University of Gothenburg
Department of Political Science

Polling accuracy of vote intentions in Sweden using different weighting and sampling strategies

Master's thesis in political science
HT 2017
30 credits
Elias Markstedt
Supervisor: Johan Martinsson
Words: 19,999

Abstract

Over and over again, vote intention polls have been reported to fail their forecasts of votes such as the UK “Brexit” referendum and the US presidential election of 2016. The trustworthiness of opinion polls are called into question and this thesis aims to provide detailed knowledge of some of the circumstances that lead to inaccurate measurements of voting intention.

In past research, the various determinants of accuracy are usually only considered separately. Using an innovative method of creating all possible weight covariate combinations—which results in a dataset with over 98,000 bias adjustments of vote intention measurements in 21 probability and nonprobability samples collected in Sweden—a more holistic analytic approach is possible and the effects of accuracy determinants may be estimated simultaneously.

Adjusting for demographic variables such as gender, age and education are found to be relatively ineffective, while employing psychographic variables, such as vote recall and political interest is more fruitful. The choice of weighting technique, cell-weighting, raking or propensity score adjustment, matters little for the resulting accuracy. Probability samples produce more consistent and higher measurement accuracy, although making a distinction between different levels of quality in nonprobability samples reveal significant variation within the nonprobability category.

The application of weights should be done with care, since there is a risk that the weights introduce more bias than they remove. For survey research, these results suggest that there is a need to find unorthodox adjustment covariates similar to that of political interest to get more accurate measurements.

Table of contents

Table of tables

Table of figures

Appendix tables

1 Introduction

Overall survey measurement accuracy is a topic rarely present in the public debate, which is not surprising given the many technicalities involved. Surveys may lack in sampling method, suffer from high nonresponse and utilize poor measurement instruments, but unless there is a readily available benchmark measurement known to be accurate, failings will go by unnoticed and unopposed. Election forecasting failures are however obvious and there are high stakes involved, which is why survey accuracy debates tend to concern why election polls deviate from final election results. Vote intention polls are also important because they supply information about the confidence in incumbent governments in between elections, and they serve as indicators of what may be expected in the future.

An oft-cited example of polling gone awry is the 1936 US presidential election where the magazine Literary Digest contacted around 10 million Americans and asked whether they would vote for Franklin Roosevelt or Alf Landon. Despite the huge sample size, the poll failed to predict Roosevelt as the winner, underestimating his support by a staggering 18 percent. It was a significant dent to the magazine's pride as they had predicted the results in the four prior presidential elections with low levels of error (Literary Digest, 1936; Lusinchi, 2015).

A common explanation is that the haphazard sampling method (or lack of method) was the main culprit, resulting in a sample with too many high-income respondents. Today this method would be referred to as a *nonprobability* or *opt-in sample*. Squire (1988) has studied the Literary Digest case via a contemporary Gallup poll conducted right after the election of 1936. Squire finds the sampling argument to be true, but argues that the 24 percent response rate too was detrimental to the accuracy[1]. Gallup in turn made a famously poor projection of the 1948 presidential election (Mosteller, Hyman, McCarthy, Marks, & Truman, 1949), incorrectly predicting Thomas Dewey as the winner by overestimating his vote proportion by 5 percent.

Poor polling is however not a thing of the past. More recent examples include the 1992 UK election predictions, which misreported the Labour-Conservative difference by 8 percent (Jowell, Hedges, Lynn, Farrant, & Heath, 1993), and the 2002 French presidential election where the polls underestimated the support for Front National's candidate Jean-Marie Le Pen by 4 percent, who was unexpectedly voted through to the second round (Durand, Blais, & Larochelle, 2004). Both cases have been attributed to poor sampling practices.

2016 also saw controversies related to polling on the European Union membership referendum in the UK ("Brexit") and the presidential election in the US, both of which were typically poorly forecasted. Comprehensive post-mortems are however not yet available, so any definitive conclusions would still be premature.

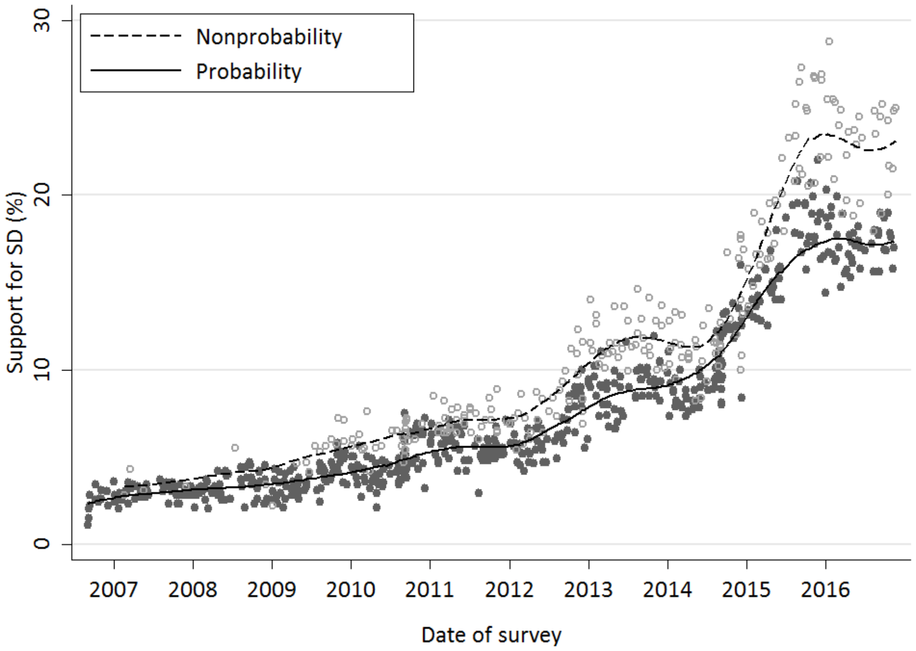
Conversely, polling in Sweden has historically been accurate, with only a few exceptions such as the 1968 election where the polling firm Sifo underestimated the Social Democrats by 5 percent (45 vs 50: Holmberg & Petersson, 1980, p. 22). Between 1944—when the first Swedish poll was collected—and 1994, there were not a single national election poll where the average

[1] Respondents in the Gallup poll in 1937 were asked whether they had received the Literary Digest poll and whether they had responded to it. The analysis concluded that had the entire sample returned their straw ballot, Roosevelt would have at least been correctly predicted as the winner, i.e. there was a majority of Roosevelt supporters among the nonrespondents. The numbers were: Roosevelt's 48 vs. Landon's 51 percent among respondents, 69 vs. 30 percent among nonrespondents.

error per party was over 2.5 percentage points (Petersson & Holmberg, 1998, pp. 137-142). The largest error since 1994 is only 1.8 percent[2].

Still, Swedish polling debates have erupted from time to time since the dawn of Swedish polling, such as the suitability of polling by the official statistics bureau in Sweden (Holmberg & Petersson, 1980, p. chapter 2) or the political bias of certain pollsters (Holmberg, 1986), but lately the debate has revolved around the issues of probability and nonprobability sampling (see e.g. Lönegård, 2016). An example of the impact of sampling is that during the past few years, nonprobability samples have fairly consistently produced greater support for the right-wing populist party Sweden Democrats (SD) than probability samples have (see Figure 1).

Figure 1. Sweden Democrat support by sample type



Comment: Based on a dataset gathered by Novus (2016) which includes most Swedish polls conducted between 2000 and early 2016 in Sweden, here showing polls between Jan 1, 2006 and Nov 26, 2016 (N=893). The lines indicate the moving average over time using Stata's -lowess- function (bandwidth: 0.2). The probability-based polls are Demoskop, Gallup, Ipsos, Ipsos/Synovate, Novus, Ruab, Statistics Sweden, SVT exit polls, Sifo, Skop, Synovate, Synovate/Temo and Temo. Nonprobability: Aftonbladet/Inizio, Sentio, United Minds, YouGov and Zapera.

In January 2016 for example, SD was reported to have 29 percent in a nonprobability-based poll by YouGov, while only 19 percent in a probability-based poll by Demoskop. It is uncertain which is closest to the true value, but both cannot be correct, assuming the there is indeed a *true* attitude and not simply something that is produced in the moment. It indicates that at least one of the two types of samples should be off the mark. In the 2014 national election, a nonprobability sample did produce the most accurate numbers for SD, while the reverse was true in 2010.

The examples illustrate that despite any improvements that have been made in the interim since the Literary Digest collapse of 1936, many of the problems persist and new challenges

[2] The number of party categories, including "other", has varied between 6 in 1944 and 9 in 2014. The largest average error was in 2002 (1.45, 7 polls) and the lowest in 1994 (0.84, 8 polls). The numbers are based on an unpublished summary by Sören Holmberg of 42 polls closest to each election.

have arisen. Another recent example is when the pollster company YouGov had Labour and the Conservatives tied at 34 percent in a pre-election poll running up to the 2015 UK election and the final result was 38 for the Conservatives and 31 for Labour. According to a post-mortem by Rivers and Wells (2015), the discrepancy was fundamentally a sampling issue with too few politically uninterested respondents in the final sample. It should be emphasized that this view ignores the fact that intentions and actions do not necessarily correlate perfectly; voters always do have an actual choice and are not necessarily predictable given available demographics or their predispositions. Nevertheless, they argue that the low accuracy could have been remedied by adjusting the levels of political interest to conform to reliable population measurements. Although this has the air of an afterthought, it is interesting nonetheless.

There are two competing views of nonprobability samples. They either just need the proper adjustments to consistently provide good measurements, as Rivers and Wells (2015) argue, or they are too unreliable to use since adjustments have no beneficial effects and only serve as superficial labels of data quality. Such a label is an even split male–female in the final sample for example. Critics such as Langer (2013) are suspicious of the methodology behind nonprobability samples: "[there is] no examination of how these estimates in fact are produced—including, potentially, their being weighted subjectively or to probability-based estimates" (p. 134). His view illuminates the problems with the non-transparency and complexity of how nonprobability samples are treated, although his points sometimes also apply to probability samples too since the methodologies and results are not always fully explained for them either.

The methods are known as *post-survey adjustments* or simply *weighting* and is used by most pollsters. As the YouGov numbers suggest, the effects of weights can be quite substantial and how they are constructed matters for polling accuracy. Particulars of those adjustments are however often largely unknown, which makes the merits of specific weighting schemes unclear. Tourangeau, Conrad, and Couper (2013, p. 31) argue that there is a need to test how well available techniques work in practice, an echo of past imperatives on the subject by Stephan and McCarthy (1958, p. 123): "[w]ithout the results of a substantial amount of empirical study, the deductive approach is purely speculative. It does not even make good progress toward a respectable theory of sampling opinion."

Using the Swedish case as a backdrop, this thesis heeds their suggestion and analyzes the impact of the three important dependent variables, sketched briefly above, on the dependent variable in this thesis: vote intention accuracy. Hypotheses regarding the effects are set up and explored: 1) how do the choice of particular covariates as well as the number and coding of those covariates affect measurement accuracy and 2) how does the specific weighting technique affect the accuracy? Finally, it examines 3) the conditions for measurement accuracy set by the sampling method.

Focusing on vote intention is a choice that rests on two foundations: first it is a central measure in political science by forecasting elections, gauging inter-party power relations and measuring perceptions of incumbent performance: the elections. Second, it is a well-known measure among the public with a constant supply of benchmarks, both the elections themselves and the survey measurements from many different sources.

The theoretical contribution of this thesis is twofold. First, it formulates new hypotheses on covariate effects on vote intention accuracy. More specifically, it theorizes that the overall efficiency of vote recall as a weighting covariate is decreased by the time that has passed between the election and the time of the survey by introducing the errors associated with recalling past behavior. Second, it introduces a qualitative gradation of nonprobability samples in this analytic setting. The gradation is theorized to produce heterogeneous effects on the level of measurement accuracy. A lower and a higher level of quality, an often overlooked aspect of nonprobability sampling, is defined based on 1) what type of channels respondents are recruited (or self-recruited) through and 2) the rate of panel attrition.

The main empirical contribution of the thesis is also twofold. First, it introduces an innovative, although computationally heavy analytic method of calculating all possible weight combinations given a set of weighting covariates (such as a number of demographics), number of covariates in each weight and the number of categories in each covariate. The advantage of the method is the resulting dataset that features the universe of possible weighting outcomes. It allows for a comprehensive analysis of how sensitive point estimates really are to the specification of weighting strategies in different samples. By varying which and how many covariates that are used and how they are categorized the study illustrate the many pitfalls of survey bias adjustments. Furthermore, it is a method which is generalizable to other types of measurements.

Second, the thesis applies these methods to the Swedish case where a unique combination of datasets with 21 different polls—both from the private and the academic sphere, and collected using various survey modes by 4 different polling organizations—are weighted in a multitude of ways in order to examine the resulting accuracy. In order to maximize the number of polls, the accuracy is determined by comparing the numbers to a high quality government-run survey (PSU), a method which is in line with comparable studies.

Results from the analyses show that demographics are fairly ineffective, while the use of psychographics, including vote recall and political interest, reduce more biases. Weighting techniques differs little in terms of resulting accuracy. Probability samples produce more consistent and higher accuracy measurements.

The application of weights should be done with care, since there is a risk that the weights introduce more bias than they remove. For survey research, these results suggest that there is a need to find unorthodox adjustment covariates similar to political interest to produce more accurate measurements.

The remainder of the thesis has the following structure: this chapter ends with a description of the general setting for surveys during the past few decades which explains the wide-spread use of nonprobability samples and why weighting is needed. Chapter 2 gives a theoretical overview of the choice of covariates, weighting techniques and survey samples. Chapter 3 describes the datasets as well as outlining the general design of the analysis presented in chapter 4. Chapter 5 contains a discussion of the implications and sums up what the conclusions are.

Survey setting background

Survey data collection have changed in many ways since the advent of polling in the first decades of the 20th century (Groves, 2011). Internet turned ubiquitous and brought the web survey, a new survey mode. Respondents turned more difficult to contact and, when found, less willing to participate (Curtin, Presser, & Singer, 2005; Kohut, Keeter, Doherty, Dimock, &

Christian, 2012), although there is considerable variation between countries, data collection modes and poll content (Baruch & Holtom, 2008; T. W. Smith, 1995).

Survey nonresponse is often higher among difficult-to-reach subgroups, such as respondents with low socio-economic status or of younger age (see especially de Leeuw & de Heer, 2002; see also Groves, Cialdini, & Couper, 1992). Survey fatigue resulting from increased polling may be one of many possible explanations (Holbrook, Krosnick, & Pfent, 2008; Porter, Whitcomb, & Weitzer, 2004), exemplified by the 900 percent increase in trial heat polls in the US between 1984 and 2000 (Traugott, 2005) and the number of election polls in the UK, which between 1945 and 2010 was 3,500 (an average of 54 a year), and 1,942 between 2010 and 2015 (323 per year: Sturgis, 2016).

Lowered costs are also likely an important factor in the increase in polling. ESOMAR (2014), a market research organization, reports that a third of all quantitative market research is conducted online, much higher than other survey modes[3]. Using web surveys is simply much cheaper than other modes (Evans & Mathur, 2005). Hill, Lo, Vavreck, and Zaller (2007) report that face-to-face interviewing may cost up to $1000 per interview, while a telephone interview may cost $200 and only $15 on the web. A discrepancy which is likely even larger today.

Survey recruitment through wide-reaching web sites has facilitated setting up large online panels of respondents willing to participate in recurring surveys. It explains why web survey-use often coincides with different types of nonprobability samples. Survey mode is a potentially confounding factor (Yeager et al., 2011, pp. 710-711): essentially all nonprobability samples make use of web surveys, while (publicly reported) probability samples are mostly collected via telephone, at least when it comes to Swedish polls.

Low costs, potentially huge samples sizes and technological flexibility—such as showing images, video and question filtering possibilities—all may explain why web surveys and nonprobability samples are increasingly popular in research, even though there is an increasing number of probability-based web panels as well (Bosnjak, Das, & Lynn, 2016). Web panels in general are often plagued by systematic error and particularly nonprobability-based ones. The validity of other modes and probability samples are however also increasingly threatened by recent developments, which underscores the need for bias adjustments in order to get accurate measurements.

2 Theory

In this chapter, hypotheses related to three important determinants of survey measure accuracy are developed and these are later tested in Chapter 4 on voting intention measurements. The three aspects will be described in following order: the choice and specification of weighting covariates, the choice of weighting technique and the conditions set by sample type. But first, what constitutes a good weight and the concept of vote intention both needs a short overview.

To theorize about the adjustment procedures, knowledge about its prerequisites is needed. Total survey error is a useful theoretical framework, the current paradigm in survey research. It brings together all the ways a “true value” in a population may be biased when brought all the way through to a final measurement in a responding sample. In an influential

[3] The percentages are recalculated to represent proportions of the *quantitative* market research only, which is 74 percent of the total, rather than all market research.

conceptualization of the framework, Groves et al. (2011) describes two main strands of inferences: measurement and representation, each with their own sources of error. Measurement refers to the inference from theoretical concept to final measurement via operationalizations such as question wording and response alternatives. On the representation side, the target population inference is made from a set of respondents. Here, the errors are those of coverage error, for example when the sampling frame does not include the whole population, sampling and nonresponse error. All the above aspects are relevant for the accuracy of vote intention, but the focus in this thesis will lie on the nonresponse error and to some extent coverage error because these are the errors that may be adjusted by weighting.

Errors are unproblematic if they are random, but inference is threatened when factors determining willingness to participate in surveys (*P*, nonresponse), or the likelihood to be included in a sampling frame and the variable of interest (*Y*) have common determinants (Z: Groves, 2006). Consider an example with nonresponse and newspaper readership (Peiser, 2000): as both are correlated with age (*Z*), then higher nonresponse will lead to lower accuracy of the newspaper readership measurements. Meta-analyses have also shown that nonresponse error seldom is predictive of nonresponse bias, but varies substantially variable to variable (Groves & Peytcheva, 2008; Sturgis, Williams, Brunton-Smith, & Moore, 2016).

Leverage-saliency theory is one of the few theories that suggests a mechanism behind the reason to participate in a survey (Groves, Singer, & Corning, 2000). Leverages are things such as cash incentives, while saliency captures the various aspects of a survey that might cause a respondent to answer or not, such as the topic of a survey. The theory is therefore useful since it gives a way of identifying covariates that predicts *P*, even though the saliency may even vary from survey to survey for a single individual.

In essence, the purpose of weighting is then to remove any biases that were the result of sampling and data collection efforts. Strata or cells are created that "are homogenous with respect to the target variable" (Bethlehem, 1988, p. 259) using a vector of adjustment variables, also known as covariates (Z; using the notation from: Groves & Peytcheva, 2008)):

$$Cov(P, Y \mid Z) = 0.$$

Simply put, the goal is to find the variables that efficiently remove the relationship between vote intention and survey participation. The usual process of creating weights has three stages (Kalton & Flores-Cervantes, 2003), countering unequal selection probabilities, nonresponse bias and making sure key variable distributions in the finalized sample looks like the population equivalents. In this thesis however, the stages are conflated into one single adjustment of coverage, sampling and nonresponse bias. The next step is to examine the *Y* variable and its determinants.

Vote intention and weighting covariates

The dependent variable, *Y*, in this study is vote intention, often measured in Sweden with the question (or similar to): *If there was an election today, which party would you vote for?* It is a measure that is generally meant to measure the future behavior in elections. Reported voting behavior and intention do correlate strongly in countries such as Sweden and the US (Granberg & Holmberg, 1990), but the US numbers correlates somewhat less strongly when using validated vote information (Achen & Blais, 2010). The relation was theorized by Ajzen (1985) in the theory of planned behavior, where there is a direct correlation between intention and behavior. Past behavior and self-identity (J. R. Smith et al., 2007)—which in this case would be electoral participation and political engagement/identification—are also found to have a direct effect on voting, independent of intention (Granberg & Holmberg, 1990). The benchmark used here, however, is a high-quality survey benchmark for all cases but one (see Chapter 3), so there is no theoretical reason attribute any effects to the attitude–behavior discrepancy, but the theory might still be informative in finding covariates that predict vote intention.

The search for a set of covariates that will be a general panacea for all biases for all *Y* variables is unfortunately futile since the determinants of different *Y* variables naturally differ. Although for specific outcomes such as vote intention, the search could turn out to be more fruitful.

Thomassen (2005, p. 6) says that "[s]tability and change in the mutual strength of political parties depend on two consecutive decisions individual citizens make. First, the decision whether to vote, and second, the choice of a particular party. " Between the 1920s and 1960s, the overall electoral volatility in most of the West European and North American democracies was low (Lipset & Rokkan, 1967): in terms of election turnout and with regard to the power structures between parties. Since then, turnout has decreased in many countries (Dalton, 2008, p. 37; though not in Sweden) and there is lower correlation between factors such as socioeconomic class on one hand and turnout and party choice on the other (Dalton, 2008, chapter 8). It suggests that elections are increasingly subjected to the saliency of political issues rather than demographics and pure sociological theories of voting and political participation have a worse fit on the data. Such an issue is refugee and immigration policy, an important issue for Sweden Democrats, a group whose support also tends to be underestimated in some polls and overestimated in others. As such, it is potentially useful as a weighting covariate.

While turnout in Sweden has no real trend at all, the predictive power of demographics on party choice in Sweden is slowly waning (Oscarsson & Holmberg, 2013, p. 77). It does not necessarily mean that demographic factors are unimportant, but rather that they are increasingly mediated via other facets of politics. By extension, it also complicates the conditions for weighting efforts. Voters with specific socio-demographics might in some political contexts coalesce around one particular party in one election, and in another it might not. The development has been attributed to many of the same things as the decline in survey participation: the individualization and "modernization" of society (Thomassen, 2005), so it might be possible to find common denominators. Education is however still fairly predictive of party choice on a bivariate level (Oscarsson & Holmberg, 2013, p. 77)

Vote recall is a particularly interesting covariate since it correlates very strongly with vote intention and as a result is also a commonly used covariate in polls. On the downside, there is ample evidence that time deteriorates memory and may, in the case of vote recall, slowly be colored by the present party preference, i.e. a sort of bandwagon effect (see e.g. Durand, Deslauriers, & Valois, 2015; van Elsas, Lubbe, van der Meer, & van der Brug, 2013). The more time that has passed since the election, the less effective it should be as a bias reduction tool.

In practice, the most commonly used weighting covariates are however often also the "lowest-hanging fruit": the demographic variables that are often available in sample frames. Gender, age, education, geographical location, employment, income and marital status are commonly found, as well as "race"/ethnicity in the US. Loosveldt and Sonck (2008) and

Shadish, Clark, and Steiner (2008, pp. 1340-1341) also find that “predictors of convenience”—i.e. the demographic variables readily available in both the frame and in the sample—are poor instruments to reduce bias. Age and education is often found to be related to response propensities, but it only accounts for the *P* side.

Other covariates, such as topic interest and similar psychographic measures are less studied, even though there are many indications that for example interest might be highly predictive of self-selection into surveys and panels that have a focus on issues the respondent is interested in (Groves et al., 2006) as the leverage-saliency theory described above predicts. In this case it would be measures such as political interest, though the lack of non-survey benchmarks is a problem for psychographic measures.

Now moving on to two other aspects of weights, namely the number and coding of covariates. They are usually not included in similar studies, perhaps since it is thought to be of little importance. Although there is little to go on in terms of previous results, but this study has a design that allow for an easy examination of the potential effects. For example, having too few categories for an age group variable, say young and old, might cluster exceedingly heterogeneous subgroups together in terms of *P* and *Y*, thus limiting the bias adjusting properties. Conversely, having too many categories might result in cross-tabulated categories with zero or few respondents, which in turn might lead to weights with high variability, but average point estimates should not be affected much. It is hypothesized that, *ceteris paribus*, more covariates and more covariate categories will improve polling accuracy.

The chapter may then be summarized in the first four hypotheses:

H1a *Vote intention accuracy is improved with each added covariate.*

H1b *The more categories a covariate variable is coded into, the more bias reduction.*

H1c *Covariates that are correlated with both vote intention and response propensities will be the most efficient in reducing bias: vote recall, political interest and education.*

H1d *Vote recall moderated by the time that has passed since the election: the longer the less effective will it be.*

Weighting techniques

The second area of accuracy determinants examined here are the specific adjustment methods, an area which has not been discussed in public as much as the sampling controversy. Data management practicalities might be viewed as a more esoteric subject. The stakes for the involved parties, economic or otherwise, are also not as high, and the dividing lines in the literature are also not as clear since the techniques are not necessarily mutually exclusive. Regardless of the reason, different types of techniques are bound to be less well known.

There is a wide array of different weighting techniques available (see the review by Kalton & Flores-Cervantes, 2003), among which some of the more commonly applied are *cell-weighting* (Kalton, 1983), *raking* (Battaglia, Hoaglin, & Frankel, 2013; Deville & Särndal, 1992), *GREG weighting* (generalized regression estimation: Bethlehem & Keller, 1987) and lately also *PSA* (propensity score adjustment, originally described by Rosenbaum & Rubin, 1984; see also

Rubin & Thomas, 1996)[4]. GREG, since it is closely related to raking, will however not be included in the analyses.

Cell-weights and raked weights

A cell-weight consists of the ratios between the proportions of each of the cells of cross-tabulated covariates (*Z*) in the final dataset and in the population. For example, cross-classifying gender and age with two categories each would result in four proportion ratios. The basic assumption is that respondents and nonrespondents in each cell are similar, and as such, the respondents' answers correspond to the answers nonrespondents would have given (see Appendix B for examples). Put in the notation from earlier, *Y* is now assumed to be independent of *P*.

Raking is similar to cell-weighting, but requires only the *marginal* population totals, that is not the joint distribution. The weight is created by iteratively adjusting marginal totals (*Z*) until they are simultaneously the same as in the target population. Using the same example as above: gender is first adjusted to the population margin totals and then the same is performed for age. The second adjustment is likely to have skewed gender once again and is therefore adjusted a second time. This is repeated until (and if) all margins converge. Two advantages of raking there is less risk to add sampling variance to the data than cell-weights and it allows for the use of population data from different sources. On the downside, it assumes no interaction between the covariates, which might undershoot in terms of adjustment.

Propensity score weights

PSA is different from cell-weighting and raking in that it is based on an explicit model of survey participation. Propensity scores are usually estimated by fitting a logit model with survey participation as the dependent variable and a covariate set as independent variables in a sample consisting of both the dataset to-be-weighted and a reference survey (register data is even better). Weights are created by balancing differences between the propensity scores in the two samples in each quantile. Cochran (1968) argues that the optimal number of quantiles, or bins, is five (quintiles), although that has not (to the author's knowledge) been tested since.

The main advantage of PSA vis-à-vis cell-weighting and raking is that many more covariates may be added to the model, including more or less continuous variables such as age or number of contact attempts. Misspecification of the model also does not seem to bias the PSA weighting effort (Stuart, 2010, p. 5). A possible disadvantage of PSA is that by using a reference survey, the method might not adjust for noncoverage, which could be detrimental to data quality. It might also run into sample matching difficulties when using too few covariates in the matching procedure, which should reduce its efficiency.

Effects of weighting techniques

Tourangeau et al. (2013, pp. 31-32) summarizes studies that looks bias reduction properties of the techniques, see Table 1 below. Four of the studies have a design which is similar to this thesis: comparing web survey estimates with estimates from a benchmark study (Berrens, Bohara, Jenkins-Smith, Silva, & Weimer, 2003; Schonlau, van Soest, & Kapteyn, 2007; Schonlau

[4] Weighting techniques are associated with several terms: *cell-weights* are often referred to as post-stratification weights, balancing weights or base weights. The term *cell-weight* has the advantage that it describes the method in practice closer. *Raking* may also be called iterative proportional fitting or random iterative method (RIM).

et al., 2004; Yeager et al., 2011), while the rest compare a subset of a dataset (e.g. Internet users) with the whole analyzed dataset (Dever, Rafferty, & Valliant, 2008; Lee, 2006; Lee & Valliant, 2009; Schonlau, van Soest, Kapteyn, & Couper, 2009).

Tourangeau et al. (2013) conclude that bias is decreased by all methods of adjustment (in most cases), but with significant portions still remaining afterwards (see the column "Mean reduction in bias" in Table 1). There are also large differences depending on which covariates are chosen. A closer examination of the studies however reveals that there are some gaps in the knowledge produced by the studies.

First, the only study that examines cell-weights does not report enough information to show the bias reduction and none of them compare cell-weights and raked weights. Second, almost all of the studies use US data exclusively, which calls the generalizability into question. Third, only Yeager et al. (2011) makes a distinction between probability and nonprobability samples. Fourth and last, all of the studies use few covariate combinations, and none of them employ different categorization of the same covariates. All but Lee (2006) and Lee and Valliant (2009) fail to provide any systematic guidance on how to decide which covariates to use.

A study not included in the original list is a report by Steinmetz, Tijdens, and de Pedraza (2009) where Dutch and German data is assessed. Even though they conclude that differences between cell-weights and PSA weights are small when the most effective configurations are used, the average bias reduction when all combinations are taken into account tells another story: their cell-weights actually increase the bias in most cases.

Table 1. Bias reduction levels by weighting techniques

	Mean proportional bias reduction (%)[a]						
	Cell-weight	*Raking*	*PSA*	*GREG*	*Number of weights[b]*		*Country of origin*
Study					*#*	*page ref*	
Berrens et al. (2003)		−10.8	−31.8		1	(p. 9)	US
		+3.0					
Dever et al. (2008)				−23.9	3	(p. 59)	US
Lee (2006)				−31.0	9	(p. 340)	US
Lee and Valliant (2009)			−62.8	−73.3	5	(p. 335)	US
Schonlau et al. (2007)			−24.2		2	(p. 14)	US
			−62.7				
Schonlau et al. (2009)			−43.7		8		US
Schonlau et al. (2004)	NA		NA		1		US
[Steinmetz et al. (2009)[c]]	+36.6		−39.6		8	(pp. 28–29)	DE/NL
Yeager et al. (2011)		−30.6			1	(p. 717)	US
		−35.3					
		−37.4					
		−38.7					
		−42.0					
		−53.3					
		−57.0					
Min/	+36.6 /	+3.0 /	−24.2 /	−23.9 /			
Max	+36.6	−57.0	−62.8	−73.3			

Comment: Adapted from Tourangeau et al. (2013, pp. 31-32) with some additions. Note that the sign is inverted. a. Reduction in bias is calculated as the mean difference between the weighted and unweighted web survey estimate in relation to a benchmark. b. Weight setup here is defined as specific combinations of adjustment covariates, which the data is changed to conform to. c. The bias number is based on approximate wage numbers that are visually procured from the figure on p. 30 since no actual numbers are reported.

Although the data and methods might be too varying in the earlier studies to make anything but tentative conclusions, there are indications that PSA might perform more consistently better than cell-weighting and raking, which might be due to a more parsimonious and dynamic categorization of nonresponse propensities in a continuum rather than the fairly rigid method of nominal cross-classification and use of many more covariates. Theoretically, raked weights using the same set of covariates but stripped of the joint distribution should produce less effective weights than cell-weights. Also, the iterative procedure involved in raking could also in some cases lead to non-convergence.

There are other ways of further decreasing bias, for example using calibrated PSA weights, i.e. an additional layer of weights that adjust a sample to have matching totals with the population totals. Lee and Valliant (2009, pp. 336-337, 340) say that "[t]he calibration step is particularly important for surveys from which totals are to be estimated. If only means or proportions are needed, then the propensity adjustment alone may be sufficient" (p. 340).

To sum up, cell-weights and raked weights should be equal in terms of bias reduction as long as there are no interactions present in the vector of covariates. Since interactions are not uncommon, raked weights should reduce bias somewhat more. PSAs main strength vis-à-vis the other methods lies in the number of covariates that may be included, so when fewer covariates are used PSA should run into matching issues, decreasing the efficiency.

H2a *Cell-weights will reduce more of the vote intention measurement biases than raked weights since it retains the most accurate information.*

H2b *PSA will reduce vote intention bias less than the two other types of weights using the same set of covariates. With an increased number of covariates, it will surpass cell-weighting and raking.*

Sample type

While weighting technique is noncontroversial, sampling methodology is the opposite. The probability versus nonprobability divide can be traced back to the late 1800s where the first fundamental building blocks of sampling inference where laid down by Norwegian statistician Anders Kiær in 1896 (Kruskal & Mosteller, 1980) as an attempt to move away from full enumeration. It was the more elusive concept of *representativeness* that was first proposed, an early variant of the quota sample, which later was developed by others to require a randomization component from which non-zero selection probabilities may be derived, a requirement when generalizing the sample results to a population (Kish, 1965). A few decades into the debate, Hansen and Hauser (1945) argue that researchers need to design a sample so that:

> "…that each element of the population being sampled […] has a chance of being included in the sample and, moreover, that that chance or probability is known. The knowledge of the probability of inclusion of various elements of the population makes it possible to apply appropriate weights to the sample results so as to yield 'consistent' or 'unbiased' estimates" (pp. 184–185).

It is also argued that it is impossible to determine probabilities of inclusion and reliability measures such as confidence intervals in nonprobability sampling (referred to as quota sampling). Proponents of nonprobability sampling admit that "…no exact solution for the statistical reliability of quota polls has been achieved," but retort that relevant demographic categories are controlled for and thus decreasing the possibility for biases, while at the same time pointing to past successes in predicting election outcomes (Meier & Burke, 1947, p. 587).

Despite the many claimed innovations in the field seven decades later, the debate has changed surprisingly little. Nonprobability sampling is still criticized for lacking a theoretical framework (see e.g. Langer, 2013), a fact which is not (entirely) disputed by proponents (Baker et al., 2013). It is however maintained that self-selection, an innate aspect of nonprobability samples, differs little from the forces driving survey nonresponse patterns (Rivers, 2007, p. 8). More importantly, it is argued that while selection probabilities are unknown in nonprobability samples, they may still be estimated from case to case (Rivers, 2013). Advocates then refer to the track record of habile (mostly election) predictions that serve as evidence for suitable practices.

The publicly available empirical evidence does support the *possibility* to create consistently accurate and reliable vote intention measurements from nonprobability samples. Studies of vote intention measurements from the US elections of 2000, 2008 and 2010 (Ansolabehere & Rivers, 2013; Rivers & Bailey, 2009; Taylor, Bremer, Overmeyer, Siegel, & Terhanian, 2001; Vavreck & Rivers, 2008) and the UK election of 2005 (Twyman, 2008) show that nonprobability samples may be used successfully, but the accuracy is not always compared with probability-based samples. A carefully constructed nonprobability sample may indeed produce as or more accurate election predictions as probability samples, but it is also clear that adjustment procedures are often very complex with up to 7 distinct steps of pre- and post-stratification and other techniques (see in particular: Ansolabehere & Rivers, 2013). However, these results may be the result of publication bias and it says little to nothing about the general consistency between samples providers.

A study of the Swedish case (Sohlberg, Gilljam, & Martinsson, 2017) where 110 polls from the 2006, 2010 and 2014 Swedish national elections campaigns are analyzed, indicates for example that nonprobability samples have a slightly lower accuracy even when controlling for sample size and temporal distance from the election. They only use aggregate data however, and thus do not disentangle sampling from bias adjustment methodology.

Vote intention might also be an “easy” case for nonprobability samples since much is known about how to model voting. The accuracy of measuring other concepts using nonprobability samples outside the confines of electoral polling is mixed at best (Baker et al., 2013, p. 5). In one of the most often cited studies in the field, Yeager et al. (2011) find that the outcome variable is driving the results. Smoking frequency, subjective health quality and possession of a driver’s license are analyzed showing that the probability samples consistently show greater accuracy than the nonprobability sample, with and without weights. Pasek (2016) illustrates that tests of accuracy may be extended to include concurrent and predictive validity, where correlations are found to be similar, but probability samples are superior when it comes to point estimates and predictions, though only through using rudimentary demographic weights.

It should however be emphasized that the qualitative divide between probability and nonprobability samples is more blurred (Baker et al., 2013). High-quality nonprobability samples might outperform less well-adjusted probability samples such as in a recent large study of the accuracy of nonprobability samples by Pew Research Center (Kennedy et al., 2016). Across 20 measurements, the bias of nonprobability samples was very varied with Pew’s own probability based panel ending up in the middle.

On one hand, more measurements from probability samples are likely to suffer from nonresponse bias. On the other hand, some of the issues often attributed to nonprobability (web) samples, such as the low Internet penetration, are less of a problem today. For example, 99 percent of the households in Sweden have potential access to some kind of broadband[5] (European Commission, 2016). Many producers of nonprobability samples such as YouGov are also, through necessity, arguably more knowledgeable about available bias adjustment techniques. Samples are therefore more diverse than ever and should therefore be nuanced to a greater degree in analyses.

Although no actual sampling frame exists for a nonprobability sample, as Groves et al. (2011, p. 84) argue, the level of likeness to a true frame should be possible to approximate based on 1) which recruitment avenues are used and 2) the level of attrition of the panel to which the recruitments were made. Consider one of the featured datasets (*CPVAA*) that was exclusively recruited via large voting advice applications (VAAs: Rosema, Anderson, & Walgrave, 2014). The VAAs were featured online on one of the most popular websites in Sweden—the tabloid Aftonbladet's site aftonbladet.se[6]—during two election campaigns in 2014. 1 and 2.3 million completed tests respectively[7], compared to the 7.3 million who were eligible to vote. About 1 percent went on to join the Citizen Panel, an academic web panel, from which the samples used here were drawn. A second sample is a convenience sample (*CPCON*) composed of many different solicitation efforts using several different recruitment avenues, a majority coming from recruitments on local newspaper website during the 2006 and 2010 election campaigns in Sweden. The CPVAA was recruited not more than one year before its last use in this study, while the CPCON was generally recruited between four and eight years prior to sampling. The CPVAA sample could be defined as coming from *wide and recent recruitment*, while CPCON originates from a *narrow and old recruitment*.

An indication that the above approximation is reasonable is that the raw CPVAA data is closer to the Swedish population than CPCON in terms of demographics, political interest and party identification. By extension, they should also display different levels of vote intention accuracy.

To sum up, earlier studies have produced mixed results in terms of how the sample type affects the accuracy of vote intention measurements, with a slight upper hand for probability. The different levels of quality between nonprobability samples should amplify the difference. The level of improvement of vote intention accuracy, i.e. the bias reduction, is however reversed since there is more initial bias to remove.

H3a The post-adjustment accuracy of vote intention in Swedish samples is the highest in probability samples, second highest in the nonprobability sample with wide and recent recruitment and lowest in samples with narrow and older recruitment.

[5] The use of Internet in Sweden is ranked 2nd in the EU.

[6] Using an online panel, reach50.com (https://reach50.com/#reach/2014/37) finds that aftonbladet.se was visited by 43 percent (6th place of all sites) at least once during the week the general election was held. Similarly, the KIA index maintains that the website was number one in Sweden in a somewhat less comprehensive list during the same week (http://www.kiaindex.se/sok/?site_name=&category=&kyear=2014&kweek=37&section=&hide_networks=&filter=1), or about 5.5 million unique web browsers.

[7] About half were collected from unique IP addresses. Not all of those are duplicates though since it is common that many users have the same IP number.

H3b The level of improvement of vote intention accuracy due to a weight is the reverse of the above, with a greater improvement in nonprobability samples and lesser in probability samples.

3 Methodology

In this chapter, the first section describes the dependent variable, polling accuracy and its choice and use of benchmarks. Second, the choice of datasets is explained, including the reference survey, followed by a discussion and outline of the covariates and choice of study design.

Data

Three groups of datasets are used in this study: the first collection is from a large university-based online panel, the Citizen Panel, a panel that includes both a probability-based sample and two diverse opt-in samples. Here, there was greater control over the collection and design of Citizen Panel datasets than the other data sources.

A second group includes several surveys from three private survey companies in Sweden: one probability-based telephone survey: Demoskop, and two nonprobability samples: Inizio and United Minds. It is second-hand use of the data, which means there was no control over what and how the data was collected, but their inclusion in the study permits greater generalization of the results.

The third group consists both of the benchmark surveys that holds the “true” vote intention measurement—Statistics Sweden’s *PSU*, described more in-depth in the next section—and the reference surveys that is needed for the joint covariate distribution: the *SOM* surveys.

The Citizen Panel is administered by the Laboratory of Opinion Research (LORE) at the University of Gothenburg (Markstedt, 2016). Two probability-based samples (*CPPROB*, N_{w13}=3,177, N_{w15}=1,575) were collected in two waves in May and November 2015 (wave 13 and 15). They were originally recruited as two separate samples in 2012 and 2013 when postcards were sent to a list of addresses randomly drawn from a population register of Swedish residents. The cumulative response rate, i.e. the recruitment and survey response rate multiplied, was 5.4 percent and 5.9 percent respectively.

Two responding samples of two different nonprobability sample *variants* were collected parallel to CPPROB: the CPVAA (N_{w13}=3,533, N_{w15}=7,579) and *CPCON* (N_{w13}=1,836, N_{w15}=4,463), they were mainly recruited online on newspaper websites; see the introduction for a discussion on their recruitment. Participation rates varied between 52 and 70 percent.

Specifically for this thesis, a number of datasets were made available by polling companies that are doing polls in Sweden: Inizio that has a nonprobability web panel (IN, N_{Nov14}=1,609, N_{May15}=4,686, RR≈63 percent) and Demoskop conducts telephone interviews with cross-sectional probability-based samples (DS, N_{min}=1,267, N_{max}=1,285, RR=16 percent). The second nonprobability sample is the United Minds data, which was available online (UM, N_{min}=954, N_{max}=1,171)[8]. Since United Minds collected data continously, the data used here was matched to the collection periods of the PSU benchmark survey. Table 2 summarizes the surveys-to-be-

[8] http://unitedminds.se/open-opinion/ (2016). United Minds discontinued its party preference surveys in October 2014.

weighted as well as the benchmark data and Table 3 illustrates the approximate period for each dataset. See Appendix table 1 for more information on each study in this thesis.

Table 2. Data sources

	Data provider (panel)	Data collection period	Sample type	Survey mode
Surveys	LORE (Citizen Panel)	Nov 2014 – May 2015	Probability & nonprobability	Web
	Inizio (Sverige Tycker)	Nov 2014 – May 2015	Nonprobability	Web
	Demoskop	Nov 2013 – May 2015	Probability	Telephone
	United Minds (Väljarbarometern)	Nov 2010 – Nov 2014	Nonprobability	Web
Bench-mark	Statistics Sweden (PSU)	Nov 2010 – May 2015	Probability	Telephone & web

Table 3. Available data points

Date	Nov 2010	May 2011	Nov 2011	May 2012	Nov 2012	May 2013	Nov 2013	May 2014	Nov 2014	May 2015	Nov 2015
CPPROB										✔	✔
CPVAA										✔	✔
CPCON										✔	✔
Inizio (IN)							✔	✔	✔	✔	✔
Demoskop (DS)										✔	✔
United Minds (UM)	✔	✔	✔	✔	✔	✔	✔	✔			

Demoskop's sample is based on number lists with both landline and mobile phone. Inizio's sample is recruited largely the same way as the CPVAA, via pop-ups on the publishing house Schibstedt's websites (such as Aftonbladet, Svenska Dagbladet and others) which is then pre-stratified by gender, age and region. United Minds use the same variables when pre-stratifying samples via Cint, an online panel aggregator that in turn draws its samples from several different nonprobability panels. More detailed information on United Mind's sampling is however unavailable.

Accuracy measure – benchmark and calculation

A polling accuracy measure consists of two parts: the *benchmark* with which an estimate is compared with and its *calculation*. This section will begin by describing the considerations surrounding the choice of benchmark.

In many studies where vote intention accuracy is measured, the benchmark is simply a contemporary election (usually the election the poll is meant to forecast). While forecasting elections is a straightforward and common use of polls, it is not their sole purpose; gauging between-election support for the incumbent and opposition is also an important use. Furthermore, the greater the distance between a poll and an election, the less sound is it using the following election as the actual benchmark, since much may change during the last few months of an election campaign.

A high quality poll can be used as a substitute benchmark in order to measure the accuracy of polls far removed from an election campaign. This is only possible when using a poll that enjoy higher response rates than the probability samples. Since many of the datasets did not coincide with an election, the chosen benchmark is the Swedish Party Preference Survey (PSU). The PSU is a gold standard survey conducted by Statistics Sweden and has very large samples (N_{min}=4,757, N_{max}=6,192) and relatively high response rates: between 50-68 percent during 2010-2015 and is weighted for sex, age, region, education, foreign born and vote recall (see e.g. Statistics Sweden, 2014).

Joint distributions of covariates, a requirement for cell-weights and PSA, are however unavailable for the PSU data so the omnibus SOM surveys is used as a population substitute (N_{min}=5,007, N_{max}=6,876, RR≈50 percent). SOM cannot be used as an overall benchmark since it is not biannual like the PSU and would therefore have limited the datasets too much. There are indications that the standard party preference question in SOM ("*Which party do you like best today?*") also does produce somewhat different results from the vote intention question used in the other surveys in this study ("*If there was an election today, which party would you vote for?*", with some variation) (Statistics Sweden, 2014, p. 9), in part likely due to strategical considerations which are more prevalent in the latter measure.

Note that the SOM dataset is itself not weighted, but deviations from the Swedish population are generally low, except for age which is biased somewhat towards older respondents (Markstedt, 2014). This might have biased the results, but there were examinations carried out which compared a whole subset of adjustments based on a weighted and an unweighted SOM dataset; only very small differences were found. It is likely a reflection of a low correlation between age and vote intention during the studied period. To use this data, a number of assumptions needs to be made. The SOM national surveys only include vote recall measurements during election years, so the 2010 dataset will be used as the benchmark for all data collected during the 2010–2014 term. 2014 SOM data is used with data collected after the general election in 2014. It is assumed that the relationship between covariates does not change during the selected time period, a likely assumption given the stability of these measurements.

The type of benchmark makes the method generalizable to other types of opinions where no viable alternative to a survey exists, a design similar to that of Ansolabehere and Rivers (2013) and Yeager et al. (2011). Using another survey as a benchmark is not without its limitations though; a survey benchmark is subject to many of the same types of biases as the survey-to-be-weighted. The accuracy of the benchmark itself is also more difficult to assess since there is no benchmark for the benchmark. In order to make sure the choice of benchmark does not bias the overall results, one poll included in this study use the 2014 national election as a benchmark (Demoskop September 2014).

A poll-of-polls, an average of all known polls, could also serve as an alternative to a benchmark survey, as suggested by Bergman and Holmquist (2014). But therein lays the fundamental issue with treating all polls equal in terms of quality. A few high quality polls could be dwarfed and mistaken as poor polls when averaged together with many polls with poor accuracy[9].

To sum up, in order to analyze the data at hand, using Statistics Sweden as a benchmark is the best method available for the specific method employed in this thesis.

[9] Irrespective of how it is measured, such as a simple average or some type of weighted average. Sample size is also a poor predictor of quality in polls; consider the case of the failure of the 1936 Literary Digest polls described earlier in the introduction section.

The accuracy measure

How you make the actual calculation of the accuracy is important, but another question needs to be addressed first. Why focus on accuracy when precision might be as important when making inferences? Precision refers to the spread of values, such as standard deviation. Three important arguments should be noted here. First, polling vote intention is essentially quantifying several dichotomous measurements; the variance in this case is only determined by sample size and the actual binomial distribution. Using precision measurements does not add much in terms of describing a single measurement[10].

Second, the use of measures such as standard deviation and confidence intervals are dependent on parametric assumptions such as having continuous measures, a known corresponding population distribution (for example normal distribution) and low non-response rate, assumptions which are all violated at least to some extent.

Third, Kish (1992) suggests there is an increase in variance when applying weights. If this is the case then there is also a trade-off between accuracy and precision every time. Proper weights would then need to strike a balance between low variance in the resulting cross-categorized cells, while still having good bias reduction properties. It has however been suggested in simulations by Little and Vartivarian (2005), that when covariates are correlated with both *Y* and *P*. When using only the necessary weights, the variance does not increase. The trade-off argument does however have some merit since perfect covariates are rarely available. The potential variance inflation is however not explored in depth in this thesis.

The concept of accuracy can be operationalized in many ways. In the wake of the erroneously called US election of 1948, Mosteller et al. (1949) listed eight different ways of showing how such errors might be described in a two-party setting (Mitofsky, 1998). As Sohlberg et al. (2017) point out, the only applicable one in a multiparty setting is the average absolute deviation (AAD) between each party's (or candidate's) percentage in a poll and the benchmark. There are however a number of problems with the measure. Undecideds are unaccounted for, any systematic party bias is largely obscured and the measure is sensitive to the number of parties, making comparisons difficult over time and across party systems.

There are alternative measures, such as the *predictive accuracy* measure initially suggested by Martin, Traugott, and Kennedy (2005). It is a more model-based approach which is argued to be more applicable to multiparty systems. Empirical application have not yet revealed any significant benefits over Mosteller's original measure (see e.g. Martin et al., 2005; Sohlberg et al., 2017).

Considering the purpose of this study, using AAD should give detailed enough information. The political context is also kept constant, and by extension the number of parties, thus limiting the problems of non-comparability. The number of parties is still an issue though since one large bias might be concealed by other smaller biases and therefore decreasing the overall bias. On the other hand, the more parties that are measured, the lower the chance of "correctly" measuring all of them by chance. As was shown in the Sweden Democrat example in Figure 1, there are cases of both systematic overestimation and underestimation of Sweden Democrats, so any results found will be applicable to both types of biases. The number of

[10] The variance of an estimate of the vote share for a party across many different samples however is of much greater interest.

categories in the vote intention variable is then nine, eight parties plus the *other party* category.

What to do with the undecideds then? Probing is one way of limiting the problem, where the undecideds are asked for leanings towards specific parties (this is done in the PSU). Finally, an alternative method is to assign them randomly among the other parties in even proportions (Visser, Krosnick, Marquette, & Curtin, 2000). It is however common practice to drop them from the analysis, so to emulate standard methods that is what is done in this thesis.

Covariate and analysis strategy

The often arbitrarily chosen weighting covariates, frequently based on availability rather than suitability since factors determining survey response (*P*) and sample inclusion are to a large extent unknown. Knowing if the same factors are also shaping the distribution of a particular *Y* variable adds to the uncertainty. Consider a survey with a skewed gender variable, a demographic unrelated to many measures. Simply having that skew might casts a shadow over the overall validity of the data and therefore the simple adjustment may at least help with greater face validity.

Given the situation above, there are a number of analytic venues that present themselves: 1) following the example of Lee (2006) where the relation between *P* and a *Y* is examined and any redundant covariates are dropped beforehand. What has not been studied however is to 2) determine the full extent to what different choices, both informed ones and the ones less so, may result in. In this thesis, the second method will be employed where all available covariates are used in all possible unique combinations in sets of between one and four covariates (disregarding covariate order, i.e. not permutations).

The covariates used in this study are gender (abbr: G), age (A), education (E), geographical region (R), marital status (M), labor market situation (L), vote recall (V), political interest (I) and a political proposal on accepting fewer refugees into Sweden (P) (see Appendix table 4 for more specifics on the covariate codings). Availability and suitability of variables have both guided the choice of covariates. Since the study aims to emulate practical weighting situations, variables that are known to be poor predictors of both *P* and/or vote intention such as marital status, are also included in the study. Furthermore, to limit the complexity of comparisons between samples, a number of covariates were excluded from the overall analysis, mainly in the non-Citizen Panel samples.

The fact that secondary data is used limits the analysis, but varying actual sample and survey providers supersedes varying covariates. It is also in order to limit the number of combinations, which increases exponentially with each added covariate as a result of studying all covariate permutations.

The number of categories of each covariate is also varied by recoding them into between two a four categories (see Appendix table 4), which is intended to capture another way of how an overall adjustment strategy may be tweaked. It should be noted that four categories is fairly low for some of the covariates, but this is also maintained in order to keep comparability across samples. Gender with its two categories[11], vote recall and vote intention are exceptions. Vote recall has ten categories: the eight parties, *other party* and *did not vote*. Note that

[11] Although an ”other” category for gender has been available since the 2014 election in the Citizen Panel, few respondents choose it, usually below half a percent, so only 2 categories are kept.

respondents who were ineligible to vote in the last election were excluded from the analysis. In most of the samples these groups are very small (due to skewed age distribution) and their vote intention does not deviate enough from the rest of the sample so that it is likely to affect the overall results.

Gender and vote recall plus the seven other covariates that are recoded into three different variants each, means 23 covariates in practice. In the Citizen Panel dataset, the setup produces 6,513 weight combinations in each wave for cell-weights and raked weights both (see Table 4 below)[12]. Since the other datasets have fewer covariates, fewer combinations are produced. All in all, the number of weights is 98,309, including the PSA weights described below.

Table 4. Availability of covariates

	CPCON	*CPVAA*	*CPPROB*	*DS*	*IN*	*UM*
Covariates						
Gender (G)	✓	✓	✓	✓	✓	✓
Age (A)	✓	✓	✓	✓	✓	✓
Education (E)	✓	✓	✓	✓	✓	✓
Region (R)	✓	✓	✓	✓	✓	✓
Marital status (M)	✓	✓	✓			
Labor market situation (L)	✓	✓	✓	(✓)		(✓)
Vote recall (V)	✓	✓	✓	✓	✓	✓
Political interest (I)	✓	✓	✓			
Refugee proposal (P)	✓	✓	✓			

Propensity score adjustment weights

PSA weights may use the same covariates as cell-weights and raked weights, but an altered variant of the cell-weight/raking analysis design is needed to assess the method's relative merit since one of the main strengths of PSA in using a larger set of covariates, including continuous ones. PSA has also been shown to not produce sufficiently good adjustments using only a few covariates.

Like the other techniques, the number of covariates and covariate categories are varied in the fitted logit models, but are kept to five groups of weight types and with a few simplifications[13]. The number of categories in each weight type is the same (except for gender and vote recall). A two-category model will therefore not include a three-category covariate. The number of bins (or quantiles) that form the basis for the PSA weights are also varied between 2 and 20, which also allows for testing the number of quintiles as well. As illustrated in Table 5, the models are built up in a stepwise manner by adding a new covariate to each new weight type.

[12] The number of combinations without repetition is calculated as: $\frac{n!}{r!(n-r)!}$, where n is the number of covariates and r the number of covariates in the set.

[13] Running one PSA model takes about 1 minute on average. Creating each combination of weights in one wave in the first weight model in Table 5 therefore takes about 13 hours ($1\,\text{min}\times\frac{12!}{5!(12-5)!}$). The overall analysis would therefore take prohibitively long to run with the current method and hardware.

Table 5. Propensity score adjustment weight setup

	No of covs	*Covariates included in the model*	*Weight acronym*	*Available in:*					
				CPCON	*CPVAA*	*CPPROB*	*DS*	*IN*	*UM*
Weight type 1	5	gender, age, education, region, labor market situation	GAER	✓	✓	✓	✓	✓	✓
Weight type 2	6	gender, age, education, region, labor market situation, marital status	GAERLM	✓	✓	✓			
Weight type 3	6	gender, age, education, region, vote recall	GAERV	✓	✓	✓	✓	✓	✓
Weight type 4	7	gender, age, education, region, labor market situation, marital status, vote recall	GAERLMV	✓	✓	✓			
Weight type 5	9	gender, age, education, region, labor market situation, marital status, vote recall, political interest, discussing politics	GAERLMVID	✓	✓	✓			

Comment: Weight type 1–3 contain the same variables as used when creating cell proportion and raked weights, but type 4 and 5 includes a continuous age variable and how often the respondent discussed politics the past year.

Two points are worth mentioning here. First, PSA models are generally fitted with more covariates and any covariates that contribute to the model goodness-of-fit is likely to be beneficial to the weights. However, in order to compare the models in different datasets, the number had to be cut down significantly. This is likely to impede the conditions for this study to examine the maximum potential of PSA.

Second, a standard routine among survey practitioners is to use several of the weighting methods listed above in tandem both before (pre-stratification) and after data collection (post-adjustment). The aim here is however to try to compare the "effect" of each weighting technique separately, even though the inherent nature of them might not lend itself to easy comparisons. Combinations of weighting techniques are therefore excluded from this study.

4 Results

The results section will provide formal tests of the eight hypotheses formulated in the theory chapter. First, it provides descriptive statistics of how large the biases are before and after adjustments has been made and then it describes bivariate relationships to examine the hypotheses. Second, multivariate analyses of the post-adjustment biases are performed to confirm or reject the hypotheses. The focus is largely on cell-weights and raked weights, but it is followed up by a shorter PSA analysis.

Descriptive statistics

Table 6 describes the weighting results in a number of ways, giving an initial indication of whether the weighting technique hypotheses *H2a* and *H2b and* sample hypotheses *H3a* and *H3b* are substantiated or not. These hypotheses are more closely examined in the multivariate regression models which follow, with the addition of covariate hypotheses *H1a*, *H1b*, *H1c* and *H1d*.

First, in order to provide comparable results across both surveys and methods, the six leftmost columns of Table 6 report the averages using a single weight variant that only contains gender, age, education and region covariates, each coded into two categories. The next set of six columns reports all weights using gender, age, education, region and vote recall in different combinations. These are the comparable results across all surveys and methods, except for when using the PSA technique. The last set reports the use of all weights including all PSA weights, but limits the data to Citizen Panel samples only.

There is considerable variation in both unweighted bias—the highest in CPCON and lowest in the Demoskop sample (DS)—as well as adjustment effects. The post-weighting absolute bias, bias reduction and proportional bias reduction together form the picture that almost all samples can be adjusted to better conform to the benchmark on average. The absolute accuracy improvement varies much, but the proportional reduction—which has a different unit of measurement than the other two measures—is fairly constant across polling organization. DS is an exception, which might be due a roof effect by already being accurate enough than many adjustments makes it more inaccurate. The table also conveys the fact that weighting technique matters little here, hinting at a rejection of *H2a*.

Looking at the comparable cell-weights (the middle six columns), the post-adjustment biases in the Citizen Panel samples follow a general pattern where the convenience sample (CPCON) is the highest with 4.12 percentage points bias on average across ten categories, followed by the VAA sample (CPVAA) with 3.48, while the probability sample is the lowest (CPPROB) at 2.83. The high level of bias was expected since none of these samples were prestratified. *H3a* is thus provided with an initial confirmation on a bivariate basis.

The prestratified samples follow a similar pattern, although at a lower level of post-adjustment bias. Inizio (IN) has the highest bias of 3.20 percentage points, followed by United Minds' (UM) at 1.89, both of which are nonprobability samples. DS's probability samples produce the lowest post-adjustment biases (1.60) of all samples, which is also in line with *H3a*.

Hypothesis *H3b*, which states that reverse is true when looking at absolute bias reduction, is confirmed, but the result is less clear in proportional terms. Looking at the same comparable results from cell-weights, the cell-weight numbers for CPCON are a reduction of bias by on average 1.57 percentage points, –0.90 for CPVAA and –0.84 for CPPROB, but the proportional reduction is similar: –28, –21 and –23 percent of the initial bias is removed, respectively. It seems as if the reduction is a function of the initial bias than sample type per se.

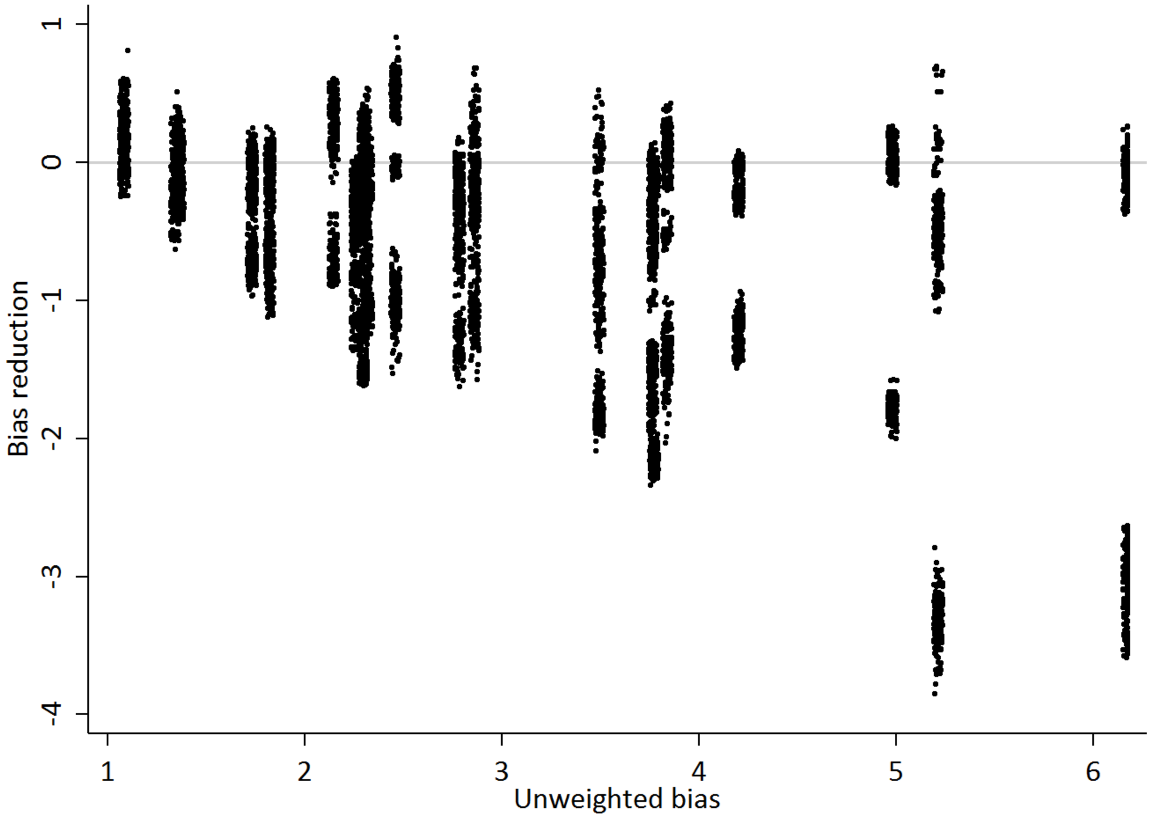
Plotting the bias reduction of all comparable weights by their initial unweighted bias does indeed reveal a negative correlation, as shown in Figure 2 (r=–0.44, *p*=0.00), but it also shows substantial heterogeneity where 20 percent of the weights are actually making the measure *less* accurate. A third point is that there is a distinct clustering, a heteroscedasticity pattern, especially on the higher end of the unweighted bias scale. The clustering is due to the presence of the vote recall covariate (see Appendix Figure 9), a fact that will be followed-up upon below. The sample type is likely to be a significant factor for determining the initial bias through sample selection mechanisms, but the specifics are unknown and the samples variation is too low to further pursue the question.

Table 6. Mean absolute bias in percentage points before and after weighting, mean bias reduction and mean proportional bias reduction

		Weight: Gender (2 cat), age (2), edu (2), region (2)						Comparable cell-weights/raked weights: gender, age, education, region, vote recall						Full set of weights					
		CPCON	*CPVAA*	*CPPR.*	*IN*	*UM*	*DS*	*CPCON*	*CPVAA*	*CPPR.*	*IN*	*UM*	*DS*	*CPCON*	*CPVAA*	*CPPR.*	*IN*	*UM*	*DS*
	Unweighted bias	5.70	4.38	3.67	3.99	2.32	1.65	5.70	4.38	3.67	3.99	2.32	1.65	5.70	4.38	3.67	3.99	2.32	1.65
Cell-weight	Bias after weighting	5.33	4.27	3.50	3.57	2.39	1.81	4.12	3.48	2.83	3.20	1.89	1.60	4.39	3.56	2.85	3.20	1.91	1.96
		(1.00)	(0.90)	(0.63)	(0.35)	(0.46)	(0.72)	(1.63)	(1.22)	(0.82)	(0.70)	(0.61)	(0.49)	(1.31)	(1.04)	(0.71)	(0.70)	(0.57)	(0.85)
	Bias reduction	−0.36	−0.10	−0.17	−0.41	+0.07	+0.16	−1.57	−0.90	−0.84	−0.79	−0.43	−0.05	−1.30	−0.82	−0.82	−0.79	−0.40	+0.31
		(0.32	(0.05)	(0.38)	(0.04)	(0.26)	(0.21)	(1.51)	(0.99)	(0.76)	(0.61)	(0.52)	(0.35)	(1.17)	(0.78)	(0.66)	(0.61)	(0.48)	(0.75)
	Prop. bias reduction	−7	−3	−5	−10	+3	8	−28	−21	−23	−20	−19	−1	−23	−19	−23	−20	−18	21
		(6)	(2)	(11)	(2)	(11)	(9)	(27)	(24)	(21)	(16)	(22)	(20)	(21)	(19)	(18)	(16)	(20)	(44)
Raked weights	Bias after weighting	5.29	4.30	3.39	3.55	2.37	1.71	4.14	3.50	2.75	3.17	1.81	1.49	4.39	3.59	2.79	3.17	1.87	1.78
		(1.06)	(0.87)	(0.59)	(0.58)	(0.44)	(0.65)	(1.59)	(1.19)	(0.83)	(0.69)	(0.66)	(0.49)	(1.25)	(1.02)	(0.66)	(0.69)	(0.64)	(0.72)
	Bias reduction	−0.40	−0.08	−0.27	−0.44	0.05	0.06	−1.56	−0.88	−0.92	−0.81	−0.51	−0.16	−1.30	−0.79	−0.88	−0.81	−0.44	+0.13
		(0.38)	(0.02)	(0.35)	(0.27)	(0.23)	(0.20)	(1.46)	(0.96)	(0.75)	(0.59)	(0.56)	(0.37)	(1.12)	(0.75)	(0.59)	(0.59)	(0.53)	(0.64)
	Prop. bias reduction	−8	−2	−8	−11	+2	3	−28	−21	−25	−21	−22	−8	−23	−19	−24	−21	−20	11
		(8)	(1)	(10)	(8)	(10)	(10)	(26)	(23)	(21)	(15)	(24)	(21)	(20)	(18)	(17)	(15)	(23)	(35)
PSA weights	Bias after weighting	5.33	4.30	3.47	3.88	2.42	2.36							3.71	3.10	2.68	4.00	1.93	2.08
		(0.72)	(0.62)	(0.42)	(0.17)	(0.42)	(0.70)							(1.57)	(1.25)	(0.77)	(0.64)	(0.61)	(0.66)
	Bias reduction	−0.37	−0.08	−0.19	−0.10	+0.10	+0.71							−1.93	−1.21	−0.97	+0.02	−0.39	+0.43
		(0.23)	(0.03)	(0.24)	(0.39)	(0.23)	(0.24)							(1.41)	(1.02)	(0.69)	(0.71)	(0.53)	(0.49)
	Prop. bias reduction	−7	−2	−6	−2	+5	43							−35	−29	−27	+1	−17	29
		(5)	(1)	(7)	(10)	(10)	(9)							(26)	(25)	(19)	(18)	(22)	(31)
N per cell/raked weight		2	2	2	5	2	8	456	456	456	865	456	1,824	13,026	13,026	13,026	2,620	456	5,376
N – PSA		38	38	38	95	38	152							513	513	513	570	570	912

Comment: Comment: Standard deviations are shown in parentheses. Bias is calculated as the average absolute bias between the percentage points of all parties in the benchmark (PSU) and the weighted survey. Bias reduction is calculated as the difference between the unweighted bias and the weighted bias. Proportional bias reduction is calculated as relative size of the bias reduction proportion in relation to the unweighted bias.

Figure 2. Scatterplot of unweighted absolute bias and bias reduction



Comment: N=9,026. The jitter (1) option of the Stata command -scatter- is used. Including all three types of weights.

Moving on to weighting techniques, it was hypothesized that cell-weighting would provide more effective weights than raked weights in *H2a*. The hypothesis is not at all substantiated, which runs contrary to the precision loss when using raked weights hypothesized by Kalton and Flores-Cervantes (1998, p. 84). If anything, raked weights are somewhat more effective, although the bias reduction of the two techniques differs little on average.

The largest absolute difference can be found in the DS sample where bias is reduced by on average 0.11 percentage points more when using raked weights than cell-weights (*p*=0.00) and the smallest difference is CPCON with a 0 point difference (p=0.30). The same is true when measuring proportional bias reduction, but more pronounced.

Table 7. t-tests of differences between the weight effects of raked weights and cell-weights, comparable weights (proportional reduction).

	Comparable sample			*Full sample*		
	Diff – bias reduction	Diff – prop bias reduction	df	Diff – bias reduction	Diff – prop bias reduction	df
CPCON	+0.02*	+0.3*	455	0.00	+0.1*	13,025
CPVAA	+0.02*	+0.5*	455	+0.03*	+0.6*	13,025
CPPROB	−0.08*	−2.2*	455	−0.05*	−1.5*	13,025
DS	−0.11*	−7.1*	864	−0.17*	−10.4*	2,619
IN	−0.02*	−0.6*	455	−0.02*	−0.6*	455

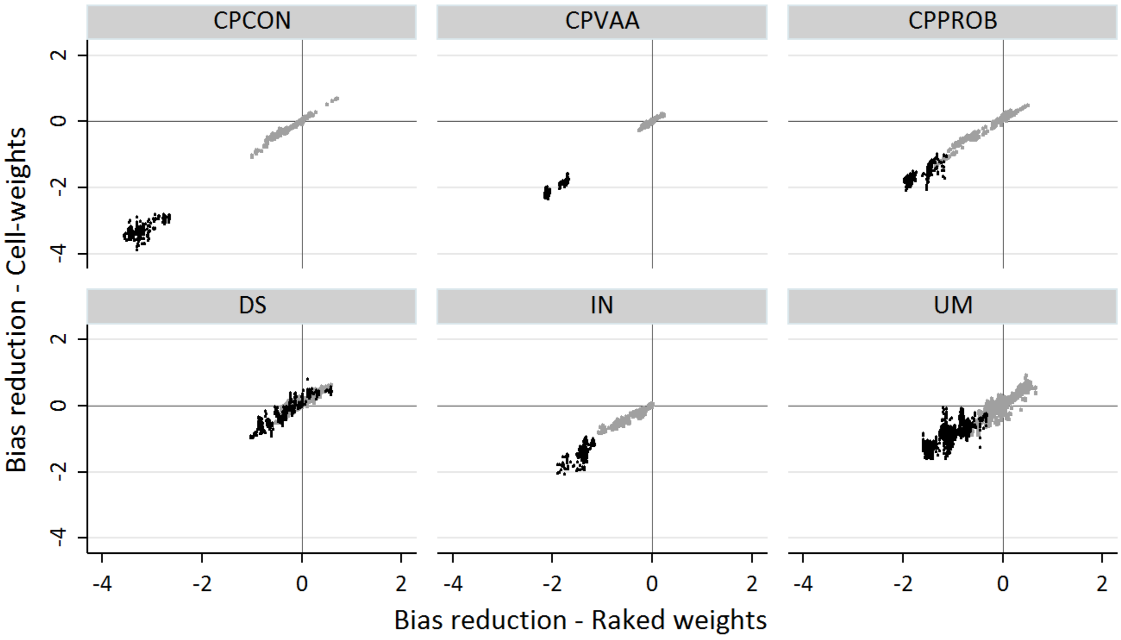
UM	−0.08*	−3.6*	1,823	−0.04*	−2.0*	5,375

Comment: The differences are measured as the proportional effects of raked weights subtracted by cell-weights. Negative differences indicate that the reduction is greater when using raked weights. * p<.001.

In the specific cases tested here, on average, the use of marginal distribution(s) or joint cell distributions matters surprisingly little for how effective a weight is in dividing up the sample into homogenous subgroups in terms of response propensities and vote intention. If it had mattered, cell-weights would have outperformed raked weights. Raked weights are however slightly better which might be explained by a "less is more" effect, where raked weights are "gentler" and therefore provides more conservative weights. From this perspective, cell-weights could be described as more prone to introducing bias when the covariates are not predictive of the bias.

Figure 3 plots cell-weights and raked weights and the corresponding correlations. It illustrates that there is at least some variation in the differences between cell-weights and raked weights, but closer scrutiny reveal normal distributions, indicating that the difference might be more random than anything else (see Appendix Figure 10).

Figure 3. Scatterplots of cell-weights and raked weights by data source



Correlations:
CPOI: r = 0.996 (0.98, 0.71), CPVAA: r = 0.999 (0.96, 0.94),
CPPROB: r = 0.986 (0.98, 0.78), DS: r = 0.929 (0.91, 0.92),
IN: r = 0.970 (0.95, 0.70), UM: r = 0.929 (0.88, 0.68)

Comment: N=27,362. Non-vote recall weights are the grey dots, while vote recall weights are black. Pearson's *r* for each data source is reported below the plots, first the overall coefficient, then followed by non-vote recall weights and vote recall in parentheses.

H2b says that PSA should provide less bias reduction when the same set of weights covariates is used, but more when a full model is employed. Looking again at the first set of columns of Table 6 where only one weight variant is used (using the PSA average across bins), the first part of the hypothesis is lent some support: weight effects are smaller using PSA than both raking (–0.22 percentage points: t(20)=–3.02, *p*=0.01) and cell-weights (–0.17 percentage points: t(20)=–2.88, *p*=0.01). When looking at the averages, the second part of the hypothesis is true for three of the six sample sources: all the three Citizen Panel samples.

In order to get an idea of what the post-adjustment distributions look like, Figure 4 through Figure 6 below show the survey by survey distribution of the post-weight biases (for the distribution of the full sample of Citizen Panel weights, see the histogram in Appendix Figure 11). Again, the clustering of vote recall and non-vote recall weights is striking in almost all cases, except a few of the Demoskop surveys, most manifested in the earlier measurements, i.e. the ones the furthest removed from its actual election. It should be emphasized that the results when using the actual vote results as a benchmark (DS in September 2014) exhibit the same pattern.

Finally, the covariate hypotheses are best analyzed under a multinomial framework, which will be done in the following subsection.

Figure 4. Histograms of post-weighting bias by Citizen Panel and Inizio surveys – comparable weights

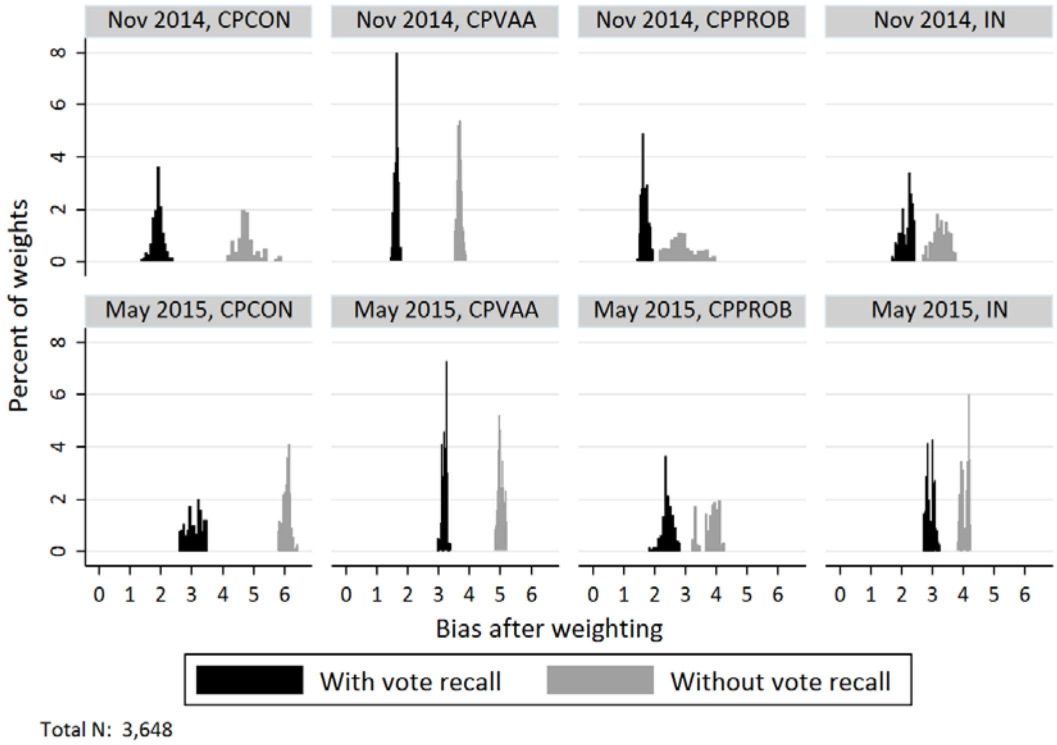


Figure 5. Histograms of post-weighting bias by Demoskop surveys

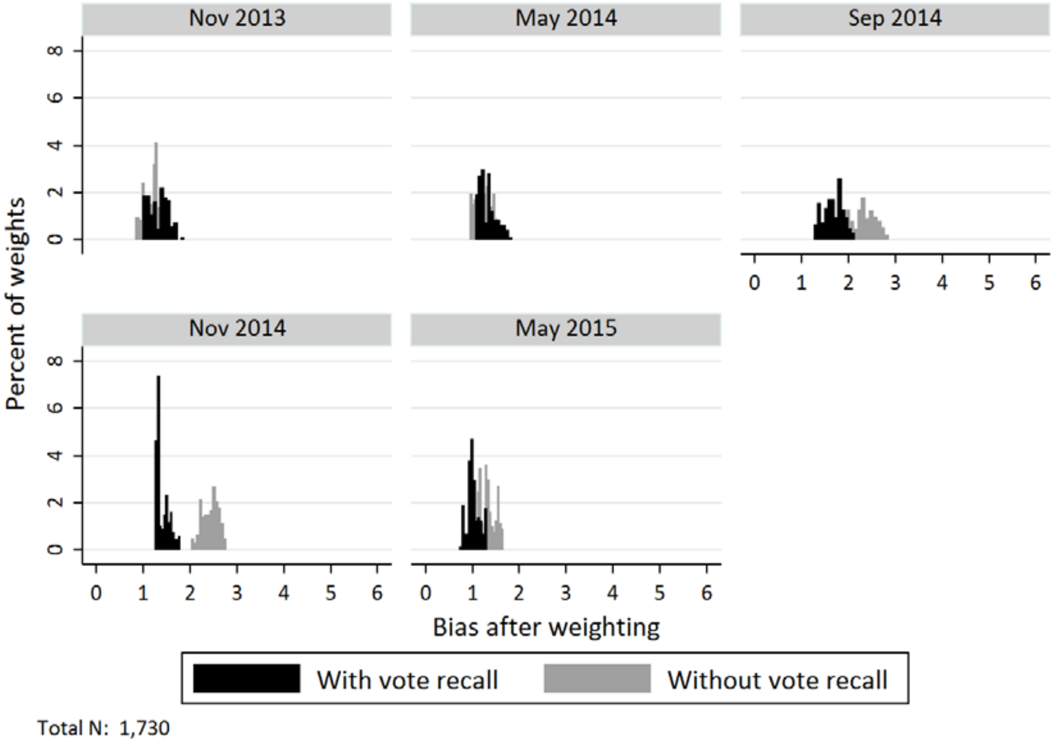
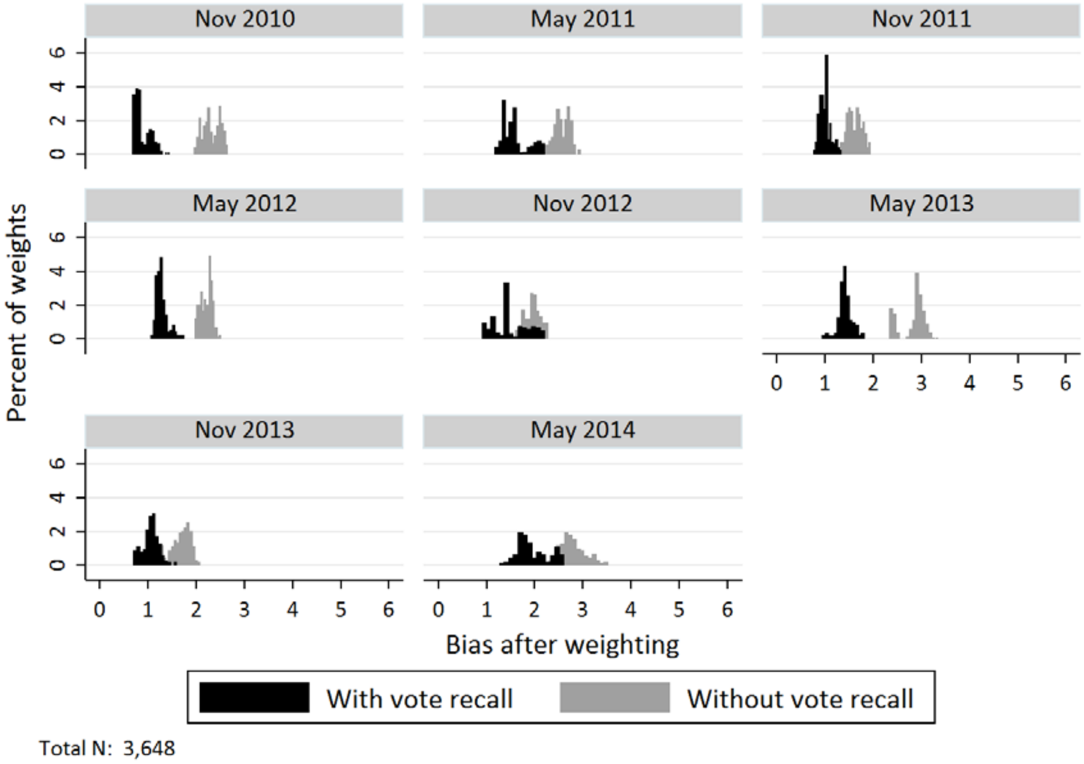


Figure 6. Histograms of post-weighting bias by United Minds surveys



Multivariate analyses

Four ordinary least squares (OLS) regressions are estimated with weight effects in proportional bias reduction as the dependent variable (DV), i.e. the bias reduction as percent of the unweighted bias level. The results are reported in Table 8 below. The proportional rather than actual percentage point reduction is chosen as dependent variable since the former is more independent of the initial bias and is easier to interpret (for the same model with absolute weight effect as DV, see Appendix table 6, a set of models which produces similar results).

At its core, the data is based on 21 vote intention estimates from as many surveys. As a result of this design, the weight effects are not independent within each survey, but instead directly linked to each survey's unweighted measurement. The weight effects could also be influenced by how the data was collected, i.e. the house effects within each sample source.

A multilevel regression (mixed) was fitted with three levels: weights are nested within surveys, which in turn are nested within sample source (result not reported here). It showed that the unique variance provided by sample source was close to zero and dwarfed by between-survey variance. Standard errors are therefore clustered by survey to produce more reliable variance estimates. A comparison between a naïve OLS and OLS with clustered standard errors does indeed produce very different variance estimates, with the former having almost only significant coefficients, while the latter has wider and likely more credible confidence intervals. To lower multicollinearity, the *months since last election* measure is mean-deviated.

Table 8 below reports four models, each serving a specific purpose. In order to illustrate how the regression constant should be interpreted, only the very basic weight covariates are analyzed in model 1. Model 2 is a full-fledged model with all data sources, but is restricted to fewer covariates than model 3 due to lower availability. Model 3 is in turn limited to Citizen Panel samples only, but has the full set of weighting covariates. Lastly, model 4 shows the results for the September 2014 DS sample which uses the election results as benchmark rather than a survey.

The reference category for all covariate variables is gender and the results in model 1 should therefore be interpreted in relation to the weight effects of the gender covariate. Model 1 features only covariate dummy variables, which means that the constant of –1.4 should be interpreted as an average of 1.4 percent reduction of the initial unweighted bias when including gender (the reference category) in a weight, on its own and in combinations with the age, education and region covariates.

A hypothetical unweighted bias of an average of 2 *percentage points* across the 10 party categories in the vote intention measure would only be decreased to 1.97 when only employing the gender covariate $(2 - 2 \times 0.014)$. None of the other three covariates do improve the accuracy much either, and age is even leaving the accuracy worse off by about 6 percent when controlling for the inclusion of the other covariates $(-0.014 + 0.079)$. The results may be surprising considering that age and education are known to often predict survey nonresponse, but it confirms earlier studies that demographics are often poor predictors of nonresponse bias. The covariates are simply not predictive enough of vote intention in the studied surveys and show the waning influence of demographics in Swedish party preference structure.

Table 8. OLS regressions with proportional bias reduction as dependent variable (percentage points from benchmark), cell-weight and raked weights.

	Model 1: Simple – all sample sources		Model 2: Full – all sample sources		Model 3: Full – Citizen Panel only		Model 4: DS Sept 2014 only – Election result benchmark	
	b	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Demographic covariates[a]								
Age (A)	+7.9***	(1.5)	+8.0***	(1.3)	+4.7**	(1.1)	+10.8***	(0.4)
Education (E)	−7.8***	(1.6)	−3.2*	(1.3)	−4.6	(2.8)	−11.6***	(0.4)
Region (R)	−2.5*	(1.1)	−0.4	(1.1)	−1.6	(1.0)	−0.4	(0.4)
Labor market situation (L)					−1.1	(0.7)	−1.0*	(0.5)
Marital status (M)					−0.8	(0.5)		
Psychographic/vote covariates[a]								
Vote recall (V)			−43.5***	(4.0)	−45.4***	(5.6)	−28.7***	(0.5)
Months since last election (2/48)			0.0	(0.1)	+1.2**	(0.2)		
Vote recall × months since last election			+0.6*	(0.2)	+1.1	(0.9)		
Political interest (I)					−10.7***	(1.1)		
Refugee policy proposal (P)					−4.3**	(1.0)		
Weight paradata[b]								
2 covariates			−1.8*	(0.8)	−1.4**	(0.2)	−0.6	(1.0)
3 covariates			−2.8*	(1.3)	−1.8*	(0.6)	+0.4	(1.1)
4 covariates			−3.5	(1.8)	−0.9	(1.1)	+2.3	(1.3)
Number of weight cells (standardized 0–1)			−1.6	(2.7)	−7.0**	(1.2)	+2.2	(1.4)
Weighting technique[c]								
Raked weight			−3.0***	(0.8)	−0.3	(0.5)	−4.0***	(0.3)
Sample[d]								
CPVAA			+6.9	(4.8)	+4.1***	(0.6)		
CPPROB			+3.5	(6.0)	−0.2	(0.6)		
IN			+7.5	(4.6)				
UM			+3.5	(3.6)				
DM			+18.1***	(4.4)				
Constant	−1.4	(0.7)	−8.4*	(3.4)	−9.6**	(2.2)	+0.7	(0.9)
R2	0.190		0.736		0.889		0.917	
N	5,014		9,026		78,156		1,048	
RMSE	10.3		12.0		6.3		4.7	

Comment: Standard errors are clustered by survey in model 1–3, but model 4 is not clustered since it uses only one suryey sample. * p<.05; ** p<.01; *** p<.001. Reference categories: [a]Gender (G). [b]1 covariate. [c]Cell-weight. [d]CPCON. The DV is the proportional effect of weighting on the average absolute deviation from the benchmark. Each observation is a unique weight using different combinations of variables. Gender through vote recall are dummy variables (0/1). The number-of-cells covariate, which indicates the number of unique cell combinations that can be made with the variables used, is standardized from 0 to 1 (it ranges in practice from 2 cells to 576).

Model 2 introduces various metadata variables as well as weights with vote recall. The comparable covariate coefficients from model 1 remain largely unaffected here, if somewhat less significant. Most of the hypotheses may be finally tested here. *H3b*, which states that the bias reduction effect is larger in nonprobability samples than probability samples, is confirmed when estimating the same models with absolute post-adjustment biases as DV (not shown) and *H2a*, which hypothesize that cell-weights reduce more bias than raked weights, is still confirmed.

Hypotheses *H1a* and *H1b* suggest that adding more covariates and more weight cells (more "finely grained" covariates) reduce bias more than fewer covariates and less weight cells. The short answer is that such effects are very small or cannot be found. For example, having 4 covariates instead of 1 only results in 3.5 percent more accurate measurements, and the difference is not significant at the 95 percent level (p=0.06). The independent effect of number of cells is also nonexistent. Both *H1a* and *H1b* should therefore be at least partially rejected according to the results from model 2. Simply relying on a larger set of covariates is not enough to provide good measurements.

Vote recall is however shown to be a very effective bias reduction tool. Including a vote recall covariate decreases a bias 2.00 percentage points bias to 1.23, a significant improvement, way outperforming all other covariates. It confirms the vote recall part of hypothesis *H1c*. It is also true that there is an interaction effect between the how far removed the election is and the bias reduction efficiency of adjusting for vote recall. Assuming a linear interaction effect[14], the 43.5 percent reduction of using vote recall would be reduced to 29.1 after two years ($-43.5 + 0.6 \times 24$) and 14.7 by the next election (the term office in Sweden is 4 years). Regardless, it shows that the further apart the measurement is from the election, the less efficient will the covariate be, an effect of the overall decrease of vote recall accuracy due to respondents misremembering their past vote or deliberate misreporting. Therefore *H1d* is confirmed, although it should be emphasized that the interaction effect results are tentative, since the within-source provider variance for *time since last election* is fairly low. Note also that all samples except IN use measurements of vote recall from the same survey. Had the panel component been utilized in the other cases, an effect would be less likely to be found.

Model 3 adds one demographic covariate, marital status, and two psychographic measures: political interest and refugee policy preference, but limits the data to Citizen Panel only. Marital status, which is known to be largely unrelated to both vote intention and to survey participation, is indeed found to be an ineffective covariate. More interestingly, political interest and refugee policy are as effective as or more effective than the (nonsignificant) education covariate, thus at least partially confirming *H1c*. It indicates that psychographics should not be dismissed as irrelevant weighting covariates. Note that the number-of-cells coefficient is significant here, while the raking coefficient is not, which differs from model 2, casting doubt on the stability of those results.

Since there is a concern that the results might only be generalizable to the survey benchmark used here, model 4 principally replicates the results using the 2014 national parliamentary election instead. Vote recall is however somewhat less effective, while education is more so.

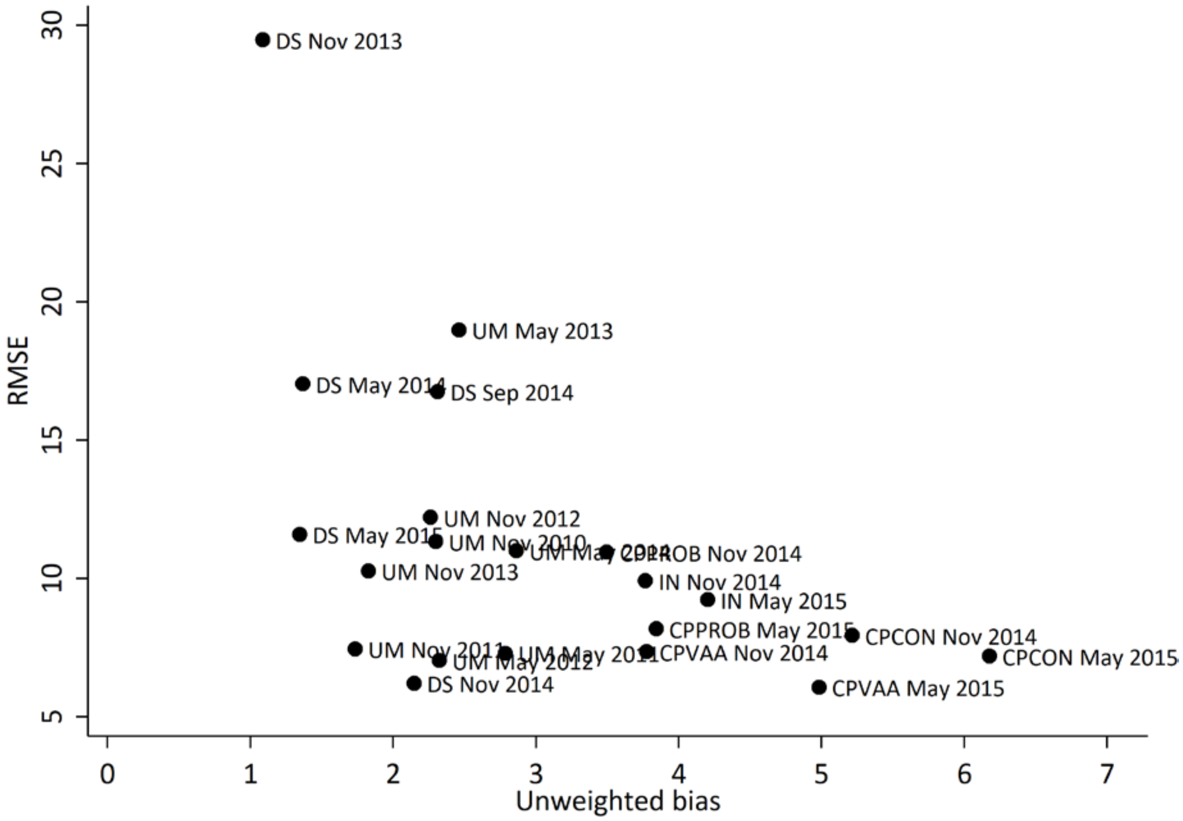
Sample size is not used as a control since the smallest sample is about 1,000 and should therefore have enough statistical power given standard random sampling theory. Furthermore, when most of the samples are derived in a nonprobabilistic fashion, sample sizes are poor

[14] The interaction seem to be exponentially decaying in an increasing form given the data, see Figure 13 in the appendix.

predictors (consider the Literary Digest discussed in the introduction) due to the ease by which nonprobability samples can collect very large samples.

To assess the validity of the OLS results, a diagnosis of model residuals indicate that they are all homoscedastic, except for the a few residuals for model 2 close to the fitted zero value (i.e. zero bias reduction, see Appendix Figure 12). A closer examination reveals that one common denominator of survey measurements with a less fitting model is—although not explaining the relation perfectly—the unweighted bias. Figure 8 plots the survey level root mean square errors (RMSE), which is basically the standard deviation of the model's unexplained variance. Still assuming that the benchmark is less biased than the weighted surveys, the plot shows that with less initial bias, the worse is the model fit (greater RMSE). It could mean that bias adjustments of surveys which are closer to the benchmark are more volatile and need be performed with more care as not to worsen the accuracy. This short diagnostic analysis indicates that any deviations found are likely not enough to risk invalidating the results, but it reveals that the full models has omitted variable bias, a lack of important predictors to explain the full range of observations. It is particularly true for a majority of the DS samples.

Figure 7. Root mean square error of model 2 for each survey by unweighted bias



Up until now, the focus in the multivariate analyses has been on cell- and raked weights only. Before moving on to the final discussion and conclusion section, an analysis of PSA follows under the next heading.

Propensity score adjustment

Table 9 reports the bias reduction properties of a set of PSA weights. As was described more in depth earlier in the methodology section, the number of weight iterations is scaled down in comparison to the analysis above. Rather than being a full-fledged study in its own right, it is designed mainly to provide a comparison point to the cell- and raked weight analyses above. The reference category is the cell-weight and raked weight GAER: gender, age, education and region, with all covariates having either 2, 3 or 4 categories. The *PSA weight types* with less covariates—number 1 and 2—are generally about as effective as or somewhat less effective than the reference category. Similarly to the cell-weight and raked weight analyses earlier, adding vote recall as in the third weight type results in a substantial reduction of bias. Interestingly, the addition of more covariates in weight type 4 and 5 does little to reduce the bias much further. The effects are by all estimates close to the results found for the other weight techniques, suggesting that the technique itself is of lesser importance.

On a side note, when creating PSA weights, the propensities are divided into quantiles, or bins, which in practice equals the number of actual different "cells". The five bins suggested by Cochran (1968) is used as the reference category in the analysis and is indeed better than 2–4 and but somewhat less useful than 6 or more bins (see marginal effects in Appendix Figure 15).

Table 9. OLS regressions with proportional bias reduction as dependent variable (percentage points from benchmark), cell-weight and PSA weights

	Model 1		Model 2	
Sample	All sample sources		Citizen Panel only	
	b	SE	b	SE
Weight type[a]				
PSA weight type 1: GAER	+9.8*	(3.7)	−1.1	(0.4)
PSA weight type 2: GAERML			−3.3**	(0.7)
PSA weight type 3: GAERV	−28.1***	(5.4)	−46.3***	(3.9)
PSA weight type 4: GAERMLV			−46.2***	(4.1)
PSA weight type 5: GAERMLVIP			−48.9***	(3.7)
Sample[b]				
CPVAA	+5.0	(5.1)	+5.8	(5.4)
CPPROB	+6.8	(6.0)	+7.9	(6.1)
IN	+30.2***	(7.0)		
UM	+13.6**	(4.5)		
DS	+58.0***	(7.2)		
Number of covariate categories[c]				
3 categories	+1.0	(2.0)	+2.2***	(0.3)
4 categories	+4.4	(2.3)	+1.0	(0.5)
Number of bins[d]				
2–4 bins	+4.2***	(0.5)	+3.2*	(0.9)
6 bins	−0.1	(0.3)	−0.4	(0.4)
7 bins	−1.3	(0.7)	−0.5	(0.5)
8 bins	−1.4	(0.8)	−0.5	(0.3)
9 bins	−2.4*	(0.9)	−1.1	(0.5)
10 bins	−2.3**	(0.8)	−1.0	(0.6)
11–20 bins	−2.4*	(0.9)	−0.8	(0.5)
Constant	−21.8**	(5.8)	−8.9	(4.6)

R2	0.757	0.903
N	2,510	1,575
RMSE	15.9	7.4

Comment: See Table 5 for the PSA weight setup. G=gender, A=age, E=education, R=region, L=labor market situation, M=marital status, V=vote recall, I=political interest, P=refugee policy proposal. Reference categories: a. GAER (cell-weight), b. CPCON, c. 2 categories, d. 5 bins.

In short, the models largely confirm *H1c, H1d, H2b, H3a* and *H3b*, partially confirm *H1a*, while rejecting *H2a* and *H1b*.

5 Discussion and summary

A number of high profile cases of polling in 2016 were wrongly called. In the wake of poorly forecasted elections, or where the framing is that they were poor, the same type of discussion crops up: what is wrong with polls and what can be done to improve the accuracy?

To sum up the results from 21 Swedish surveys, measurements of vote intention are generally more biased in nonprobability than probability samples, but biases can be significantly decreased through weighting adjustments. The specific weighting technique—cell weighting, raking or propensity score adjustment—matters little. However, the choice of what parameters, covariates, to adjust is paramount: vote recall is by far the most efficient covariate, as was expected, while demographics such as education and age do little good, relatively speaking. How many covariates and how many categories in each covariate also do not have any impact.

Although adjustments are sometimes skillfully implemented and biases are indeed decreased, it is arguably more often the case that adjustments are not so well-thought-through: when the end-users, researchers and analysts have little knowledge what makes people answer surveys or what predicts the variable of interest (*Y*), the implementation of weights is bound to be a gamble. A unique feature of this study and the first contribution is that it explores the arbitrariness of adjusting a variable of interest by actually making every possible weighting implementation and examines the results.

Consider the case of the European Social Survey omnibus survey that comes with post-stratification weights to nonresponse bias (using variants of a GAER weight: European Social Survey, 2014). As the weights need to be as general as possible due to the diverse measurements (*Y*) contained in that type of survey, weighting covariates (*Z*) can only be chosen based on what might explain response propensities (*P*) and not all the *Y*s. *Z*s should however be correlated with both *P* and *Y* to decrease the *Y* bias. Those weights cannot be a good match for all, since the same *Z* vector will not be equally correlated with all *Y* measurements.

Earlier studies have focused on varying *Y* (see e.g. Yeager et al., 2011), but have typically kept *Z* constant or almost constant. This study is instead concentrated on the full impact of various combinations of *Z* vectors in different samples, while keeping *Y* constant. The results show that there is an actual risk involved in misspecifying weights, particularly when the accuracy is already fairly good. When looking at comparable weights (*N*=9,026) 20 percent of the weights are actually making the accuracy worse off (*Median* proportional bias increase=7 percent). These results stress the importance of making informed decisions about how to make bias adjustments, not only because there is a risk of increasing variance estimates, but because point estimates may be severely biased.

Similar to studies such as Tourangeau et al. (2013), this investigation proves that it is possible to remove a good portion of the bias. What is striking, however, is that vote recall so systematically outperforms all other covariates, irrespective of sample type: the average bias reduction using a weight with vote recall removes about half of the bias. Among the comparable weights, the overall best weight removed 72 percent of the bias of one sample (in the November 2014 CPCON sample using GVE3R4; proportional bias reduction: −71.6%; average percentage point bias before/after adjustment: 5.22/1.48; see also Appendix Figure 14 for a visualization of the weight effects). Even though this study is not designed to cherry-pick results, that specific combination of covariates could be taken as the weight *par préférence*, at least during this period: the median percentile rank of the weight's bias reduction across all 21 surveys was 89 (the 89th percentile), see Appendix table 8 for more details.

A second contribution concerns psychographics in general and vote recall in particular, a measure which often is argued to be unreliable (Cohn, 2016). The argument is that when there are large movements in the electorate between elections, these will bias the relation between the recalled past electoral behavior and the current intended vote. It assumes that that 1) the respondent misremembers past vote, willingly or no, and that 2) vote recall is measured in the same survey. The first point has some merit, this study does indeed suggest that it is gradually less effective over time, but the net bias reduction is still far from removed even after 4 years. Although, it is possible that the consistency is also decreasing with time, with the risk of larger errors. Regarding the second point, most panel studies actually employ measurements that where collected right after the election, and thus remove most of the worries about bandwagon effects.

While vote recall seem to be an important puzzle piece to solve the low accuracy of vote intention measurements, it is perhaps surprising how small the improvements by most demographic covariates are. Psychographic covariates, such as political interest and refugee policy attitudes seem to have more potential. Three aspects could provide an explanation. First, consider that demographics can be argued to be antecedents to most psychographics. Second, the idea of individualization of society (Thomassen, 2005) theorizes that the relation between social groups and party preferences will weaken. Third, the leverage-saliency theory of survey response (Groves et al., 2000) describes a situation of gradually fragmented reasons for answering survey questions, thus limiting the predictive capacity of demographics and potentially increasing the influence of certain psychographics—such as interest in the topics featured in the survey respondents are asked to participate in. Such variables might better predict what is salient when deciding whether to respond to a survey or not. Interestingly, the general idea of individualization of society and the leverage-saliency theory do to some extent describe the same development, which promises a theory synthesis in future research.

An example of the use of psychographics is that the Swedish branch of Ipsos recently started to use GAL-TAN covariates (Green, Alternative, Liberal, Traditional, Authoritarian and Nationalistic) to adjust their probability web vote intention estimates.

A third contribution of this thesis is that the specific weighting technique used matters little for the accuracy outcome; cell-weighting and raked weights produce very similar results. A possible reason behind this result is that the covariate interactions are not strong enough to impact vote intention. An alternative explanation is non-converging raking procedures, i.e.

when the raking algorithm cannot find equilibrium where all margins are concurrently equal to the benchmark margins. The issue is likely to appear more as the total number of margin cells increase, but it has not been sufficiently documented, and therefore not controlled for in the analyses. The corresponding unchecked problem when using cell-weights is empty cells, which is usually fixed by collapsing them with close-by cell. This is also not corrected here, and it is therefore likely that some weights might be affected by it.

It could also be that raking produces more conservative weights in general, glossing over any misinformed use of weights. This could be of some importance when digging deeper into the reason behind the cell-weight/raked weight difference found here. There are however no indications that the difference between the methods diverge more when the number of cells increase, so the consequences of expanding the analysis is not certain to produce diverging results. PSA also seem to produce results close to that of the other two techniques, although the comparison is not as easily done. It should however also be reiterated that the methods are not at odds, cell-weighting, raking and PSA might very well be used to adjust the same dataset.

The fourth contribution is about sampling method, which is the most controversial part of polling. The aggregate results found in Sweden by Sohlberg et al. (2017) and on an individual level in the US (Pasek, 2016; Yeager et al., 2011) are confirmed: on average, nonprobability samples produce less accurate and less precise measurements. This is not very surprising, taking into account both theories behind sampling as well as the current empirical evidence of the field. However, the thesis highlights a few important caveats that deserves to be lifted forward, since they indicate that not all is said in the debate about the importance of probability samples: 1) often, the sample variation is too low to claim generalizability, and this is the case also in my study, 2) several of the samples in my study were not pre-stratified, and 3) there is substantial variation in accuracy both before and after applying weights. The variation indicated that some nonprobability samples can be at least as good as some probability samples. This means that the quality categorization of nonprobability samples is an important distinction to be made, a distinction that may explain many of the diverging results found in studies where non-probability samples are conflated into one large category, thereby being jointly refuted, rather than individually evaluated.

There are a few notable drawbacks of the design employed in this thesis. The fact that the main benchmark is a survey itself, and therefore a subject to many of the same types of biases is a valid objection. Yet the overall results are unlikely to change, both since the benchmark is known to be reliable, have a high response rate and the results were largely reproduced using election results as the benchmark. The fact the probability sample DS' produced very poor results in some cases could be caused by the fact that they were too similar to the benchmark to begin with.

A second point is that the study features a convenience sample of samples. A consequence is that not all surveys from different pollsters overlap time-wise and there is a higher density of surveys closer to the election of 2014. It would have been desirable with more variation in terms of probability samples, of varying levels of response rates as well as the data collection mode. Mode is relevant due to mode effects such as social desirability bias, which is usually a greater problem in telephone surveys. Another concern regarding survey mode is that both the benchmark and the DS surveys, the set of surveys which produced the most accurate measurements, are telephone survey and therefore both affected by the aforementioned social desirability bias—a bias that would likely affect the support for the Sweden Democrats negatively (see e.g. Knigge, 1998). On average, among the 8 other party categories, the effect is however likely to be small. There are also still relatively few commercial probability-based web panels and none that collect publicly available vote intention data and second, there are no true nonprobability telephone surveys. Although a case could be made that random digit dialing (RDD) may be cutting it close depending on the quality of the telephone number list and whether noncontact results in calls to another number or whether that same number is called over and over. This is sample of samples is likely to bias some of the results.

A third point is that the results indicate that each of the more efficient covariates have a sufficiently high correlation with both *P* and *Y*. However, this is not necessarily true for any of the inefficient covariates, since it is not known whether it is the *Z*–*P* or *Z*–*Y* relation (or both) that causes the problem. The fact that only Swedish data is used could also be a problem for the generalizability, since the relation between survey participation and vote intention might potentially be a Swedish idiosyncrasy. This is likely to be true to some extent, but it is also true that some aspects of survey participation, such as the low participation among the young, and electoral participation and party choice seems to be more universal, at least in the Western democracies.

Final remarks and future research

The implication of the thesis is that while surveys with nonprobability samples will lead to lower consistency and lower accuracy, there is still a place for good nonprobability samples in describing societal trends and relationships or even point estimates when somewhat lower accuracy is acceptable.

This is an important conclusion as high-quality probability samples become increasingly difficult to come across. The thesis has also shown that it is possible to detect crucial variation in the quality of different non-probability sampling. Or put differently; scholars can (and probably should) put effort into evaluating what non-probability sampling that can be trusted and under what circumstances. This being said, although vote intention is an important measure in political science, the results should be seen in a wider perspective of overall public policy processes where surveys often are the basis for decision making. When the variable-of-interest will be much less explored than vote intention (which will be the case for most other variables), even more care will be needed when adjusting these variables. If weights are not correlated to both the outcome variable and to the survey response, then there is a risk that the weights introduce more bias than they remove.

For survey research, these results suggest that there is a need to find more unorthodox adjustment covariates, such as political interest, to get accurate measurements. Focus should be on finding covariates that are stable. Rivers (2016) gives two examples from the US: third party identification and respondents who cannot place themselves on an ideological scale.

Acknowledgements and disclaimer

I would like to express the sincerest gratitude to Karin Nelson at Inizio and Peter Santesson at Demoskop for making the raw datasets available for analysis. The transparency of United Minds was also commendable since they shared their data online for all to see.

The following is a disclaimer from Inizio regarding the data used:

"Schibsted/Inizios opinionspanel samlar dagligen in data om röstningsbeteende. Det är den här datan som är utlånad till Laboratory of Opinion Research (LORE). Notera att det alltså inte är verktygen som producerar Aftonbladet/Inizios väljarbarometer som lånats ut. De metoder och procedurer som Aftonbladet/Inizio använder för att skatta svensk väljaropinion är alltså inte desamma som LORE använder för att utvärderna Schibsted/Inizios opinionspanel. De mått och steg som LORE använder för att utvärdera Schibsted/Inizios opionionspanel resulterar inte nödvändigtvis i de siffror som en gång publicerats under rubriken 'Aftonbladet/Inizios väljarbarometer'."

6 References

Achen, C., & Blais, A. (2010). Intention to vote, reported vote, and validated vote.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior *Action control* (pp. 11-39): Springer.

Ansolabehere, S., & Rivers, D. (2013). Cooperative survey research. *Annual Review of Political Science, 16*(1), 307-329. doi:10.1146/annurev-polisci-022811-160625

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M. P., Couper, M. P., Dever, J. A., . . . Tourangeau, R. (2013). *Report of the AAPOR task force on nonprobability sampling*

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations, 61*(8), 1139-1160. doi:10.1177/0018726708094863

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2013). Practical considerations in raking survey data. *Survey Practice, 2*(5).

Bergman, J., & Holmquist, B. (2014). Poll of polls: A compositional loess model. *Scandinavian Journal of Statistics, 41*(2), 301-310.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. (2003). The advent of Internet surveys for political research: A comparison of telephone and Internet samples. *Political Analysis, 11*(1), 1-22.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics, 4*(3), 251-260.

Bethlehem, J. G., & Keller, W. J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics, 3*(2), 141-153.

Bosnjak, M., Das, M., & Lynn, P. (2016). Methods for probability-based online and mixed-mode panels: Selected recent trends and future perspectives. *Social Science Computer Review, 34*(1), 3-7.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

Cohn, N. (2016). A favorable poll for Donald Trump seems to have a problem. *New York Times*. Retrieved from http://www.nytimes.com/2016/08/09/upshot/a-favorable-poll-for-donald-trump-has-a-major-problem.html

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly, 69*(1), 87-98.

Dalton, R. J. (2008). *Citizen politics - Public opinion and political parties in advanced industrial democracies* (Fifth ed.). Washington, D.C.: CQ Press.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. *Survey nonresponse*, 41-54.

Dever, J. A., Rafferty, A., & Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods, 2*(2), 47-62.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(418), 376-382.

DiSogra, C., & Callegaro, M. (2015). Metrics and design tool for building and evaluating probability-based online panels. *Social Science Computer Review*, 0894439315573925.

Durand, C., Blais, A., & Larochelle, M. (2004). Review: The polls in the 2002 French presidential election: An autopsy. *Public Opinion Quarterly, 68*(4), 602-622.

Durand, C., Deslauriers, M., & Valois, I. (2015). Should recall of previous votes be used to adjust estimates of voting intention? *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach.*

ESOMAR. (2014). *Global market research 2014*. Retrieved from ESOMAR, Amsterdam: http://www.mrsnz.org.nz/webfiles/MarketResearchSocietyNZ/files/ESOMAR_GMR2014_FullReport.pdf

European Commission. (2016). Digital Economy and Society Index (DESI) 2016 - Sweden. Retrieved from https://ec.europa.eu/digital-single-market/scoreboard/sweden

European Social Survey. (2014). *Weighting European Social Survey data* https://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research, 15*(2), 195-219. doi:10.1108/10662240510590360

Granberg, D., & Holmberg, S. (1990). The intention-behavior relationship among US and Swedish voters. *Social Psychology Quarterly*, 44-54.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646-675.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861-871.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). The decision to participate in a survey. *Public Opinion Quarterly, 56*(4), 475-495.

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly, 70*(5), 720-736.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561): John Wiley & Sons.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public Opinion Quarterly, 72*(2), 167-189.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly*, 299-308.

Hansen, M. H., & Hauser, P. M. (1945). Area sampling—some principles of sample design. *Public Opinion Quarterly, 9*(2), 183-193.

Hill, S. J., Lo, J., Vavreck, L., & Zaller, J. (2007). The opt-in Internet panel: Survey mode, sampling methodology and the implications for political research. *Unpublished manuscript at the University of California, Los Angeles, California*.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2008). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (Vol. 538, pp. 499-528): John Wiley & Sons.

Holmberg, S. (1986). Politiska opinionsmätningar i Sverige *Källa 23: Makten över opinionen*: Forskningsrådsnämnden (FRN).

Holmberg, S., & Petersson, O. (1980). *Inom felmarginalen*. Stockholm: Liber.

Jowell, R., Hedges, B., Lynn, P., Farrant, G., & Heath, A. (1993). Review: The 1992 British election: The failure of the polls. *Public Opinion Quarterly, 57*(2), 238-263.

Kalton, G. (1983). *Compensating for missing survey data*. Retrieved from Ann Arbor, Michigan: http://www.psc.isr.umich.edu/dis/infoserv/isrpub/pdf/CompensatingforMissingSurveyData_OCR.PDF

Kalton, G., & Flores-Cervantes, I. (1998). *Weighting methods*. Paper presented at the International Conference of Association for Survey Computing, Chilworth Manor, Southampton, UK.

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

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys - Vendor choice matters; widespread errors found for estimates based on blacks and Hispanics* http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/

Kish, L. (1965). *Survey sampling*: John Wiley & Sons.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*(2), 183-200.

Knigge, P. (1998). The ecological correlates of right–wing extremism in Western Europe. *European Journal of Political Research, 34*(2), 249-279.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Christian, L. (2012). Assessing the representativeness of public opinion surveys. *Pew Research Center, Washington, DC*.

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, 169-195.

Langer, G. (2013). Comment. *Journal of Survey Statistics and Methodology, 1*(2), 130-136. doi:10.1093/jssam/smt011

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics, 22*(2), 329.

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research, 37*(3), 319-343. doi:10.1177/0049124108329643

Lipset, S. M., & Rokkan, S. (1967). Cleavage structures, party systems, and voter alignments: an introduction. In S. M. Lipset & S. Rokkan (Eds.), *Party systems and voter alignments: Cross-national perspectives* (pp. 1-64). New York: Free Press.

Literary Digest. (1936). What went wrong with the polls? *Literary Digest, November 14,* 7-12.

Little, R. J., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology, 31*(2), 161-168.

Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods, 2*(2), 93-105.

Lusinchi, D. (2015). Straw poll journalism and quantitative data: The case of the Literary Digest. *Journalism Studies, 16*(3), 417-432.

Lönegård, C. (2016). Mäthatet [Online (tr/p)olling]. *Fokus.*

Markstedt, E. (2014). *Representativitet och viktning - Riks-SOM som spegel av det svenska samhället 1986-2013.* Retrieved from University of Gothenburg: som.gu.se

Markstedt, E. (2016). *Annual report – LORE Citizen Panel 2015.* Retrieved from Gothenburg:

Martin, E. A., Traugott, M. W., & Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *Public Opinion Quarterly, 69*(3), 342-369.

Meier, N. C., & Burke, C. J. (1947). Laboratory tests of sampling techniques. *Public Opinion Quarterly, 11*(4), 586-593.

Mitofsky, W. J. (1998). Review: Was 1996 a worse year for polls than 1948? *The Public Opinion Quarterly, 62*(2), 230-249.

Mosteller, F., Hyman, H., McCarthy, P. J., Marks, E. S., & Truman, D. B. (1949). *The pre-election polls of 1948: The report to the Committee on Analysis of Pre-election Polls and Forecasts*

Novus. (2016). *Samtliga svenska väljarbarometrar.* Retrieved from: http://novus.se/valjaropinionen/samtliga-svenska-valjarbarometrar/

Oscarsson, H., & Holmberg, S. (2013). *Nya svenska väljare [New Swedish voters].* Stockholm: Norstedts Juridik.

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*(2), 269-291.

Peiser, W. (2000). Cohort replacement and the downward trend in newspaper readership. *Newspaper Research Journal, 21*(2), 11.

Petersson, O., & Holmberg, S. (1998). *Opinionsmätningarna och demokratin*: SNS.

Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research, 212,* 63-73. doi:10.1002/ir.101

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings.*

Rivers, D. (2013). Comment. *Journal of Survey Statistics and Methodology, 1*(2), 111-117. doi:10.1093/jssam/smt009

Rivers, D. (2016, May 13, 2016). Pew Research: YouGov consistently outperforms competitors on accuracy. Retrieved from https://today.yougov.com/news/2016/05/13/pew-research-yougov/

Rivers, D., & Bailey, D. (2009). Inference from matched samples in the 2008 US national election *Proceedings of the Joint Statistical Meetings* (pp. 627-639).

Rivers, D., & Wells, A. (2015). *Polling error in the 2015 UK general election: An analysis of Yougov's pre and post-election polls* https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/x4ae830iac/YouGov%20%E2%80%93%20GE2015%20Post%20Mortem.pdf

Rosema, M., Anderson, J., & Walgrave, S. (2014). The design, purpose, and effects of voting advice applications. *Electoral studies, 36*, 240-243.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249-264.

Schonlau, M., van Soest, A., & Kapteyn, A. (2007). Are 'webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring?

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research, 37*(3), 291-318. doi:10.1177/0049124108327128

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., . . . Berry, S. H. (2004). A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review, 22*(1), 128-138. doi:10.1177/0894439303256551

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.

Smith, J. R., Terry, D. J., Manstead, A. S., Louis, W. R., Kotterman, D., & Wolfs, J. (2007). Interaction effects in the theory of planned behavior: the interplay of self-identity and past behavior. *Journal of Applied Social Psychology, 37*(11), 2726-2750.

Smith, T. W. (1995). Trends in non-response rates. *International Journal of Public Opinion Research, 7*(2), 157-171.

Sohlberg, J., Gilljam, M., & Martinsson, J. (2017). Determinants of polling accuracy: the effect of opt-in Internet surveys. *Journal of Elections, Public Opinion and Parties*, 1-15.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly, 52*(1), 125-133.

Statistics Sweden. (2014). *The party preference survey in November 2014* http://www.scb.se/Statistik/_Publikationer/ME0201_2014M11_BR_ME60BR1402.pdf

Steinmetz, S., Tijdens, K., & de Pedraza, P. (2009). *Comparing different weighting procedures for volunteer web surveys* (Working Paper 09-76)http://www.uva-aias.net/uploaded_files/publications/WP76-Steinmetz,Tijdens&Pedraza.pdf

Stephan, F. F., & McCarthy, P. J. (1958). *Sampling opinions: An analysis of survey procedure*. Oxford, UK: John Wiley.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics, 25*(1), 1.

Sturgis, P. (2016). *Quality of Internet Surveys: what can we learn from election polling?* Paper presented at the Web Panel Surveys, Methods and Experiences Statistics Sweden, Stockholm Sweden. http://www.scb.se/Grupp/Produkter_Tjanster/_Dokument/Web-Panel-Surveys-Patrick-Sturgis.pdf

Sturgis, P., Williams, J., Brunton-Smith, I., & Moore, J. (2016). Fieldwork effort, response rate, and the distribution of survey outcomes: a multi-level meta-analysis. *Public Opinion Quarterly*.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W., & Terhanian, G. (2001). The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 US elections. *International Journal of Market Research, 43*(2), 127-136.

Thomassen, J. (2005). Introduction. In J. Thomassen (Ed.), *The European voter. A comparative study of modern democracies* (pp. 1-21): Oxford University Press.

Tourangeau, R., Conrad, F. G., & Couper, M. (2013). *The science of web surveys*: Oxford University Press.

Traugott, M. W. (2005). The accuracy of the national preelection polls in the 2004 presidential election. *Public Opinion Quarterly, 69*(5), 642-654. doi:10.1093/poq/nfi061

Twyman, J. (2008). Getting it right: YouGov and online survey research in Britain. *Journal of Elections, Public Opinion and Parties, 18*(4), 343-354.

van Elsas, E. J., Lubbe, R., van der Meer, T. W. G., & van der Brug, W. (2013). Vote recall: A panel study on the mechanisms that explain vote recall inconsistency. *International Journal of Public Opinion Research*, edt031.

Vavreck, L., & Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties, 18*(4), 355-366.

Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. *Election polls, the news media, and democracy*, 224-260.

Yeager, D. S., Krosnick, J., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, nfr020.

7 Appendix

Appendix A. General dataset information

Appendix table 1. Information on the data sources used in the study

Survey	Response/ participation rate	Invited sample	Participating sample	Data collection period
Citizen Panel wave 13 – opt-in*	58	3,178	1,836	Nov 11–Dec 21, 2014
Citizen Panel wave 13 – VAA	52	6,816	3,533	Nov 11–Dec 21, 2014
Citizen Panel wave 13 – probability	53 (CUMR: 5.4)	6,068	3,177	Nov 11–Dec 21, 2014
Citizen Panel wave 15 – opt-in	70	6,421	4,463	May 11–June 1, 2015
Citizen Panel wave 15 – VAA	59	12,802	7,579	May 11–June 1, 2015
Citizen Panel wave 15 – probability	63 (CUMR: 5.9)	2,516	1,575	May 11–June 1, 2015
Demoskop	14***		1,287	Nov 1, 2013
Demoskop	14		1,267	May 1, 2014
Demoskop	14		1,274	Sept 1, 2014
Demoskop	14		1,265	Nov 1, 2014
Demoskop	14		1,276	May 1, 2015
Inizio	61	n.a.	1,839	Oct 29–Nov 20, 2014
Inizio	64	n.a.	5,996	Apr 4–May 27, 2015
United Minds	n.a. **	n.a.	1,109	Oct 31–Nov 25, 2010
United Minds	n.a.	n.a.	1,155	May 1–26, 2011
United Minds	n.a.	n.a.	1,171	Nov 1–27, 2011
United Minds	n.a.	n.a.	1,100	May 2–27, 2012
United Minds	n.a.	n.a.	1,003	Nov 1–25, 2012
United Minds	n.a.	n.a.	1,049	May 2–27, 2013
United Minds	n.a.	n.a.	1,084	Nov 3–27, 2013
United Minds	n.a.	n.a.	954	May 2–25, 2014
SOM Institute National Survey 2010	56		5,007	Sep 18, 2010–Feb 27, 2010
SOM Institute National Survey 2014	51		6,876	Sep 18, 2014–Feb 27, 2014
Party Preference Survey	68		6,192	Oct 31–Nov 25, 2010
Party Preference Survey	67		6,147	May 1–26, 2011
Party Preference Survey	65		5,907	Nov 1–27, 2011
Party Preference Survey	61		5,473	May 2–27, 2012
Party Preference Survey	61		5,479	Nov 1–25, 2012
Party Preference Survey	56		5,098	May 2–27, 2013
Party Preference Survey	58		5,267	Nov 3–27, 2013
Party Preference Survey	52		4,757	May 2–25, 2014
Party Preference Survey	56		5,072	Oct 29–Nov 25, 2014
Party Preference Survey	50		6,067	Apr 27–May 27, 2015

Comment: *In the Citizen Panel studies actual response rates (for example the cumulative response rates suggested by DiSogra & Callegaro, 2015) are usually not reported, but instead a so called participation rate is used, which corresponds to AAPOR RR5 had the panel included the entire target population. **No response rate can be calculated since no sampling in a traditional sense occurs. ***14 is the usual RR during this period. Note that the response rate in general is not comparable between the data providers since the original sample sizes are differently determined. See the technical reports for Citizen Panel 13 and 15.

Appendix B. Weighting examples.

Weighting example 1: Cell weighting

The cell weighting procedure is simple: covariates in the target population are cross-tabulated and divided by the same proportion in the realized sample. For example, if women with low education were 34.4 percent of the target population and 16.7 percent of the sample population, then the weight would be 2.06 (34.4 / 16.7 ≈ 2.06, see the table below). Note that the same procedure can be used using the acutal population figures if known.

Appendix table 2. Cell-weighting example

Sample to be weighted (cell proportions):

	Low education	High education	Total
Woman	16.7	28.0	44.7
Man	24.2	31.1	55.3
Total	40.9	59.1	100

Target population (cell proportions):

	Low education	High education	Total
Woman	34.4	17.8	52.3
Man	35.9	11.8	47.7
Total	70.4	29.6	100

Final cell-weights:

	Low education	High education
Woman	2.06	0.64
Man	1.49	0.38

Weighting example 2: Raking

Raking is similar to cell weighting in the sense that it relies directly on univariate distributions, but only uses the margin totals of an aggregated frequency distribution in a population. In Appendix table 3. Raking examplebelow, the population margin totals are known while the joint distribution is unknown. In the first iteration the procedure creates a weight that balances the sample to conform to the distribution of one of the margin totals (e.g. the distribution of different levels of education). In the second iteration the same thing is done for the second now adjusted margin totals (gender in the example), which in turn makes the first distribution to diverge from the population again. The procedure is then repeated until all margin totals are correct simultaneously. One of the greatest benefits of raking is that you can easily combine information from different sources without knowing the multivariate cell proportions.

Appendix table 3. Raking example

Sample:

	Low edu	High edu	Total
Woman	16.7	28.0	44.7
Man	24.2	31.1	55.3
Total	40.9	59.1	100

Population:

	Low edu	High edu	Total
Woman	unknown	unknown	52.3
Man	unknown	unknown	47.7
Total	70.4	29.6	100

First iteration:

Second iteration:

Final raked weights:

	Low edu	High edu
Woman	1.72	0.50
Man	1.72	0.50

	Low edu	High edu
Woman	2.01	0.59
Man	1.48	0.43

	Low edu	High edu
Woman	2.06	0.64
Man	1.49	0.38

Weighting example 3: Propensity score adjustment (PSA)

PSA is originally designed to create a control group for observational studies where treatment group assignment is not random, which is the case in most observational settings. The basic idea is to gather information about the respondents in the treatment group that are related to both the outcome and the likelihood of being in the treatment group then find “untreated” individuals who are as similar to treated respondents as possible except the fact that they are in the treatment group.

The same basic idea is used when reducing biases in surveys. First, the respondents from the biased survey and the benchmark reference survey are merged to a single dataset. Second, the likelihood of being “treated” according to a set of covariates is predicted using a logit regression or a similar method. “Treatment” here is answering the survey to be weighted, i.e. the combined likelihood that the individual is covered in the sampling list, selected from it, is availble for interviewing and answers. Third, all individuals are sorted by likelihood and divided into equal parts, for example quintiles (originally proposed by Cochran, 1968). These parts, usually called “bins”, are the basis for each individual weight:

$$psa\_weight = \frac{n_b^R/n^R}{n_b^S/n^S}$$

Where n_b^R is the number of reference survey respondents in a specific bin and n^R is the total number of respondents in the reference survey. n_b^S and n^S are the equivalents in the biased survey. The weight then serves to adjust the biased survey to have the same distribution of likelihoods.

Appendix C. Covariates

Appendix table 4. Covariates, question wording, response alternatives and coding.

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
Gender (G)					Är du kvinna eller man?	[Register data]	Är du man eller kvinna?	Är du...	Are you woman or a man?
					Kvinna	Kvinna	Kvinna	Kvinna	Woman
					Man	Man	Man	Man	Man
Age (A)					Vilket år är du född?	Hur gammal är du?	Vilket år är du född?	Vilket år är du född?	What year were you born?
					1920 eller tidigare/1999 eller senare	18-89	1918-1999	1910-1997	
Age groups	1	1	1		18-29	18-29	18-29	18-29	
	1	2	2		30-44	30-44	30-44	30-44	
	2	3	3		45-59	45-59	45-59	45-59	
	2	3	4		60-70	60-70	60-70	60-70	
Education (E)					Vilken skolutbildning har du? Markera det svar som du anser bäst stämmer in på dig.	Vilken skolutbildning har du? Har du...?	Vilken skolutbildning har du? Om du ännu inte avslutat din utbildning, markera den du genomgår för närvarande.	Vilken är din högsta avslutade utbildning?	What is your education?
	1	1	1		Ej fullgjort grundskola		Ej fullgjort grundskola (eller motsvarande obligatorisk skola)		Not completed elementary school
	1	1	1		Grundskola	Grundskolenivå	Grundskola (eller motsvarande obligatorisk skola)	Folkskola, grundskola	Elementary school

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
	1	2	2		Gymnasium eller motsvarande, kortare än 3 år	Gymnasienivå	Studier vid gymnasium, folkhögskola (eller motsvarande)	Gymnasieutbildning kortare än tre år	High school, less than three years
	1	2	2		Gymnasium eller motsvarande, 3 år eller längre		Examen från gymnasium, folkhögskola (eller motsvarande)	Gymnasieutbildning tre år eller längre	High school, three years or more
	1	2	3		Eftergymnasial utbildning, ej högskola, kortare än 3 år				Studies after high school (not college/
	1	2	3		Eftergymnasial utbildning, ej högskola, 3 år eller längre		Eftergymnasial utbildning, ej högskola/universitet	Eftergymnasial utbildning upp till tre år	Studies after high school (not college/
	1	2	3		Högskola/universitet, kortare än 3 år	Universitetsnivå	Studier vid högskola/universitet		University/college less than three year
	2	3	4		Högskola/universitet, 3 år eller längre		Examen från från högskola/universitet	Eftergymnasial utbildning längre än tre år	University/college three years or more
	2	3	4		Examen från forskarutbildning		Examen från/vid studier vid forskarutbildning		PhD
Region (R)					Vilken kommun bor du i?	[Register data]	Vilket län bor du i -> Vilken kommun bor du i?	[H-region]*	What municipality do you live in?
					Ale-Övertorneå	Ale-Övertorneå	Ale-Övertorneå	1-9	
Region groups (number of inhabitants)	1	1	1		1-29 999 inhabitants	1-29 999 inhabitants	1-29 999 inhabitants	5-6	
	1	2	2		30 000-69 000 inhabitants	30 000-69 000 inhabitants	30 000-69 000 inhabitants	4	
	1	2	3		70 000-199 999 inhabitants	70 000-199 999 inhabitants	70 000-199 999 inhabitants	3	
	2	3	4		200 000+ inhabitants	200 000+ inhabitants	200 000+ inhabitants	1, 8-9	
Marital status					Är du...:				

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
(M)									
	1	1	1		Ensamstående - och har aldrig varit gift				Single - and have never been married
	2	2	2		I förhållande - och har aldrig varit gift				In a relationship - and have never been
	2	2	2		Sambo - och har aldrig varit gift				Cohab - and have never been married
	2	3	3		Gift				Married
	1	1	1		Skild - och ensamstående				Divorced - and single
	2	2	2		Skild - och i nytt förhållande				Divorced - and in a new relationship
	1	1	4		Änka/änkling - och ensamstående				Widow/widower - and single
	1	1	4		Änka/änkling - och i nytt förhållande				Widow/widower - and in a new relationsh
Labor market situation (L)					Hur ser din arbetsmarknadssituation ut?				Which of these groups do you currently belong to?
	1	1	1		Egen företagare				Entrepreneur
	1	1	1		Anställd heltid				Gainfully employed, full time
	1	1	1		Anställd deltid				Gainfully employed, part time
	2	2	2		Arbetsmarknadspolitisk åtgärd/-utbildning				Participating in labour market policy measures
	2	2	2		Arbetslös				Unemployed
	2	2	3		Student				Student
	2	3	4		Pensionär				Pensioner
	2	2	2		Hemmavarande				Homeworker

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
	2	2	2		Annat, ej yrkesaktiv:				Other
Political interest (I)					Hur intresserad är du i allmänhet av politik?				Generally speaking, how interested are you in politics?
	1	1	1		Inte alls intresserad				Not at all interested
	1	1	2		Inte särskilt intresserad				Not particularly interested
	1	2	3		Ganska intresserad				Rather interested
	2	3	4		Mycket intresserad				Very interested
Refugee proposal (P)					Nedan finns ett antal förslag som har förekommit i den politiska debatten. Vilken är din åsikt om vart och ett av dem?				Below is a number of proposals that have appeared in the political debate. What is your opinion on each of them?
	1	1	1		Mycket dåligt förslag				Very bad proposal
	1	1	1		Ganska dåligt förslag				Rather bad proposal
	1	2	2		Varken bra eller dåligt förslag				Neither good nor bad proposal
	2	3	3		Ganska bra förslag				Rather good proposal
	2	3	4		Mycket bra förslag				Very good proposal
Party preference					Vilket parti skulle du rösta på om det vore riksdagsval idag?	Vilket parti skulle du rösta på om det var riksdagsval idag?	Om det vore val till riksdagen idag, vilket parti skulle du rösta på då?	Hur skulle du rösta om det vore val till riksdagen i dag?	What party would you vote for if it was election day today?
				1	Vänsterpartiet	Vänsterpartiet	Vänsterpartiet	Vänsterpartiet	Left Party
				2	Socialdemokraterna	Socialdemokraterna	Socialdemokraterna	Socialdemokraterna	Social Democrats
				3	Centerpartiet	Centern	Centerpartiet	Centerpartiet	Centre Party

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
				4	Folkpartiet	Folkpartiet	Folkpartiet	Folkpartiet	People's Liberal Party
				5	Moderaterna	Moderaterna	Moderaterna	Moderaterna	Moderates
				6	Kristdemokraterna	Kristdemokraterna	Kristdemokraterna	Kristdemokraterna	Christian Democrats
				7	Miljöpartiet	Miljöpartiet	Miljöpartiet	Miljöpartiet	Green Party
				8	Sverigedemokraterna	Sverigedemokraterna	Sverigedemokraterna	Sverigedemokraterna	Sweden Democrats
				9	Feministiskt initiativ	Feministiskt Initiativ	Feministiskt Initiativ		Feminist Initiative
				9		Piratpartiet			
				9	Annat parti, nämligen:	Annat parti	Annat, vilket?	Annat	Other party
					Skulle rösta blankt	*Skulle rösta blankt*		*Tänker rösta blankt*	Blank ballot
					Skulle inte rösta	*Skulle inte rösta*		*Skulle inte rösta*	*Would not vote*
					Vet ej	*Vet ej vilket parti*	*Tveksam, vet ej*	*Tveksam, vet ej*	*Don't know*
						Vägrar uppge/Ej svar		*Vill ej uppge*	*Don't want to answer*
Vote recall (V)					Vilket parti röstade du på i riksdagsvalet 2014?	Vilket parti röstade du på i senaste riksdagsvalet?	[Röstning i riksdagsvalet 2014]	Om du ser tillbaka till det senaste riksdagsvalet 2010 - röstade du i det valet och i så fall på vilket parti?	What party did you vote in the 2010/2014 elections?
				1	Vänsterpartiet	Vänsterpartiet	Vänsterpartiet	Vänsterpartiet	Left Party
				2	Socialdemokraterna	Socialdemokraterna	Socialdemokraterna	Socialdemokraterna	Social Democrats
				3	Centerpartiet	Centern	Centerpartiet	Centerpartiet	Centre Party
				4	Folkpartiet	Folkpartiet	Folkpartiet	Folkpartiet	People's Liberal Party
				5	Moderaterna	Moderaterna	Moderaterna	Moderaterna	Moderates
				6	Kristdemokraterna	Kristdemokraterna	Kristdemokraterna	Kristdemokraterna	Christian Democrats
				7	Miljöpartiet	Miljöpartiet	Miljöpartiet	Miljöpartiet	Green Party
				8	Sverigedemokraterna	Sverigedemokraterna	Sverigedemokraterna	Sverigedemokraterna	Sweden Democrats
				9	Feministiskt initiativ	Feministiskt initiativ	Feministiskt initiativ		Feminist Initiative
				9	Annat parti, nämligen:	Annat parti	Annat parti	Annat	Other party

Covariate	**Coding**				**Survey**				
	2 cat	*3 cat*	*4 cat*	*9 cat*	*Citizen Panel (CP)*	*Demoskop (DS)*	*Inizio (IN)*	*United Minds (UM)*	*English translation*
				10	Röstade blankt	Röstade med valsedel utan partibeteckning	Röstade blankt	Röstade blankt	Blank ballot
					Hade inte rösträtt	*Hade inte rösträtt då*	*Var inte röstberättigad*	*Var för ung då*	*Not eligible to vote*
				10	Röstade inte	Röstade inte	Röstade inte	Röstade inte	*Did not vote*
					Minns ej/vill inte svara	*Minns ej*	*Minns inte*	*Minns ej*	*Don't remember*
						Ej svar		*Vill ej uppge*	*Don't want to answer*
								Utländsk medborgare då	*Foreign citizen at the time*

Appendix D. Tables

Appendix table 5. Mean absolute bias before and after weighting, mean absolute bias reduction and mean absolute proportional bias reduction by sample and weighting technique

		CPCON	*CPVAA*	*CPPROB*	*IN*	*UM*	*DS*
Weighting technique							
	Unweighted bias	5.46	4.36	3.56	4.00	2.32	1.65
Cell-weight	Bias after weighting	4.07	3.49	2.71	3.14	1.98	1.70
		(1.56)	(1.24)	(0.86)	(0.76)	(0.59)	(0.56)
	Bias reduction	−1.39	−0.88	−0.84	−0.86	−0.34	+0.05
		(1.52)	(1.05)	(0.85)	(0.69)	(0.49)	(0.42)
	Proportional bias reduction	−26	−21	−24	−22	−15	+5
		(28)	(25)	(24)	(18)	(21)	(25)
Raked weights	Bias after weighting	4.05	3.46	2.65	3.08	1.93	1.59
		(1.59)	(1.24)	(0.86)	(0.79)	(0.65)	(0.52)
	Bias reduction	−1.41	−0.91	−0.91	−0.91	−0.39	−0.06
		(1.54)	(1.06)	(0.84)	(0.71)	(0.55)	(0.41)
	Proportional bias reduction	−26	−21	−26	−23	−17	−1
		(29)	(25)	(24)	(18)	(24)	(24)
PSA weights *(other set of weights)*	Bias after weighting	3.74	3.30	2.63	3.61	1.97	2.26
		(1.47)	(1.23)	(0.73)	(0.44)	(0.57)	(0.65)
	Bias reduction	−1.72	−1.06	−0.93	−0.39	−0.35	+0.61
		(1.42)	(1.03)	(0.71)	(0.56)	(0.49)	(0.53)
	Proportional bias reduction	−32	−25	−26	−9	−15	+42
		(26)	(25)	(20)	(14)	(21)	(35)
N cell/raked weight		1,344	1,344	1,344	456	5,376	2,620
N PSA		228	228	228	228	912	570

Comment: Standard deviations are shown in parentheses. Only the results for comparable weights are reported for cell-weights and raked weights, i.e.: Bias is calculated as the average absolute bias between the percentage points of all parties in the benchmark (PSU) and the weighted survey. Bias reduction is calculated as the difference between the unweighted bias and the weighted bias. Proportional bias reduction is calculated as relative size of the bias reduction proportion in relation to the unweighted bias.

Appendix table 6. OLS regressions with mean absolute bias reduction in terms of percentage points as dependent variable, cell-weight and raked weights.

	Model 1: Simple – all sample sources		Model 2: Full – all sample sources		Model 3: Full – Citizen Panel only		Model 4: DS Sept 2014 only – Election result benchmark	
	b	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Demographic covariates[a]								
Age (A)	+0.20***	(0.04)	+0.20***	(0.02)	+0.21**	(0.04)	+0.25***	(0.01)
Education (E)	−0.20***	(0.05)	−0.07	(0.04)	−0.18	(0.11)	−0.27***	(0.01)
Region (R)	−0.08*	(0.03)	−0.01	(0.03)	−0.07	(0.04)	−0.01	(0.01)
Labor market situation (L)					−0.04	(0.04)	−0.02*	(0.01)
Marital status (M)					−0.03	(0.03)		
Psychographic/vote covariates[a]								
Vote recall (V)			−1.60***	(0.23)	−1.82*	(0.48)	−0.66***	(0.01)
Months since last election (2/48)			−0.01*	(0.00)	+0.03	(0.02)		
Vote recall × months since last election			0.03**	(0.01)	0.00	(0.09)		
Political interest (I)					−0.49***	(0.06)		
Refugee policy proposal (P)					−0.18**	(0.04)		
Weight paradata[b]								
2 covariates			−0.06*	(0.02)	−0.07**	(0.01)	−0.01	(0.02)
3 covariates			−0.09*	(0.04)	−0.09*	(0.03)	+0.01	(0.03)
4 covariates			−0.12	(0.06)	−0.05	(0.06)	+0.05	(0.03)
Number of weight cells (standardized 0–1)			−0.08	(0.08)	−0.33**	(0.06)	+0.05	(0.03)
Weighting technique[c]								
Raked weight			−0.06**	(0.02)	−0.01	(0.02)	−0.09***	(0.01)
Sample[d]								
CPVAA			+0.68***	(0.12)	+0.50***	(0.01)		
CPPROB			+0.68**	(0.19)	+0.45***	(0.02)		
IN			+0.76***	(0.15)				
UM			+1.02***	(0.11)				
DM			+1.35***	(0.11)				
Constant	−0.05*	(0.02)	−0.84***	(0.13)	−0.60**	(0.14)	0.02	(0.02)
R2	0.173		0.785		0.850		0.917	
N	5014		9026		78156		1048	
RMSE	0.28		0.39		0.35		0.11	

Comment: Standard errors are clustered by sample source (Model 4 is not clustered since it uses only one sample). * p<.05; ** p<.01; *** p<.001. Reference categories: [a]Gender (G). [b]1 covariate. [c]Cell-weight. [d]CPCON. The DV is the effect of weighting on the average absolute deviation from the benchmark. Each observation is a unique weight using different combinations of variables. Gender through vote recall are dummy variables (0/1). Number of cells, which indicates the number of unique cell combinations that can be made with the variables used, is standardized from 0 to 1 (it ranges in practice from 2 cells to 576).

Appendix table 7. Comparison between using an unweighted or weighted SOM benchmark (OLS with proportional bias reduction as dependent variable, cell-weights)

	Cell-weights, Citizen Panel 13 Simplified model 5			
	Unweighted SOM		Weighted SOM	
	b	*SE*	*b*	*SE*
Ref cat: Gender				
Age	+5.3	(2.0)	+12.3	(3.1)
Education	−8.2	(4.1)	−8-.5	(3.0)
Region	−3.2	(1.4)	−0.3	(0.5)
Labor market situation	−2.2	(0.6)	+9.7**	(0.8)
Vote recall	−48.2**	(4.6)	−52.5**	(1.7)
Marital status	−1.7	(0.6)	−2.1	(1.3)
Political interest	−9.7*	(1.7)	−7.8	(3.1)
Refugee policy proposal	−5.8	(1.5)	−4.6	(1.9)
Number of cells	−5.5	(1.6)	+17.4*	(3.5)
Ref cat: Cell-weight				
Number of variables 2	−1.5*	(0.2)	−4.5**	(0.4)
Number of variables 3	−1.4	(0.3)	−6.7**	(0.4)
Number of variables 4	+0.2	(0.7)	−9.2*	(1.0)
Ref cat: CPCON				
CPVAA	+2.9***	(0.0)	+4.7***	(0.0)
CPPROB	−0.4***	(0.0)	−6.4***	(0.0)
Constant	−4.3*	(0.9)	−6.9	(2.3)
R2	0.901		0.626	
N	19,539		19,539	
RMSE	6.7		17.7	

Comment: This analysis uses a dependent variable with 9 categories instead of 10 as in the main analysis (excluding non-voters), but the results are by and large similar. In all analyses, an unweighted SOM survey is used as a stand-in population in order to get the joint distribution of a number of psychographics. The weighted SOM-column reports the results when using a weighed benchmark. It was weighted using a 4 covariate cell-weight with 32 cells in total: gender (2 cat) × age (3 cat) × education (3 cat) × region (2 cat) and then raked with vote recall (9 categories). Using a weighted benchmark created more volatilite (and large) adjustments with a few cases where the bias was increased threefold.

Appendix table 8. Rank of weights sorted by median percentile of the proportional bias reduction across 21 different surveys.

Rank	Weight	Median percentile	Mean percentile	Percentile SD	Greatest proportional bias reduction	Lowest proportional bias reduction
1	VE3R3	91.3	82.8	16.5	−70.7	4.9
2	GVE3	91.3	87.0	11.8	−66.9	2.3
3	GVE3R2	91.2	87.8	9.2	−68.1	5.0
4	GVE3R3	90.8	85.4	16.7	−70.8	8.9
5	VE3	90.4	85.4	14.0	−63.6	4.2
6	*GVE3R4*	89.9	86.4	14.4	−71.6	13.0
7	VE3R4	89.9	83.7	15.5	−70.5	12.7
8	GVE2R4	88.6	84.1	14.6	−68.9	10.4
9	VE3R2	88.2	83.3	14.4	−67.8	5.1
10	GVE2R3	87.7	83.4	13.5	−68.2	−0.8
11	GVE2	86.8	83.6	10.5	−62.9	−7.3
12	GVR4	86.8	85.1	10.6	−66.0	10.4
13	VE2R4	86.4	80.8	15.1	−69.1	10.8
14	GVE2R2	86.1	83.6	11.2	−68.2	−1.7
15	GV	86.1	84.6	10.6	−64.1	−0.7
16	VE2	86.0	83.0	11.8	−63.9	−2.1
17	VE2R3	86.0	80.7	14.8	−68.4	5.6
18	V	85.5	82.8	10.6	−63.4	−4.0
19	GVR2	85.0	85.7	7.3	−65.3	0.4
20	VA3E3R4	84.2	73.8	28.2	−65.4	41.0

Comment: The median percentile is calculated using the rank percentiles for each weight's proportional bias reduction in 21 different surveys. Only comparable weights are described here and the total number of weight are 173. For example, the 5 highest ranking weights have proportional bias reductions which are among the top 10 percent most efficient weights—median-wise—across all 21 surveys. The weight with overall best performance in one individual survey, GVE3R4, ends up on the 6th spot. G=gender, A=age, E=education, R=region, V=vote recall.

Appendix E. Figures

Figure 8. Scatterplot of unweighted absolute bias and bias reduction by the inclusion of a vote recall covariate

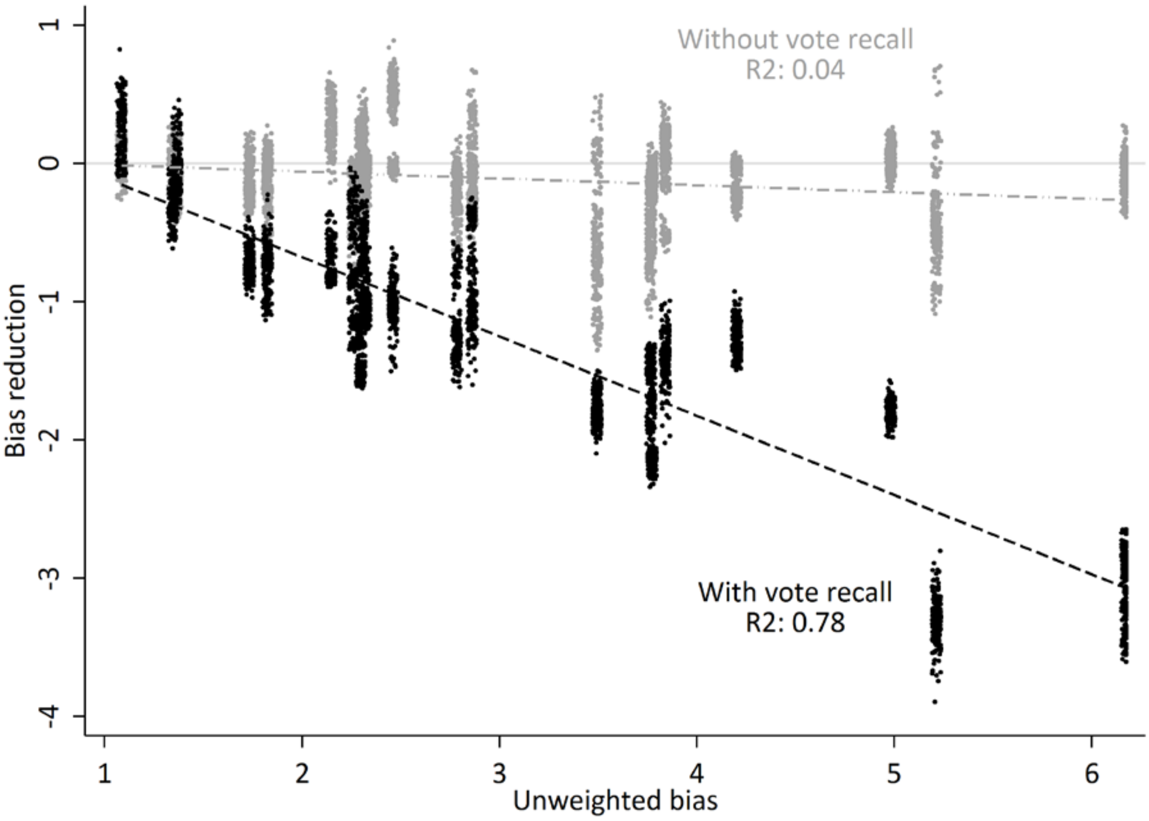
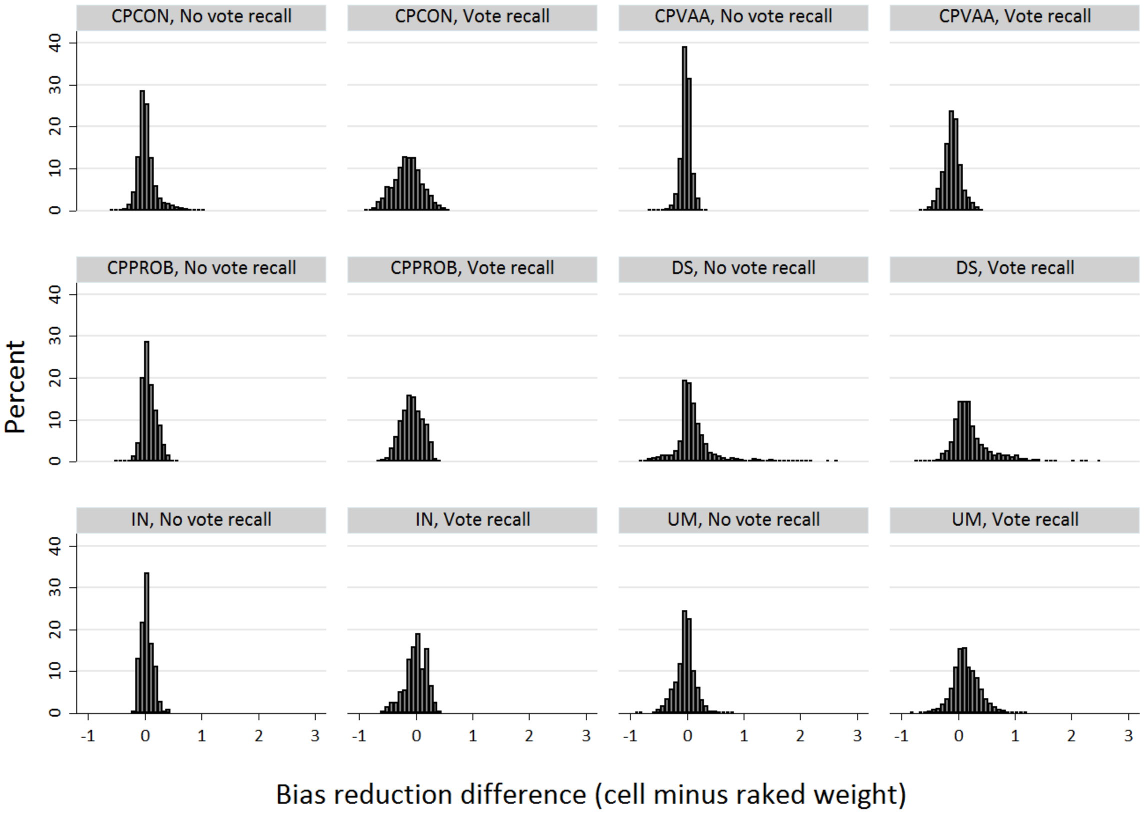


Figure 9. Differences of bias reduction



Comment: The difference is calculated as the difference between the cell-weight subtracted by the raked weight, which means that positive values means a greater negative for cell weights than raked weights (i.e. the better weight) and negative values means that raked weights are better.

Figure 10. Histograms of post-weighting bias by Citizen Panel surveys – all weights

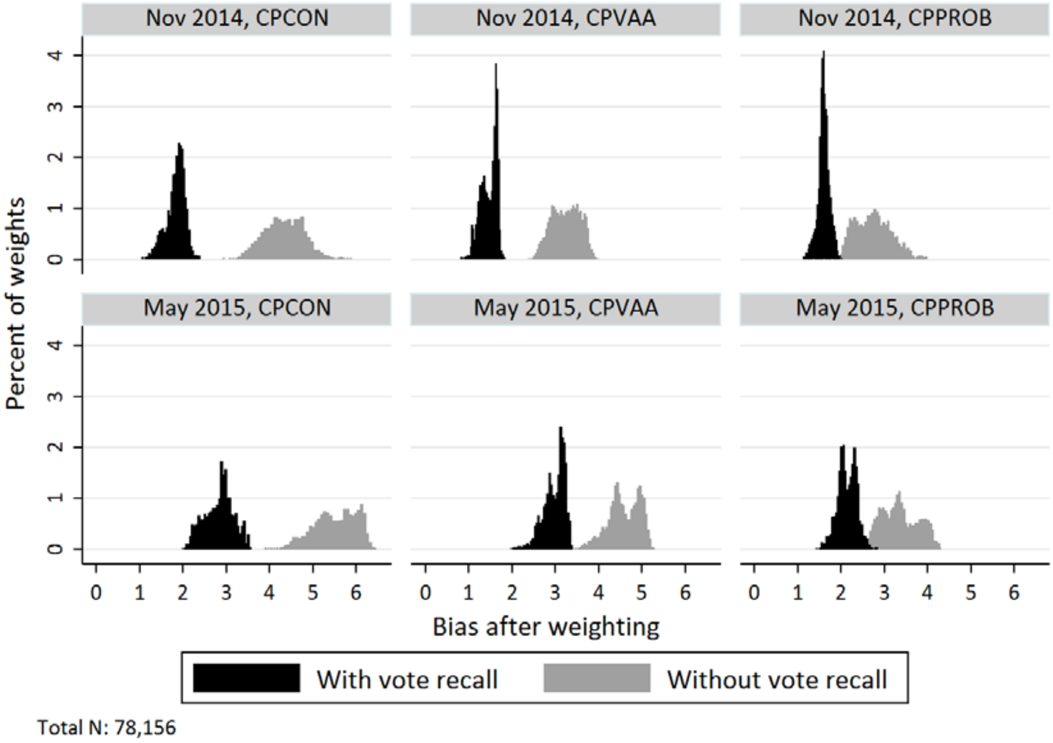
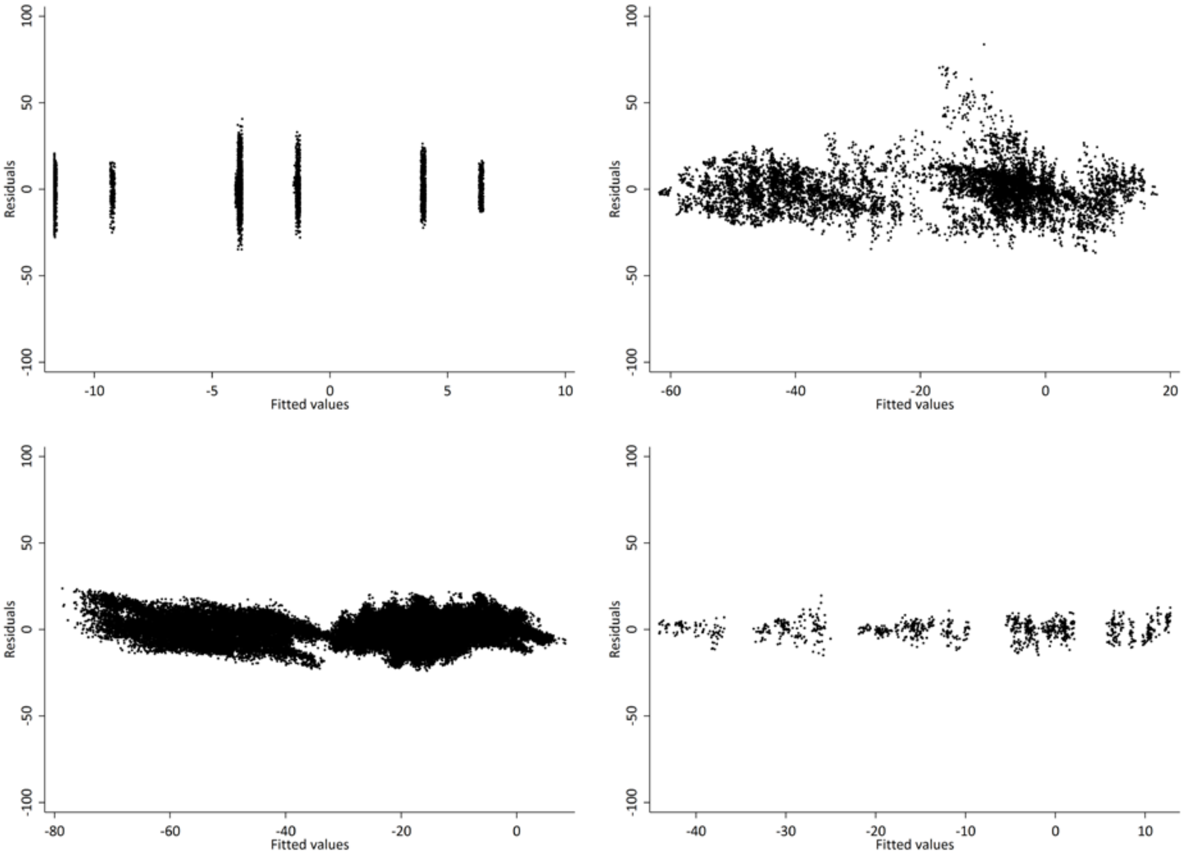


Figure 11. Regression diagnostics of OLS model 1–4 in Table 8. Residuals plotted by fitted weight effect.



Comment: The Stata option -jitter- is used (1). Upper left: model 1; upper right: model 2; lower left: model 3; lower right: model 4.

Figure 12. Estimated bias reduction as a function of an interaction between the vote recall covariate and the distance between election and measurement

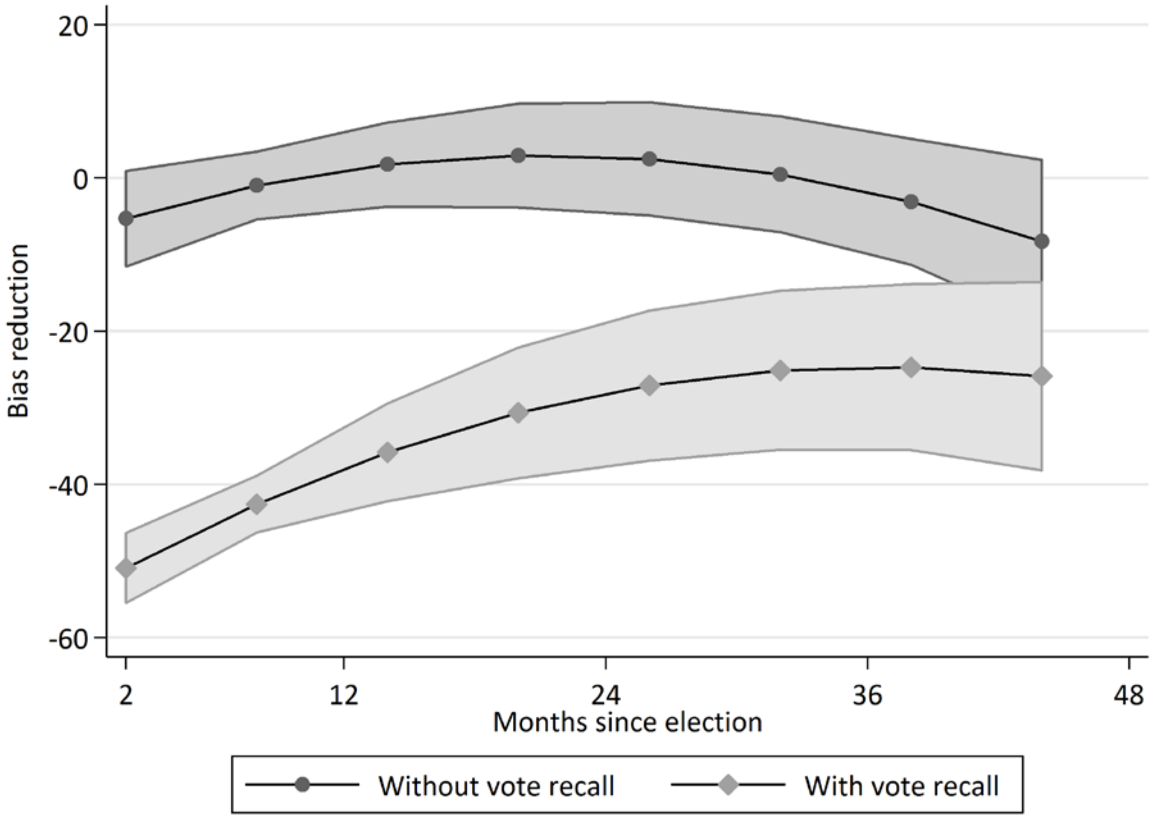
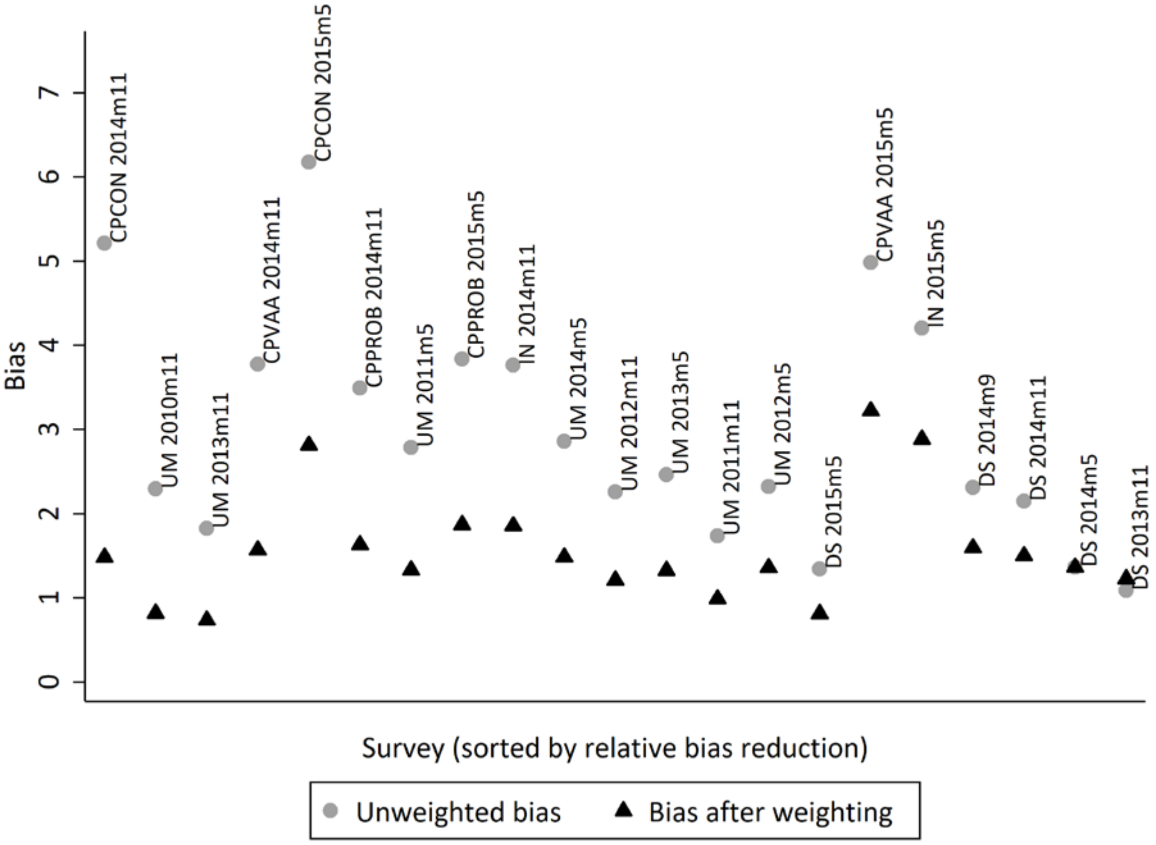
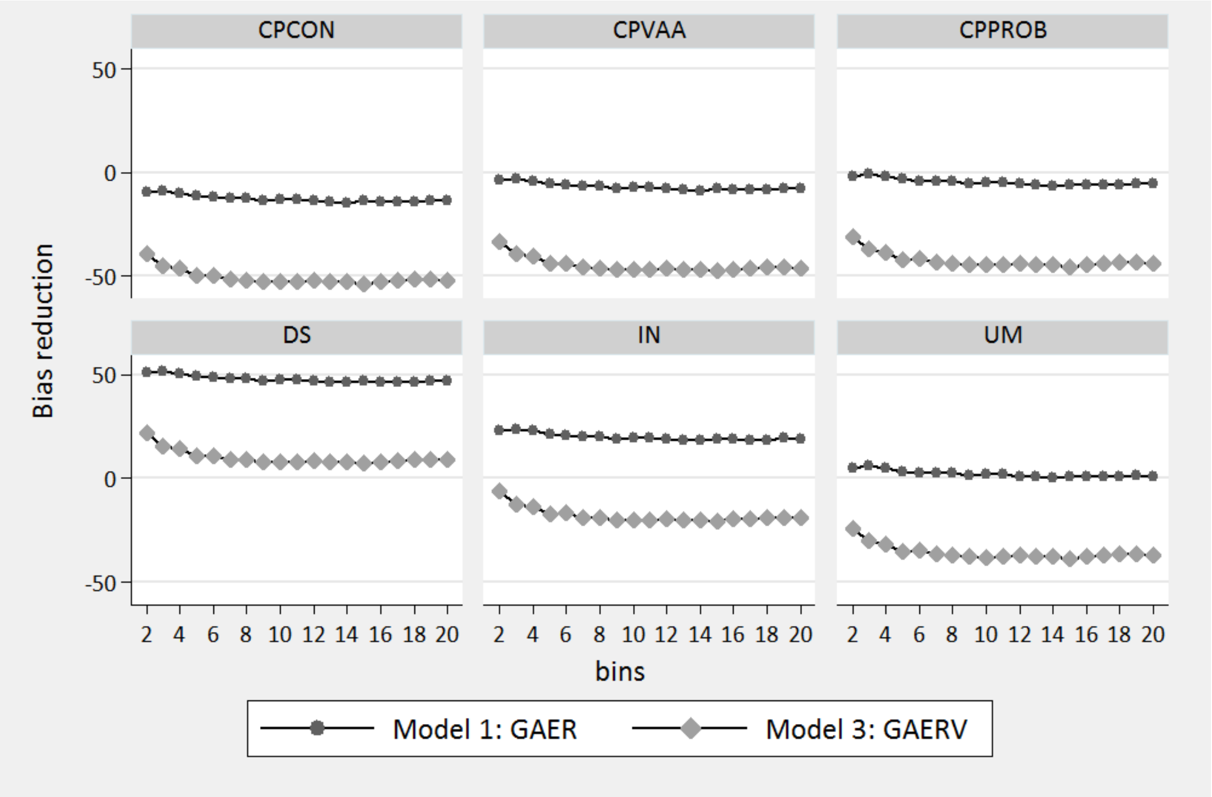


Figure 13. Weight example of GVE3R4 (gender, vote recall, education [3 cat], region [4 cat])



Comment: The examples are sorted by proportional bias reduction in descending order from left to right. The largest bias reduction was 71.6 (CPCON Nov 2014), from 5.22 to 1.48.

Figure 14. Marginal effect of the number of bins on bias reduction in PSA weights


CPCON
CPVAA
CPPROB
DS
IN
UM
Bias reduction
bins
Model 1: GAER
Model 3: GAERV