research.

Stoop, I., J. Billiet, A. Koch, & R. Fitzgerald (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester: John Wiley & Sons.

Sturgis, P. Baker, N. Callegaro, M. Fisher, S. Green, J. Jennings, W. Kuha, J. Lauderdale, B. & Smith, P. (2016). *Report of the Inquiry into the 2015 British general election opinion polls*. London: Market Research Society and British Polling Council.

Tourangeau, R., M.P. Couper, & F.G. Conrad (2013). “Up Means Good”: The Effect of Screen Position on Evaluative Ratings in Web Surveys, *Public Opinion Quarterly*, 77/S1, 69-88.

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, 15/2, 335-350.

Vreese, C.H. de, H.G. Boomgaarden & H.A. Semetko (2008). Hard and soft public support for Turkish membership in the EU, *European Union Politics*, 9/4, 511-530.

Zaller, J.R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.