CSES ROUND 6

MEMO

Subcommittee on methods and technical developments

Wouter van der Brug, Orit Kedar, Laura Stephenson & Eric Yu

Executive summary

The CSES aims to enable cross-national comparative research into electoral processes. For that purpose, it collects survey data based on a shared questionnaire among representative samples of members of the electorates in each of the countries that participates in the project and appends them with macro-level and contextual data. High quality, combined with high coverage are the principles guiding the praxis of this path-breaking enterprise.

Given the focus of the project on voter decision making, the surveys are conducted after the election took place. For a long time, the golden standard for collecting these post-election data has been to conduct a face-to-face survey among a randomly drawn sample of eligible voters. Up until this point, the CSES has only accepted studies that have been gathered based on a probability sample of the population due to concerns about undercoverage of specific segments of the population, and because any statistical inferences about the population from which the sample is drawn assumes probability sampling. This policy is consistent with AAPOR statements about non-probability samples (Baker et al. 2010, Baker et al. 2013). In addition, some opt-in internet panels used for non-probability samples are of dubious quality, which adds to the reluctance to open the door to such studies.

We hold, however, that because of changes in the industry, it is time to take stock and re-evaluate the ways to currently obtain high-quality, high-coverage data. We do not propose to re-open discussions on the *modes of interviewing*. While it is well-known that such mode-effects exist, there is no strong empirical evidence suggesting that one mode provides systematically more valid survey data than other modes. The CSES Round 5 subcommittee on Methods advised to allow for multiple modes of interviewing and we follow this advice. Our discussion focuses primarily on *sampling procedures*. We note that consistently declining and uneven response rate and coverage issues (e.g. related to cell phones vs. landlines), among other factors, have become major concerns that lead to varying quality of studies based on probability samples. At the same time we note that it is possible to gather high-quality non-probability data via internet studies in some countries and with some firms.

In view of these developments, we make three proposals for how the CSES should proceed:

1. Since no cross-country comparable standards exist for evaluating the quality of non-probability samples, we propose to maintain the structure of accepting only probability samples in Round 6 of the CSES. This means that only studies that are based on

probability samples of persons or households should be included. Similarly, we agree with the previous committee that all modes of interviewing are acceptable.

2. We think it is urgent to continue following the developments in the survey industry, as well as the academic research on the quality of non-probability studies. We think it should be an objective of the CSES to pursue information about how to evaluate the quality and representativeness of all surveys, probability and non-probability, with the goal that we might be able to re-evaluate the criteria for inclusion in the CSES datasets in light of technological changes in surveying. We recommend that the CSES planning committee charge the next methodology committee with this task.
3. In the meantime, we recommend that the CSES develops a separate repository of 'supplementary files' that provides access to election studies that have fielded substantial parts of the CSES modules, but that are not based on a probability sampling frame of the population. This includes opt-in internet panels as well as surveys with a severely limited geographical coverage (excluding more than half of the eligible population from the sampling frame). It could also host studies that are not included in the official CSES dataset because the study does not include the full module. The CSES secretariat will not be asked to function as gatekeepers of this repository (except for ensuring that all required documentation is provided), so the decision to make a study available via the repository lies with the PI. PIs must provide adequate documentation (more on this below) to be included in this section. Users will then be able to decide for themselves whether to include some of these datasets in their research. Yet, to be clear, such datasets will not be included in the official CSES data file for Module 6.

Introduction

The CSES is a large, path-breaking, multinational project that recently celebrated its 25th anniversary. The goal of the CSES is to enable cross-national comparative research into electoral processes. For that purpose, it gathers post-election survey data from representative samples of the electorates in countries around the world. Since the project began, the number of countries included has increased from 33 countries to 39 in Module 4 (Module 5 has yet to conclude).

In this memo we address the issue of how to best get high-quality, high-coverage data given changes in the polling industry in recent years. The main question this creates is whether the CSES should revisit its position on non-probability samples. We see ourselves as building upon the excellent work conducted by the previous subcommittee and the recommendations made in their report for round 5:
http://www.cses.org/plancom/module5/CSES5_NewTechnologySubcommittee_FinalReport.pdf
We think that the recommendations made by the previous subcommittee are still valid and for the most part we see no reason to revisit most of their discussions. However, as a consequence of societal changes and technological developments, the field of survey research is rapidly changing. In particular, survey research (including academic research) is increasingly being conducted by drawing samples from large internet panels because of the ease, speed and dominance of online communication. Some of these internet survey panels are based on a probability sample of households; surveys drawn from these meet the current criteria for inclusion in the CSES. Others are based on a pool of respondents who opt-in voluntarily and from which samples are drawn based on specific demographics. These non-probability samples do not allow one to draw inferences about the wider population, which is a good argument for excluding them from the CSES. At the same time, one should keep in mind that the quality of opt-in internet panels is improving in many countries, and more and more sophisticated methods are being developed to better approximate the sampling techniques of probability samples. Based on data from a decade ago, Ansolabehere and Schaffner (2014) concluded that a web opt-in sample performed just as well as telephone or mail surveys based on probability sampling, when compared to benchmark values based on population statistics. This is a new finding compared to earlier work that noted the comparability of non-probability studies for assessing correlational relationships, such as vote choice (Stephenson and Crête 2011; Sanders et al. 2007; Bytzek and Bieber 2016). More recently, Dassonneville et al. (2020), using data from Belgian face-to-face probability and online non-probability studies, noted that accurately estimating descriptive statistics remains a challenge for non-probability surveys while, again, vote choice models perform just as well as they do with probability samples.

While non-probability surveys are modernizing and improving in quality, more traditional survey designs are becoming more of a challenge due to decreasing response rates and undercoverage in sampling frames that jeopardize the potential to draw inferences about the population (Sala and Lillini 2015; Brick 2011). Needless to say, these increasingly thorny issues affect costs. And while cost in itself is not a scientific concern, the increasing challenges that in practice translate to higher costs lead to a situation in which the principle of high-quality, high-

coverage data is harder to fulfil using traditional methods. Breton et al. (2017), for example, find that a non-probability internet sample was better able to achieve benchmark estimates compared to a landline-only probability phone survey, and that the phone survey was subject to both self-selection and social desirability bias. So, in view of the rapid changes in the industry, we wish to open up the discussion on the advantages and disadvantages of including (certain) election studies based on non-probability samples in the CSES.

Background

Ideally, large cross-country comparative studies, such as the CSES, the WVS, the EVS, and the EES would like to maximize their standards on three separate goals: 1) cross-national comparability 2) high quality and 3) wide coverage. While the first two goals are in several ways related, there is a tension between these goals and third one. When being less restrictive in setting standards for comparability and quality, more country studies can be included. Yet, if this means that some studies are not really comparable to others in the dataset due to variation in quality, the goal of the project is compromised. Different cross-national studies have dealt with this trade-off in different ways. Some of the most centralized and well-funded studies, like the ESS, are the most restrictive, for instance demanding that all studies use the same mode (face to face) and have specific designs for sampling and thresholds for acceptable response rates (even though the latter are not always strictly enforced). Other studies, like the WVS, emphasize cross-country coverage. The CSES has managed to beautifully saddle these decisions. Although the CSES does not compromise its scientific quality standards, it has managed to cautiously pursue a more liberal acceptance policy compared to the ESS.

For a long time, the gold standard for collecting CSES survey data has been face-to-face surveys among a randomly-drawn sample of eligible voters. When a random sample of individuals could not be drawn, a representative sample could be obtained by sampling households. This was the 'norm' in the first rounds of the CSES, and is noted in the instructions provided to collaborators: "Interviews should be conducted face-to-face, unless local circumstances dictate that telephone, Internet, or mail surveys will produce higher quality data." The CSES has long-recognized that local circumstances may call for creative sampling choices to gather high-quality data. This is evident in its acceptance of different modes, mixed modes, and even panels as necessary. Mode of interviewing, for example, has been extensively studied and found to have an effect on non-response and social desirability (Couper 2011), as well as openness to experience and political attitudes (Valentino et al. 2020), among other findings. We see the CSES's recognition that no single mode of interviewing is the superior way to get high-quality, high-coverage data as a testimony to its innovative approach. At the same time, one cannot escape a fact that comes with this nuanced approach: the inclusion of different modes makes cross-country and over-time comparability lower than if the same mode of interviewing had been conducted. Tables 1 and 2 provide an overview of the modes used in different rounds of the CSES.

Table 1: Modes used in CSES Rounds

	Round 1	Round 2	Round 3	Round 4	Round 5
Face to face	69,2%	70,7%	72,0%	64,4%	38,7%
Telephone	10,3%	9,8%	16,0%	8,9%	19,4%
Internet	0,0%	0,0%	0,0%	0,0%	3,2%
Mail	7,7%	7,3%	4,0%	0,0%	0,0%
Multiple modes	12,8%	12,2%	8,0%	26,7%	38,7%
Total	100,0%	100,0%	100,0%	100,0%	100,0%

*: based on the surveys included until September 2021. New election studies are still being added

Table 2: Modes used, in full or in part

	Round 1	Round 2	Round 3	Round 4	Round 5*
Face to face	76,9%	78,0%	76,0%	75,6%	48,4%
Telephone	20,5%	14,6%	20,0%	20,0%	35,5%
Internet	0,0%	0,0%	0,0%	11,1%	38,7%
Mail	15,4%	19,5%	12,0%	20,0%	19,4%
Total	112,8%	112,2%	108,0%	126,7%	141,9%

*: based on the surveys included until September 2021. New election studies are still being added to round 5.

Table 1 shows the modes used in the different rounds, including multi-mode as a separate category. In Table 2, the modes that are used in the multi-mode studies are included in the counts, so that the numbers add up to more than 100%. The tables show face to face and telephone used to be the norm in the first three rounds. Since then, there is an increase in the number of multi-mode studies, mainly at the expense of face to face. While internet based studies are hardly ever used as the only mode (see table 1), this mode is increasingly popular in combination with other modes. Even though face to face and telephone surveys remain to be important, these two ways of doing fieldwork are becoming increasingly difficult as a result of societal and technological developments. There is a continuing trend in the numbers of non-responses in face to face surveys, particularly as a result of non-contacts (De Leeuw et al. 2018). Telephone surveys are even more problematic, particularly because many people have switched from landlines to using cell phones. Even if cell phones can be included in the sampling frame, one is likely to contact people at inconvenient moments, so response rates tend to be extremely low, and they are often not clearly reported.

Low response rates are not necessarily problematic, as long as respondents are not systematically different from non-respondents. However, this is certainly not the case. Certain groups of citizens are much more likely to participate in surveys than others. Young men in urban areas, ethnic minority groups and people over 75 years old tend to be underrepresented. They are difficult to reach and if contact has been established the former two groups are not eager to participate. Without measures to reduce bias, middle-aged women in rural areas will be substantially overrepresented. These biases are not just problematic for those who wish to describe population statistics, but also for estimating relationships between variables. Drawing on University of Michigan's Survey of Consumers, Heffetz and Rabin (2013), for instance, demonstrate that the gender differences in happiness are reversed between easy-to-reach and hard-to-reach respondents. Thus, the number of repeated attempts to reach a respondent may affect our results. Similarly, in a separate study, Heffetz and Reeves (2019) show that the rate of obesity, labour force participation, and household expenditures vary by number of attempts to reach respondents either across or within demographic categories.

Measures to reduce bias often take the form of offering payments to respondents who initially refused, over-representing specific hard to reach groups in the original sample, or hiring extra interviewers to reach these types of respondents. In the last round of the ANES (2016) the response rate was 50% for the pre-election wave, of whom 90% participated in the post-election survey. To reach this 45% response rate, respondents were paid up to $100 in the last round of refusal conversion. In the Dutch National Election Studies of 2017, which consisted of a post-election study only, there were similar experiences. In the third and last round of refusal conversion, respondents were offered 40 euro in cash for participating, which is substantially more than the interviewers make. Still, the response rate for the face to face part of the survey was only 48.8%. Three decades ago, response rates of 70% were often obtained with fewer measures. As a consequence, face to face surveys with fresh random samples of the population are becoming prohibitively expensive, while the samples are increasingly biased. Mail surveys

are a much cheaper alternative, and can be combined with self-completion via an online questionnaire. When some remuneration is offered, these studies can reach response rates of about 40 percent (Rekker et al. 2020).

The latest developments: internet panels

In countries where the majority of the population has access to the internet, online surveys are becoming increasingly popular. Even if respondents receive a fee for completing a survey, the efficiency (in terms of time and cost) remains much higher than for any kind of survey that is conducted by live interviewers. Because the survey industry is switching to this way of data collection in many democracies around the globe, it is important to discuss which types of internet surveys, if any, would be acceptable to include in the CSES.

When discussing the quality of data obtained by means of internet panels, we should distinguish between mode effects and sampling effects. Since the CSES has decided to accept all modes of interviewing, we do not address these types of effects here. As noted above in Table 1, previous rounds of the CSES have included data derived from self-completed questionnaires, either in the form of drop-off questionnaires (after a face to face component), or in the form of a mail survey. An advantage of self-completion is that there are no interviewer effects, so that social desirability plays a lesser role. In particular with more sensitive questions, this may even be an advantage. There are also mode consequences for how and what types of questions can be presented to respondents, most importantly in comparison to telephone surveying.

The bigger issue with internet surveys, and the one we focus on here, is the sampling procedure. As inferential statistics rely upon assumptions about the relationship of a sample to the population, this is of fundamental concern for being able to conduct quality research about a population. Indeed, how and whether sampling issues with online surveys can be mitigated is a focus of Baker et al.'s (2013) report for the AAPOR and Pasek (2016).

Surveys conducted on the internet are not always non-probability surveys. For one, it is possible to recruit individuals through other means to take surveys that are administered online (such as through mail or RDD telephone recruitment, for example). Second, established internet panels can facilitate probability surveys if they have been developed through intentional sampling strategies. In a few countries, internet panels exist that are based on a random sample of households. The Dutch LISS-panel at the University of Tilburg is one, and it was set up by a grant from the Dutch Science Foundation (NWO) for the purpose of conducting scientific research. Comparisons of data collected through this existing panel and data based upon two fresh random samples (one interviewed online and one face to face) show that the main differences are due to mode effects (self-completion vs face to face) rather than the sample itself or the fact that people are members of a panel and are thus 'experienced respondents' (Rekker et al. 2020). In line with the current standards followed by the CSES and the previous subcommittee, we recommend that an online survey collected from a sample of an internet panel that is randomly selected from the target population be acceptable for inclusion in the CSES dataset.

However, most internet surveys cannot claim such clear sampling logic. This is true of both methods that are typically employed in non-probability internet samples: river sampling and online panels. River sampling means that respondents are recruited by asking them to click on a web link that is placed on websites, which potential respondents visit for other purposes. Obviously, this way of sampling leads to all kinds of biases, resulting from the fact that those who visit the websites are not usually a cross section of the target population and the willingness to click on a link is biased towards those interested in the topic. The second method is to draw a sample from an internet panel. The participants in the original panel are often (partially) self-selected. Survey bureaus that maintain the panel try to correct for biases by drawing samples from this panel based on background characteristics. Non-probability internet studies do not have a good reputation, mainly because the representativeness of the sample can be questionable. Yet, of the two methods, river sampling seems to be the most problematic, while sampling from a larger panel looks more promising (e.g., Lehdonvirta et al. 2020). However, there is substantial variation in how firms manage their panels. Some are quite diligent, consistently monitoring for attrition and updating contact information and targeting recruitment to achieve certain representativeness goals based on official statistics. The degree to which a firm undertakes such activities is often reported in the form of responses to a set of questions established by ESOMAR (https://www.esomar.org/what-we-do/code-guidelines/28-questions-to-help-buyers-of-online-samples).

In order for the CSES to achieve its objective of gathering high-quality surveys that can facilitate comparative analysis, we believe it remains vitally important that the data included in the CSES modules are based on representative samples of the electorate. In 2013 the American Association for Public Opinion Research (AAPOR) report on the use of non-probability samples, one of the main concerns detailed is that non-probability samples lack an underlying theory for drawing inferences about the population sampled (Baker et al. 2013). It is an obvious truth that a randomly drawn sample from a registry of eligible voters is an ideal way of producing a representative sample, as long as respondents and non-respondents are not systematically different. For online studies, this can be a significant challenge.[1]

[1] While not the focus of this report, we do want to recognize that it is becoming increasingly true that few of the election studies that are based on probability sampling and that are currently included in the CSES datasets realize this ideal. First, for face to face surveys, most countries do not have a registry of eligible voters from which a sample can be drawn. Therefore, many probability samples are drawn from a sample of addresses, rather than individuals. Further, these addresses often do not distinguish between private homes, second homes and businesses. Moreover, geographical circumstances often limit the areas that are included in the samples. Similarly, modern practices for conducting telephone surveys are not truly RDD, in that lists of active and non-business numbers are often used as the sampling frames for both landline and cellphones. The particular mix of landline and cellphone surveys can also introduce bias. Second, probability surveys are experiencing increasing difficulties in achieving high response rates. This is challenging because non-respondents differ systematically from respondents in several ways; for instance, non-respondents tend to be less-educated and less likely to participate in elections than respondents. The fact that studies in the existing CSES datasets may not be perfectly representative does not mean that we should drop the requirement that the samples should be representative. Yet, it does prompt the question of

When internet panels started to become popular in Western countries, very few people had access to the internet; on that basis, any internet survey conducted with a panel would be automatically very biased. Nowadays, however, the percentage of households with access to the internet ranges in the European Union from 67% in Bulgaria to 98% in the Netherlands. This is very similar to the percentage of households that had a telephone connection 20 years ago, when telephone surveys were an acceptable standard. It is important to recognize that, in many postindustrial countries, (the lack of) access to the internet on its own does not introduce a bias into the representativeness of online samples. The main source of bias in non-probability online surveys is self-recruitment. Two main approaches have been used to address this challenge: conducting some sort of intentional sampling, and weighting. There is an extensive discussion in Baker et al. (2013) about these techniques for improving the representativeness of online non-probability surveys.

Intentional Sampling of Online Panels

The extent to which self-recruitment in online surveys is a problem depends in part upon how the survey firms draw their samples from their larger pool of self-recruited potential respondents. Typical existing internet panels of commercial survey companies consist of a large pool of individuals that have indicated their willingness to participate in surveys. Quite often this pool consists of more than 100,000 households that have been recruited or that have volunteered to be included in their panel. If survey institutes were to simply randomly select a sample of respondents from this pool they run the risk of introducing all kinds of biases into the sample, unless the panel itself is perfectly representative of the broader population (such as is achieved in probability online panels). The common practice for survey firms is to draw a sample which is stratified on the basis of several relevant background characteristics of which the distribution in the population is known. Quota sampling in this way produces samples that are often more representative of the population on the basis of specific background characteristics than random samples. Of course, concerns remain: we do not know whether the low educated 55 year old carpenter, who volunteered to be included in an online survey panel, is representative of the other 55 year old blue collar workers in the population. This problem is quite similar to the concerns about non-response bias; in a random sample with a 45% response rate we have similar concerns about the representativeness of the specific individuals who were willing to participate. Even if the respondents are representative of the target population in terms of certain background characteristics, we do not know how different the respondents are from the non-respondents.

A more sophisticated solution to this is the matching procedure developed by Rivers (2006). This procedure involves drawing a random sample from a large, pre-existing probability survey. Several characteristics of the individuals in that sample, usually going beyond basic demographics like age, gender and education, are identified and used to establish targets for a

---

whether we could/should rethink our traditional approach to representativeness and whether we can think of criteria to evaluate representativeness.

desirable sample in an internet panel. This procedure is more elaborate than simple quota sampling and acknowledges the importance of representativeness in terms of attitudes as well as demographics.

Weighting

The second strategy that can be used to improve the representativeness of an internet survey is to apply post-estimation weights. This practice involves determining the importance of specific respondents in a sample according to their frequency distribution in the target population. Good-quality official statistics are typically used for developing these sorts of weights, such as those in a census. The number of factors used in the weights can vary significantly, from a couple of demographics to a longer set of respondent characteristics that can be reliably benchmarked. Some studies even weight to correct the proportion of supporters of different parties, while others weight to correct for the proportion of non-voters in a sample. There are several different statistical techniques that have been developed to establish such weights, such as iterative proportional fitting (also known as raking).

Typically, researchers will weight their respondents in such a way that the sample mirrors the population. If 10% of the sample did not vote, and 30% of the population did not vote, we give the non-voters a weight of 3 and the voters a weight of 0.78 and then the sample mirrors the population in terms of turnout. However, by creating the weights in this way, we assume that there are no systematic differences between the respondents and the non-respondents who did not vote in the last election. It is important to realize that this is a theoretical assumption that we usually cannot test. How the attitudes of people who never participate in research differs from the attitudes of respondents is thus unknown, so that we do not know how biased our samples are, nor whether we can correct for these biases by weighting.

An approach to partially address this concern is propensity score adjustment (Terhanian and Bremer 2000). This practice involves identifying a sample characteristic(s) that is orthogonal to any variables of interest and that has an official or reliable benchmark. Propensity scores can be included as part of weighting procedures to improve the chances that the sample does more than just mimic demographics of the target population.

But do adjustments work?

The proof of the pudding is in the eating. In the Netherlands, all major survey companies publish election polls of different institutes, some based on fresh randomly drawn household samples and some based on volunteer internet panels. The differences are within the margin of sampling error. This suggests that, at least in the Dutch context, a self-recruited sample can provide a distribution of vote intentions that comes pretty close to that of a representative sample. In the American context, Ansolabehere and Schaffner (2014) provide a systematic comparison of the results of a web opt-in sample with those of exit polls and telephone or mail surveys based on probability sampling. Regarding the CCES (their opt-in web survey), these authors conclude: “In 2006, 2008, and 2010, there is no evidence of systematic bias in the projected election outcomes arising from the CCES samples. More striking still, the CCES yields nearly identical summary statistics and regression estimates for models of turnout and vote

choice as did the ANES, and the CCES is statistically indistinguishable from exit poll results on key indicators of vote choice and its correlation with basic demographics." (p. 326). While we realise that the situation may be different in other countries, we have to closely follow the developments in the field of survey research and see whether more evidence becomes available that online opt-in surveys can produce data of the same quality as more traditional approaches. If that should be the case, the CSES should be willing to accept these data.

When the previous committee advised not to include studies based on self-selected internet panels, it stated that "even if within a country a strategy for approximating findings from non-probability surveys to those of surveys based on random-sampling would be developed, its traveling capacity to other countries (with different conditions for survey research) would need to be additionally ascertained in the context of an internationally comparative study like CSES" (page 3 of its report). Like the previous committee, we do not support the inclusion of non-probability surveys in the CSES dataset without further information about their quality, and existing research does not provide any clear direction on this point. For the most part, the body of evidence tends to be mixed and the default position taken by many, unsurprisingly, is that non-probability surveys cannot be seen as equivalent to probability surveys. However, given the rising challenges of conducting high-quality probability surveys on the one hand, and the proliferation of non-probability surveys of increasingly high quality on the other hand, we believe that a universal exclusion of the latter is becoming increasingly less viable/practical. The industry has changed dramatically over the past five years. We thus suggest that the CSES takes action to set guidelines and conditions for including such surveys.

On this count, Baker et al. (2013) and Baker et al. (2016) are extremely helpful. Both suggest several elements to be considered when evaluating the quality of non-probability surveys. Unfortunately, the industry has yet to establish a threshold beyond which a non-probability survey might be considered "good". Baker et al. (2013) suggest that surveys should be thought of on a continuum of their ability to achieve accurate estimates in this regard. For this reason, we recommend that the CSES should not accept non-probability studies, at least not until a more generally accepted set of quality criteria can be agreed upon. We believe that the next methods subcommittee should undertake this task in consultation with survey design experts along with a re-evaluation of the dynamic survey industry.

At the same time, we do think that it is important to recognize that non-probability surveys may be quite appropriate for use depending upon the intent of the researcher. Baker et al. (2013) discuss this as "fit for purpose" and "fit for use". They note the distinction that Groves (1989) makes between "describers" and "modellers". As discussed above, most studies that have compared non-probability surveys with probability surveys for understanding vote choice have not found evidence that significantly different conclusions would be reached with either sample (Sanders et al. 2007; Stephenson and Crête 2011; Dassonneville et al. 2020). Thus we believe that the CSES should recognize that in some situations, having access to a broader set of responses to CSES module questions that cover more countries could be beneficial for research. To this end, we recommend the development of a separate repository, located on the CSES

website, of non-probability surveys that ran CSES modules *and* provide detailed information about the conduct and quality of the study. Since no accepted quality standards exist for non-probability samples, we do not make recommendations as to which non-probability based studies can be included in this repository and which cannot be included. Yet, we think it should be made clear to the PIs that they should only deposit datasets based on samples that they believe to be "representative of the population eligible to vote". By this we mean that relevant subsections of the population, in terms of geographical spread, demographics and electoral behaviour, are represented in the dataset in similar proportions as in the population at large, and that weight variables needed to correct biases (provided in the dataset) are of similar magnitudes as those needed in probability samples. We believe PIs can be trusted to upload good quality datasets, because they are normally interested in doing good research and because submitting a low-quality dataset will be harmful to their reputation. *It is essential that these supplementary files are submitted with easily accessible and lucid documentation that enables users to evaluate whether the datasets are suitable for their research purposes.* We offer recommendations as to the documentation that should be provided in the next section.

Standards to evaluate datasets

As is clear from this report and more explicitly discussed in Baker et al. (2013), all non-probability studies are not equal. Therefore, we propose that the CSES develops a set of questions that must be addressed by PIs before a non-probability survey will be accepted for the repository. We note that the practice of gathering information about survey methodology is well-established in the CSES community. There is an extensive questionnaire that must be completed by PIs before submitting any study for the CSES. It includes questions about sampling frame coverage, exclusions, response rate, etc. In practice, the CSES accepts post-election studies that use probability sampling and that cover most of the populated geographical areas of a country. It is acceptable to exclude some areas from the initial sampling frame, but this needs to be limited. There is no strict rule regarding the non-response rate.

For the CSES to include in its repository non-probability surveys as part of what it offers to the academic community (but notably, still separate from the surveys included in the official dataset), we believe that two things are crucial. First, since we ask the PIs to act as gate keepers, we must provide them with clear instructions about the quality of the study that they are depositing. PIs should be instructed to only deposit datasets based on samples that they believe to be representative of the voting population. Second, PIs should provide adequate documentation. When submitting surveys based on samples from a non-probability internet panel, PIs should be transparent about the procedures by which the panel was recruited and how respondents were selected from the sample (see also Baker et al., 2013, 2016). Therefore, we recommend that documentation is provided regarding the following topics:

1. Sampling frame (note, appropriate questions are already part of the standard CSES questionnaire):
    a. How the full panel was initially recruited and how it is maintained

    b. The procedures used to draw a sample from the larger panel
2. Compensation paid to respondents (note, this is already part of the standard CSES questionnaire)
3. Steps taken to ensure the quality of responses
    a. The time each respondent took to complete the questionnaire (and whether speeders were excluded)
    b. Checks to ensure unique responses
    c. Checks for attention

The specific questions included in Baker et al. (2016) provide a solid basis for modifying the existing CSES questionnaire in this regard. In the appendix, we provide a draft of a modified questionnaire that we believe could be used to gather details on probability and non-probability studies.

Moreover, given the above discussion of the representativeness of the survey datasets, we believe it would be worthwhile for the CSES to encourage all PIs, from probability and non-probability studies, to provide a variety of weights for their studies. Currently, the CSES datasets contain a sampling design weight variable, a demographic weight variable (weighting respondents on things like age, gender, etc) and a political weight variable (weighing respondents by their electoral behaviour), which can be used to ensure that the sample reflects the population on certain known parameters. For all studies, but non probability studies in particular, very high weights (say, over 7, although a standard threshold does not seem to exist in the literature) are indications that a specific group of citizens is hugely underrepresented. This appears to be a useful criterion that researchers can use to evaluate the representativeness of a survey on certain important characteristics.

Finally, we are aware of emerging concerns about respondent consent. We believe that it would be worthwhile for the consent procedures to be further documented for each study that is deposited with the CSES. Each country has relevant consent guidelines as well as expectations about how that consent translates into the future usability of surveys. At the very least, such information must be provided to researchers as information in available survey documentation. To that end, we have incorporated a new section into the revised CSES questionnaire.

Conclusion and proposals

The survey industry has been rapidly changing in the last decade, affecting the way in which the goal of high-quality, high-coverage data is met in practice. Internet studies, especially those drawn from established panels, are becoming increasingly popular. These types of surveys are usually not based on probability samples; respondents can usually self-select into a large pool of potential respondents. As these ‘opt-in’ panels are not randomly drawn, there is no theory for drawing inferences about the population based on sample statistics.

CSES has always held high data standards, and these standards are an asset that should be preserved. But given changes in the industry, the way of meeting these standards should be revisited. We believe that we need to recognize the current state of surveying and be able to

adjust as technological advancements (especially in online surveying) develop. Moreover, we should realize that decreasing response rates may become a serious threat to representativeness of probability samples as societies continue to evolve, so we should not naively believe that a traditional probability survey is always superior to what can be achieved with a nonprobability sample. Eventually, we think that the CSES should allow for the inclusion of high-quality surveys based on non-probability samples, but only when such surveys can be effectively evaluated on the basis of sufficient evidence of data quality, including representativeness.

Therefore, until clear standards have been established, we propose that the CSES take the following steps:

1. The CSES dataset should include election studies conducted among samples of eligible voters that can reasonably be considered to be representative of the population. No modes of conducting surveys should be excluded. All studies conducted in the spirit of CSES guidelines that are based on a probability sample of persons or households should be included. Such studies should include documentation on the sampling frame and its coverage of the population, as well as response rates, as is currently gathered by the CSES Secretariat. We also think the CSES should consider encouraging PIs to provide multiple weights, including those that include political variables.
2. The CSES website should develop a separate repository of 'supplementary files' which archives election studies that have fielded (substantial parts of) the CSES modules, but that are not gathered through probability sampling methods. This includes opt-in internet panels as well as surveys with a very limited geographical coverage which excludes more than half of the eligible population from the sampling frame. It might also contain probability-based studies that have not been included in the main CSES dataset because it only included parts of the modules. On the basis of the documentation that must be provided by the PIs of these studies, scholars may decide whether to include some of these datasets in their research. Yet, for Round 6 of the CSES, we recommend that these nonprobability studies are not included in the official CSES dataset.
3. Since the CSES is expanding beyond established democracies, the contexts in which electoral researchers collect their data will be increasingly diverse. While the goal should always be to include studies based on representative samples of the target population, local conditions will dictate what options exist to obtain such a sample. It is important that the PIs of each election study provides the appropriate documentation, so that researchers can decide whether they want to include the data in their study or not. We have provided a draft revision of the existing CSES questionnaire to addressed issues included in Baker et al. (2016) for evaluating the quality of non-probability surveys.

References

Ansolabehere, S., & Schaffner, B. F. (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis*, 285-303.

Baker, R., Blumberg, S., Brick, J.M., Couper, M.P., Courtright, M., Dennis, M., Dillman, D. Frankel, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lee, S., Lavrakas, P.J., Link, M., Pierkarski, L., Rao, K., Rivers D., Thomas, R.K., & Zahs, D. (2010), 'Report on Online Panels,' *AAPOR Report.*

Baker, R. Brick, J. M., Bates, N.A. Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. & Tourangeau, R. (2013), 'Report of the AAPOR Task Force on Non-Probability Sampling', *AAPOR Report*.

Baker, R., Brick, M, Keeter, S., Biemer, P., Kennedy, C., Kreuter, F., Mercer, A. & Terhanian, G. (2016) "Evaluating Survey Quality in Today's Complex Environment." *AAPOR Report*.

Breton, C., Cutler, F., Lachance, S. and Mierke-Zatwarnicki, A. (2017). Telephone versus Online Survey Modes for Election STudies: Comparing Public Opinion and Vote Choice in the 2015 Federal Election. *Canadian Journal of Political Science* 50(4): 1005-1036.

Brick, J.M. (2011). The Future of Survey Sampling. *Public Opinion Quarterly*. 75(5): 872-888.

Bytzek, E. & Bieber, I.E. (2016). "Does survey mode matter for studying electoral behaviour? Evidence from the 2009 German Longitudinal Election Study." *Electoral Studies* 43: 41–51.

Dassonneville, R., Blais, A., Hooghe, M. et al. (2020). The effects of survey mode and sampling in Belgian election studies: a comparison of a national probability face-to-face survey and a nonprobability Internet survey. *Acta Politica* 55, 175–198.

De Leeuw, E., Hox, J., & Luiten, A. (2018). International nonresponse trends across countries and years: An analysis of 36 years of labour force survey data. *Survey Methods: Insights from the Field (SMIF).*

Heffetz, O. & Reeves, D.B. "Difficulty of reaching respondents and nonresponse Bias: Evidence from large government surveys." *Review of Economics and Statistics* 101.1 (2019): 176-191.

Heffetz, Ori, and Matthew Rabin. "Conclusions regarding cross-group differences in happiness depend on difficulty of reaching respondents." *American Economic Review* 103.7 (2013): 3001-21.

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2020). Social media, web, and panel surveys: using non-probability samples in social and policy research. *Policy & Internet*.

Pasek, J. (2016). "When will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28: 269–91.

Rekker, R., van der Meer, T., & van der Brug, W. (2020). *Dutch Parliamentary Election Study 2017: A comparison of three different survey modes*. Universiteit van Amsterdam. https://doi.org/10.13140/RG.2.2.17927.42400

Sala, E. & Lillini, R. (2015). "Undercoverage Bias in Telephone Surveys in Europe: The Italian Case." *International Journal of Public Opinion Research* 29 (1): 133–56.

Sanders, D., Clarke, H.D., Stewart, M.C. & Whiteley, P. (2007). "Does mode matter for modeling political choice? Evidence from the 2005 British Election Study." *Political Analysis* 15: 257–85.

Stephenson, L.B. & Crête, J. (2011). "Studying Political Behavior: A comparison of internet and telephone survey." *International Journal of Public Opinion Research* 23 (1): 24–55.

Terhanian, G. & Bremer, J. (2000). Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research. *Research paper: Harris Interactive*.

Valentino, N.A., Zhirkov, K., Hillygus, D.S., & Guay, B. (2020). The Consequences of Personality Biases in Online Panels for Measuring Public Opinion, *Public Opinion Quarterly*.

Appendix

Comparative Study of Electoral Systems (CSES)
Module 5: Design Report (Sample Design and Data Collection Report)

Note:
Questions that seem especially relevant for NP studies are highlighted in yellow. New questions are marked with 'new'. In particular, those taken from AAPOR Baker et al. 2016 report are highlighted in blue.

Tale of Contents

Note: Because representativeness if becoming more of a challenge across all mode, this question will ensure that all teams consider this issue and present information for users.

Note: Given the changes to survey modes, we propose deleting Q32, which should repeat information provided in Q31 appropriate for mode.

Note: Given increased attention to ensuring that respondents have properly consented to their participation, we suggest that the Design Report includes a statement to this effect. Therefore, we suggest that the design report includes new Q41.

Note: Given the inclusion of mixed modes of surveying in the CSES, we propose that the Design Report includes some questions that provide users with additional information about the survey design and implementation. Some of these can be incorporated throughout by asking for answers for each mode, and others are included in this section.

Note: If the survey was conducted with mixed modes, please answer each question with information for each mode separately when appropriate.

Study Design

1. Timing of the study that the CSES Module was included in:

[ ] Post-Election Study (with interviewing starting within 6 months after the election)
[ ] Post-Election Study (with interviewing starting more than 6 months after the election)
[ ] Pre-Election/Post-Election Panel Study
[ ] Between Rounds

2a. Date Post-Election Interviewing Began:

2b. Date Post-Election Interviewing Ended:

3a. Mode of interviewing for the post-election survey in which the CSES Module appeared: (If multiple modes were used, please mark all that apply.)

[ ] In person, face-to-face - using a questionnaire on paper
[ ] In person, face-to-face - using an electronic/computerized questionnaire
[ ] Telephone
[ ] Mail or self-completion supplement
[ ] Internet

3b. Was there a mode change within interviews (e.g., selected self-completion elements within the questionnaire)?

[ ] No
[ ] Yes; please provide details:

4a. Was the survey part of a panel study?

[ ] Yes
[ ] No

4b. If the survey was part of a panel study, please describe the design of the panel study, including the date at which interviewing for each prior wave began and ended:

4c. If the survey was entirely or partly conducted via the Internet, please indicate whether it was based on an access panel (i.e. respondents were selected from a group of pre-screened panelists):

[ ] Yes, from a panel
[ ] No

4d. If the survey was based on an Internet access panel, please describe the access panel (company, population [does it include persons without initial access to the Internet and how are they interviewed], method of recruiting members, total size of access panel, method of selecting survey respondents from the panel):

Translation

Please provide copies of questionnaires in all languages used as part of the election study deposit. For questionnaires in a language other than English, please also provide a version of each translated back into English. Note: Questions are based on those developed for the ISSP.

5. Was the questionnaire translated?

[ ] Yes, translated by member(s) of research team
[ ] Yes, by translation bureau
[ ] Yes, by specially trained translator(s)
[ ] No, not translated

6. Please list all languages used for the fielded module:

7a. If the questionnaire was translated, was the translated questionnaire assessed/checked or evaluated?

[ ] Yes, by group discussion
[ ] Yes, an expert checked it
[ ] Yes, by back translation
[ ] Other; please specify: _____
[ ] No
[ ] Not applicable

7b. If the questionnaire was translated, was the questionnaire pre-tested?

[ ] Yes
[ ] No
[ ] Not applicable

7c. If the questionnaire was translated, were there any questions which caused problems when translating?

[ ] Yes
[ ] No
[ ] Not applicable

7d. If the questionnaire was translated, please provide a list of all questions which caused problems when translating. For each question listed, describe what problems were encountered and how they were solved:

Sample Design and Sampling Procedures

8a. Please describe the population that your sample is meant to be representative of:

8bnew. What steps were taken as part of the sampling and/or data collection process to ensure that the sample is representative of the target population?
-How can their success be determined?
-What is the sampling error of an estimate from this data? If a non-probability study, what is the credibility interval?

Eligibility Requirements

9a. Must a person be a certain age to be interviewed?
[ ] Yes
[ ] No
If yes, what ages could be interviewed?

9b. Must a person be a citizen to be interviewed?
[ ] Yes
[ ] No

9c. Must a person be registered to vote to be interviewed?
[ ] Yes
[ ] No

9d. Please list any other interviewing requirements or filters used:

Sample Frame

10a. Were any regions of the country excluded from the sample frame?
[ ] Yes
[ ] No
If yes, what percent of the total eligible population did this exclude from the sample frame? _____ %
If yes, please explain:

10b. Were institutionalized persons excluded from the sample?

[ ] Yes

[ ] No

If yes, what percent of the total eligible population did this exclude from the sample frame? _____ %

If yes, please explain:

10c. Were military personnel excluded from the sample?

[ ] Yes

[ ] No

If yes, what percent of the total eligible population did this exclude from the sample frame? _____ %

If yes, please explain:

10d. If interviews were conducted by telephone, what is the estimated percentage of households without a phone? _____ %

Please explain:

10e. If interviews were conducted by telephone, were unlisted telephone numbers included in the population sampled?

[ ] Yes

[ ] No

If no, what percent of the total eligible population did this exclude from the sample frame? _____ %

10f. If interviews were conducted via the Internet, what is the estimated percentage of households without access to the Internet? _____ %

10g. If interviews were conducted via the Internet, were provisions taken to include members of the population without access to the Internet? And if so, which?

[ ] Yes

[ ] No

If "Yes", please explain:

If "No", what percent of the total eligible population did this exclude from the sample frame? _____ %

10h. Were other persons excluded from the sample frame?

[ ] Yes

[ ] No

If yes, what percent of the total eligible population did this exclude from the sample frame? _____ %
If yes, please explain:

10i. Please estimate the total percentage of the eligible population excluded from the sample frame: _____ %

Sample Selection Procedures

11. Please describe, in your own words, how the sample for the study was selected. If the survey is part of a panel study and/or based on an Internet access panel, please also describe the original sample, from the beginning of the study.

12a. What were the primary sampling units?

12b. How were the primary sampling units selected?

12c. Were the primary sampling units randomly selected?
[ ] Yes
[ ] No
Please explain how the units were randomly selected. If the units were not randomly selected, please provide a justification for why the units were not randomly selected.

13. Were there further stages of selection?
[ ] Yes
[ ] No

13a. If there were further stages of selection, what were the sampling units at each of the additional stages?

13b. If there were further stages of selection, how were the sampling units selected at each of the additional stages?

13c. If there were further stages of selection, were units at each of these stages randomly selected?
[ ] Yes
[ ] No
Please explain how the units were randomly selected. If the units were not randomly selected, please provide a justification for why the units were not randomly selected.

14a. How were individual respondents identified and selected in the final stage?

14b. Could more than one respondent be interviewed from a single household?

[ ] Yes

[ ] No

If yes, please explain:

15. Did the sample design include clustering at any stage?

[ ] Yes

[ ] No

If yes, please describe:

16. Did the sample design include stratification?

Definition: Stratification involves the division of the population of interest according to certain characteristics (for instance: geographic, political, or demographic). Random selection then occurs within each of the groups that result.

[ ] Yes

[ ] No

If yes, please describe (please include the list of characteristics used for stratification, and in the case of multi-stage selection processes the stage[s] at which stratification occurred):

17. Was quota sampling used at any stage of selection?

[ ] Yes

[ ] No

If yes, please describe:

18. Was substitution of individuals permitted at any stage of the selection process or during fieldwork?

[ ] Yes

[ ] No

If yes, please describe:

19. Under what circumstances was a household designated non-sample? Please check all that apply:

[ ] Non-residential sample point

[ ] All members of household are ineligible

[ ] Housing unit is vacant

[ ] No answer at housing unit after _____ callbacks

[ ] Other (Please explain):

20. Were non-sample replacement methods used?

[ ] Yes

[ ] No
Please describe:

21a. For surveys conducted by telephone, was the sample a random digit dial (RDD) sample?
[ ] Yes
[ ] No

21b. For surveys conducted by telephone, was the sample a listed sample?
[ ] Yes
[ ] No

21c. For surveys conducted by telephone, was the sample a dual frame sample?
[ ] Yes
[ ] No
If yes, what % list frame_____ and what % RDD_____

22. For surveys conducted by mail, was the sample a listed sample?
[ ] Yes
[ ] No
Please describe:

23. For surveys conducted on the Internet, did respondents self-select into the survey, at any stage?
[ ] Yes
[ ] No
Please explain:

Incentives

24a. Prior to the study, was a letter sent to the respondent?
[ ] Yes
[ ] No
(If yes, please provide a copy of the letter.)

24b. Prior to the study, was a payment sent to the respondent?
[ ] Yes
[ ] No
If yes, please describe (including amount of payment):

24c. Prior to the study, was a token gift sent to the respondent?

[ ] Yes
[ ] No
If yes, please describe:

24d. Did respondent receive an additional payment after their participation? (Do not include any payment made prior to the study.)
[ ] Yes
[ ] No
If yes, please describe (including amount of payment):

24e. Were any other incentives used?
[ ] Yes
[ ] No
If yes, please describe:

Interviewers

25. Please describe the interviewers (e.g., age, level of education, years of experience):

26. Please provide a description of interviewer training. If possible please differentiate between general interviewer training and study-specific components:

26a. Please provide a description of the content, structure and time used for general training of interviewers:

26b. Please provided a description of the content, structure and time used for training interviewers in the specifics of the study within which CSES was run:

Contacts

27a. What was the average number of contact attempts made per household, for the entire sample?

27b. For households where contact was made, what was the average number of contact attempts prior to first contact?

27c. During the field period, how many contacts were made with the household before declaring it a **non-sample?**

27d. During the field period, how many contacts were made with the household before declaring it a **non-interview?**

27e. During the field period, what were the maximum number of days over which a household was contacted?

27f. During the field period, did interviewers vary the time of day at which they re-contacted the household?

[ ] Yes
[ ] No
If yes, please describe:

Refusal Conversion

28a. Were efforts made to persuade respondents who were reluctant to be interviewed?

[ ] Yes
[ ] No
Please describe:

28b. Were respondents who were reluctant to be interviewed sent a letter persuading them to take part?

[ ] Yes
[ ] No
(If yes, please provide a copy of the letter or letters.)
If yes, please describe:

28c. Was payment offered to respondents who were reluctant to take part?

[ ] Yes
[ ] No
If yes, how much?

28d. Were respondents who were reluctant to take part turned over to a more experienced interviewer?

[ ] Yes
[ ] No

28e. What was the maximum number of re-contacts used to persuade respondents to be interviewed?

28f. Were any other methods used to persuade respondents reluctant to be interviewed to take part?

[ ] Yes
[ ] No
If yes, please describe:

Interview/Survey Verification

Definition: Interview/survey verification is the process of verifying that an interview was conducted and that the survey was administered to the correct respondent, for quality control purposes.

29. Was interview/survey verification used?

[ ] Yes
[ ] No
If yes, please describe the method(s) used:
If yes, please indicate the percent of completed surveys that were verified: _____ %

30new. What steps, if any, were taken to ensure that respondents were providing truthful answers to the questions? Were any respondents removed from the final dataset (e.g., identifying speeders, satisficers, multiple completions)? If so, please provide details.

Response or Participation Rate

31a. If a probability survey, what was the response rate of the survey that the CSES Module appeared in? If a non-probability survey, what was the participation rate? Please show your calculations. (If the CSES Module appeared in a panel study, please report the response rate of the first wave of the study, even if the CSES Module did not appear in that wave.)

Definition: "the number of respondents who have provided a usable response divided by the total number of initial personal invitations requesting participation" (ISO 2008)

31bnew. Were steps taken to mitigate the impact of non-response in the dataset? If yes, how do the adjustments affect the survey results?

32. If the CSES Module appeared in a panel study, how many waves were conducted prior to the wave that included the CSES Module?

33. If the CSES Module appeared in a panel study, what was the total panel attrition between the first wave of the study and the wave that included the CSES Module? Please show your calculations.

34. If the CSES Module appeared in a panel study, please provide the number of completed interviews for the wave that included the CSES Module:

Post-Survey Adjustment Weights

35. Are weights included in the data file?
[ ] Yes
[ ] No

36. If weights are included in the data file, are these designed to match known characteristics of the population?
[ ] Yes
[ ] No
If yes, please describe which ones:

37. Please indicate the sources of the population estimates in the prior question. English language sources are especially helpful. Include website links or contact information if applicable.

Consent

38. Please describe the consent procedures that were followed in gathering the data. Please provide any Letters of Information that were shown to respondents.

39new. Did respondents give their consent that their responses be shared as part of the CSES data, according to local human rights regulations and laws?

Mixed Mode

Note: If the survey was conducted using mixed modes, please fill out the items about sampling frame (Q8new, Q10 above), sample selection (Q11-Q23), incentives (Q24), contacts/attempts to reach made (Q27-Q28), refusal conversion (Q29), survey verification (Q30, Q30new), response rate (Q31), post survey adjustment weights (Q35-Q37) *about each mode separately*.

Also: Please include a survey variable that includes the mode for each observation in the dataset.

40new. Was the mixed-modes design used to address a particular problem? Please elaborate.

41new. Were individuals assigned a mode after recruitment or was this part of the recruitment?

42new. If there are substantial differences in the distribution of key demographics along modes, please note them.

43new. Please list any additional advice you might have about how the survey should be used (e.g., important weights).