Tackling Self-Selection into Treatment and Self-Selection into the Sample Biases in VAA Research

Joëlle Pianzola
Andreas Ladner
IDHEAP Lausanne, University of Lausanne

6th ECPR General Conference, University of Iceland, Reykjavik, August 25th-27th, 2011

Abstract. Self-selection into treatment and self-selection into the sample are major concerns of VAA research and need to be controlled for if the aim is to deduce causal effects from VAA use in observational data. This paper focuses on the methodological aspects of VAA research and outlines omnipresent endogeneity issues, partly imposed through unobserved factors that affect both whether individuals chose to use VAAs and their electoral behavior. We promote using Heckman selection models and apply various versions of the model to data from the Swiss electorate and smartvote users in order to see to what extent selection biases interfere with the estimated effects of interest.

Author Notes: This conference paper was written as a part of the project IP16 "smart-voting 2.0" (http://www.nccr-democracy.uzh.ch/research/module5/smart-voting/smart-voting). IP16 belongs to the National Center of Competence in Research "Challenges to Democracy in the 21st Century" (NCCR Democracy; http://www.nccr-democracy.uzh.ch) and is funded by the Swiss National Science Foundation (SNSF).

Introduction

Whenever we are trying to figure out whether VAA use has an effect on users, we are comparing an outcome of interest between voters who used a VAA prior to the elections versus those who did not use the tool. In that sense, VAA use functions as treatment condition, and the difference between treated and non-treated subjects on an outcome of interest gives us the effect or impact the VAA has on individuals. Often, this scenario is an ideal which we cannot approach with the data generating process we employ in VAA research.

In terms of establishing causal relationships between VAA use and voting behavior, an experimental setting would be the preferred approach. The main reason for this being that random assignment to treatment and control group in experiments ensures that on average the difference found between treated and control group can be solely ascribed to having received the treatment (Morton and Williams 2010). In randomly assigning participants to VAA use and thus exerting control over who uses the tool prior to elections would allow us to treat VAA use as an exogenous predictor for an outcome of interest that we would like to compare among users and non-users.

Experiments are, however, still the exception in VAA research than the norm (cf. Vassil 2011b, Ruusuvirta 2011). Most of the time, data gathered by VAA researchers stem from surveys conducted among the electorate or specifically among VAA users. Since random treatment assignment is not feasible in observational studies (Morgan and Winship 2007: 41, Guo and Fraser 2010: 3), the aim is to absorb as many confounding variables to the relation of interest as possible in order to still be able to deduce effects of interest. This endeavor is, however, often limited. Especially if we deal with questions surrounding human behavior, the chance to get to the bottom of certain behavioral patterns by asking people about why they do what they do can easily lead to unreliable or unobserved measurements (cf. Mutz 2007: 92). Hence, the problem of measurement error and omitted variable bias is omnipresent in social behavior research. There is, however, a further major issue surrounding the problem of endogeneity in VAA research, one that intertwines with measurement error and omitted variable bias; and that is selection bias. Whenever a random survey among a population of interest is conducted, the part of the sample in the category of VAA users does not end up there by chance but by self-selection. Furthermore, if we conduct surveys among VAA users themselves, we not only deal with a sample of self-selected users but, if the survey is non-randomly conducted, with a sample of self-selected survey participants. Thus, self-selection into treatment as well as self-selection into the sample constitute major issues for VAA research and these issues need to be tackled if we are trying to detect effects of VAA use.

In this paper, we outline the counterfactual argument and the problem of endogeneity in VAA research and discuss possible detours around the problem. At the same time, we elaborate on the difficulties of going down these roads. We will apply proposed methods to our own data and discuss the outcomes. The aim of outlining these approaches is to improve the future of VAA research.

Measuring VAA Effects

Counterfactual argument

The Neyman-Rubin counterfactual framework of causality is at the heart of investigating causality (Guo and Fraser 2010: 24). The basic idea is to create a scenario where we can infer what would have happened to an outcome of interest if the treatment condition were absent. In general, whenever we are trying to make casual claims, we are trying to compare groups and extract differences that are only due to a change in a main variable of interest. Thus, if we want to see whether VAA use had an effect on users, we ideally want to compare two groups of voters who only differ with regard to whether they had used the VAA for their electoral decision making. Groups should thus be, on average, identical on all other observed and unobserved characteristics. Only then can we deduce changes to a specific cause and effect scenario. In the context of experiments, the random assignment of individuals to groups should ensure such a scenario. All individuals have the same probability of being assigned to either the control or treatment group. Given a large enough sample, the groups should on average be interchangeable with regards to their characteristics. This condition ensures the central component of experimental work; we create a scenario of counterfactuals.

Counterfactuals work with the "potential outcomes framework" (Morgan and Winship 2007: 4), where "what-if" scenarios build the baseline for making causal claims. Scenarios are created where each individual in the population has the same chance for a potential outcome, but we only get to observe the individual in one treatment condition. Given that groups only differ with regard to the treatment condition in randomized settings, we can assume that we observe in the control group what the treatment group would have done had it not been given the treatment and vice versa. Hence, treatment and control group are seen as interchangeable. Although this assumption only holds on the aggregate level (groups), the counterfactual framework allows detecting casual effects (Guo and Fraser 2010: 25). In randomly assigned control and treatment groups, the causal effect is the difference in the means of an outcome of interest for the two groups, and the reliability of this difference can then be statistically estimated (Antonakis and Lalive 2004).

The difference between the two groups on an outcome of interest is the average treatment effect in the Neyman-Rubin counterfactual framework. A condition that has to be met for consistently estimating the average treatment effect is the ignorable treatment assignment assumption (Rosenbaum and Rubin 1983, Morgan and Winship 2007: 41). As Guo and Fraser (2010: 31) put it, "the assumption says that conditional on covariates X, the assignment of study participants to binary treatment conditions is independent of the outcome of nontreatment and the outcome of treatment". In other words, if we hold all observed confounders constant, we assume that the outcome of interest is independent of the treatment *assignment* mechanism (Morgan and Winship 2007: 40). Assigning participants randomly to the treatment or non-treatment condition in classical experiments ensures that the condition is statistically independent of all other variables, confounders as well as outcomes. The important point here is that the average treatment effect only holds if we can assume that there is no selection bias in treatment assignment. In other words, as soon as we have reason to believe that the treatment variable is related through *unobservables* to the outcome variable, the independence of treatment assignment with regard to the outcome of interest no longer holds (Morton and Williams 2010: 113).

The impossibility of randomly assigning survey participants to treatment conditions and the likelihood for unobserved confounders is at the heart of the challenge that observational studies face when trying to detect causal effects. Through the lack of control over the treatment condition, the independence assumption is hard to maintain and the way individuals end up in the treatment condition needs to be investigated (Morgan and Winship 2007: 41). In VAA research, the use of the tool constitutes a treatment condition that is used as an explanatory variable for changes or differences in voting behavior, political preferences or political attitudes. If the data on VAA use is gathered through surveys, it is the individual itself rather than the researcher who selects whether the treatment condition applies or not (Gelman and Hill 2007: 181). This explicit choice of whether to use a VAA or not is a non-random process, systematically distinguishing VAA users from non-users. It is reasonable to assume that voters with a high interest in politics and a high motivation to engage in politics, for example, are more receptive towards using VAAs as a source of information for their electoral choices whereas disengaged and uninterested citizens are less likely to take the time and deal with questions surrounding their political positions and preferences. Hence, VAA user can be assumed to differ substantially from non-users and this difference needs to be accounted for when estimating the potential effects of such tools on voters.

So far, the situation at hand does not differ from a normal regression analysis where the outcome of interest is regressed on the treatment condition and a set of observed control variables. However, as mentioned before, the situation changes when we either cannot or do not observe all possible covariates that predict both the treatment and the outcome of interest (Gelman and Hill 2007: 215). Moreover, some observed variables may not have been measured accurately. In the case of a random sample among the electorate where we know whether participants used a VAA prior to the elections, we have to investigate how participants end up choosing to be part of the "treated" group (those who used a VAA) versus being part of the nontreated group (those who did not use a VAA) and whether this is potentially associated with an outcome of interest. Hence, the self-selection into treatment has to be accounted for in order to ensure that the estimates we retrieve are unbiased. Since we can only make assumptions about the underlying selection mechanism that makes people choose to use VAAs, we might most likely not be able to observe and accurately measure all components that are at play. Think of personality traits or cognitive abilities that might both incline people to make use of VAAs and also the choices they make at elections or the attitudes they have about politics. If this hypothetical scenario turns out to be true, we have a situation where unobserved variables affect both the treatment condition and the outcome of interest, which leads us back to the point where ignorability of treatment assignment is violated. In this case, the effect we measure from regressing or predicting an outcome of interest from the treatment condition and observable confounders will be biased (Antonakis et al. 2010). The logic behind this is the same as dealing with endogeneity in ordinary least squares (OLS) regressions.

Endogeneity in VAA Research

If we know that our main variable of interest is endogenous rather than exogenous, we need to apply appropriate measures to correct for this problem, otherwise the effects we find are not consistent. The importance of consistency for casual inference is of outmost importance, without it the relationship between two variables will never reflect the (true) causal relation. Consistency of estimates basically means that with an increasing sample size, the estimate converges to the true population estimate. Thus, we achieve certain accuracy in the estimation process. Efficiency, on the other hand, is also a desired aspect of estimation, where estimates gain in precision due to smaller estimations of the variance of parameters. In general, we prefer consistent estimates over efficient ones, since there is no scientific gain from shooting over and over at the same point but in the wrong direction. Hence, efficient estimates are worthless if they are biased.

In observational data, treatment assignment is not ignorable and using the treatment condition as a dummy variable might easily lead to endogeneity bias (Guo and Fraser 2010: 32). Through the nonignorable treatment assignment, factors underlying the selection might cause the dummy variable in the model to correlate with the error term of the equation, which leaves us with biased and inconsistent results. If VAA use is seen as the treatment condition, we only observe whether a particular individual used the tool, but what we do not observe is the underlying utility function that leads an individual to the choice of using the VAA. For every observed individual, there is a cutoff value, a tipping point where the decision is made to go one way or the other. Some aspects of that utility function can be observed and controlled for. But naturally, it is often impossible to observe them all since they are either not measured, cannot be measured or are simply unknown. It is exactly this dilemma of modeling causal relations that led Heckman (1978, 1979) to the development of the sample selection model and Maddala (1983) to the treatment effect model (Guo and Fraser 2010: 32). Both models tackle the endogeneity bias and provide statistical means to correct for them.

In the following section, we outline the workings of both the Heckman sample selection model and the treatment effect model and apply these models to data we have collected in Switzerland among smartvote users and data that has been gathered from the Swiss electorate for the 2007 Swiss federal elections.

Heckman Models

Heckman sample selection model

The general research motivation is to collect data from a subsample of a population of interest and draw inferences about the underlying population from the sample collected. As soon as the data collection process prevents us from obtaining a subsample from which we can infer to the population of interest, cautionary measures need to be taken. Sample selection or incidental truncation occurs when the sample data at hand is nonrandomly selected (Greene 2008: 883) and the challenge then lies in modeling the sample selection process in order to make valid claims with regard to the population of interest.

A sample selection model is characterized by the fact that we observe the outcome of interest only for those participants who have self-selected themselves into the sample. If we conduct a survey among VAA users, we only observe the voting behavior (i.e. swing voting) among users who first of all have chosen to make use of the tool and second of all have chosen to participate in the survey. These selection processes do very likely not happen randomly but follow a systematic path, and it is exactly this path towards self-selection that we need to take into account when analyzing the data.

The problem associated with sample selection goes back to the issues discussed in the absence of ignorable treatment assignment. If we want to measure an effect based on two groups of individuals and these groups differ systematically from one another, interchangeability of these individuals is no longer granted and a comparison of these two groups likely yield inconsistent estimates. The only way to retrieve consistent estimates given these two groups is to model the selection mechanisms that led to the systematic difference of the two groups in the first place. James Heckman (1979) won the Nobel Prize in Economics for his work on selection models in 2000, where he found a way to control for the unobserved heterogeneity due to selection in predicting variables of interest. Heckman's sample selection model was a pioneering approach for correcting selection biases (Guo and Fraser 2010: 85).

The logic behind the sample selection model is to estimate the probability of a participant to be in the sample at hand and then use that information for estimating the outcome of interest. The reason to do this is that we assume that factors leading to the choice of being part of the sample or being part of the treatment condition in the sample are unobserved and correlate with our outcome of interest. In following Morgan and Winship (2007: 185), the estimation procedure can be described in the following way.

We specify two equations, a selection equation where mechanisms determining the selection process are modeled and an outcome equation where mechanisms determining an outcome variable of interest are modeled. In the selection equation, participation in the sample or treatment condition is specified based on some observed variables that determine the selection plus an error term that includes all unobserved selection factors. In predicting the selection condition, the dummy variable indicating whether participants have self-selected themselves into the sample or treatment is treated as an endogenous latent variable, and its expected value is estimated based on both observed and unobserved factors. Since we do not observe the unobserved factors, their expected values are calculated[1] and then used as a control variable for estimating consistent effects in the outcome equation of interest. Guo and Fraser (2010: 96) label the error term of the selection equation as "a case of unobserved heterogeneity determining selection bias (which) is treated as a true omitted-variable problem and creatively taken into consideration when estimating the parameters of the (outcome) equation". This procedure is Heckman's lambda method for correcting selection bias, where the inverse Mill's ratio is estimated in the selection equation based on the probability of choosing the treatment, including all unobserved characteristics. In including the inverse Mill's ratio in estimating the outcome equation of interest, we proceed as taking an omitted variable into account (Wooldrige 2002: 567) which removes variance in the error term that is due to selection (Antonakis et al. 2010: 1110) and thus ensures that the errors of the selection equation and the outcome equation no longer correlate.

To apply these theoretical elaborations on VAA research, consider the following Swiss specific scenario. We conducted a survey among smartvote users during the 2007 Swiss federal election campaign. After receiving a voting recommendation from the website, users where asked whether they would participate in our survey. Given that data, we would now like to find out whether the voting recommendation had an effect on users' voting behavior. In modeling such an effect, we need to account for the fact that the data at hand constitutes a self-selected, non-random sample. Hence, in estimating the effect of interest we have to take the selection bias inherent in our data into account. The selection we face here is twofold: first, individuals self-select themselves into becoming smartvote users out of the total population of voters (self-selection into treatment) and then self-select themselves into becoming a participant in the survey (self-selection into the sample). Chances are that those enthusiastic about the opportunities offered by such tools and with a general openness towards new information are first of all prone to use the tool and might also have a higher tendency to answer the survey (cf. Vassil 2011a). Hence, we might end up with a sample of highly enthusiastic and convinced smartvote users who report stronger effects of the tool on their voting behavior than if we had had the chance to survey a representative sample of smartvote users. The likelihood for

[1] To do so, the strong assumption of normal distribution of the unobserved factors needs to be made (Morgan and Winship 2007: 185).

overreporting the effect of the tool on a user's behavior is therefore great. In using a Heckman sample selection model, the aim is to control for the selection mechanism that might end up aggravating effects in the sample compared to the actual population.

In addressing the question of whether the voting recommendation had an impact on users' voting behavior, we are interested to see whether those affected by the tool in their vote choice ended up voting for a different party at the 2007 election compared to the last one. Since we only observe a change in vote choice (dependent variable of our outcome equation) among the self-selected sample of smartvote users, we need to specify the selection equation. The tricky part is that we can only assume the selection mechanism at play, and due to the data structure have to make additional assumptions with regard to the underlying self-selection process. Since we assume that the effect of smartvote use is overreported in our sample, we want to control for the specific differences between smartvote users and non-users in the outcome equation. The rationale behind this is that we assume that the self-selection mechanisms into the treatment and into the sample are similar.

How do we go about modeling the probability of becoming a smartvote users versus not becoming a smartvote user in a sample consisting only of smartvote users? The advantage at hand is that we have a separate randomly sampled survey that was conducted among the Swiss electorate after the elections and which asked participants whether they have used smartvote prior to the elections. Since a lot of question items in the Swiss electorate survey are identical to the questions items in the smartvote survey, we can append the two data sets. First, we dropped all smartvote users in the Swiss electorate survey in order to get a sample of non-users. This non-user sample was then appended to the smartvote survey, creating a data set where we can distinguish between users and non-users based on respective characteristics.

The estimation procedure takes the following steps. First, we specify the selection equation which contains a vector of factors known and observed to influence the probability of becoming a smartvote user as well as an error term which contains the unmeasured characteristics in the selection equation. Since we assume that exactly those unmeasured characteristics of voters both determine the probability for smartvote use as well as the outcome of interest (e.g. openness to persuasion, availability to incorporate new information in decision making etc.), the error terms of the selection equation and the outcome equation are most likely correlated. If this is indeed the case, applying a Heckman sample selection model will take the correlation into account and provide consistent estimates in the outcome equation. Therefore, the aim is to model our outcome equation (with the effect of interest) while controlling for the selection mechanism (Vassil 2011a: 6). If we were to skip the selection equation and the error terms are correlated, the estimates in the outcome regression will be biased. Important to note here is that if the unmeasured characteristics influencing the selection equation do not correlate with the unmeasured characteristics of the outcome equation, the selection can be left out and we can still obtain consistent estimates (Kennedy 2007:

286). Hence, theoretical argumentation and correct model specifications are at the heart of such estimation procedures.

In the following example, we illustrate both estimates obtained from an ordinary probit regression and a Heckman sample selection model. The outcome of interest is swing voting, a binary variable that requires a bivariate probit model[2] since we have a probit model in the selection equation and a probit model in the outcome equation. Our main effect of interest is whether a user was affected by the voting recommendation in his or her vote choice, while we control for factors such as being surprised by the voting recommendation, multiple vote propensities, party attachment, political ideology, number of candidates running per seat as well as age (for variable details, see Appendix A). In the selection equation, we treat the dummy variable for smartvote use as an endogenous dependent variable and a function of the following observed variables: age, education, income, gender, multiple vote propensities and visiting political homepages of candidates and parties prior to the elections. From the representative Swiss electorate study SELECTS[3], we know that smartvote users are generally younger, well educated, with high incomes and male. At the same time, visiting political websites is a proxy for online affinity and for engaging in and being interested in gathering political information online. Multiple vote propensities are an indicator for openness to a vote change or indecision with regard to one's vote choice. Except age and multiple vote propensities, the other variables are not included in the outcome equation since they can be argued not to be directly linked to the outcome variable of interest but directly linked to the likelihood for smartvote use (called the exclusion restriction). In applying a Heckman sample selection model we assume that unobserved factors that determine whether a voter uses smartvote and ends up in the survey sample are correlated with unobserved factors that are associated to swing voting. Given our data structure, the selection equation is fully observed, while we only have a selected (censored) sample for the outcome equation. Table A illustrates the direct marginal effects of the coefficients for both a regular probit model and a probit model with sample selection. The variables listed with empty cells have been exclusively used to estimate the selection equation.

[2] Using STATA's command *heckprob*
[3] http://www2.unil.ch/selects/?lang=en

Table A. The effect of the smartvote voting recommendation on vote choice – Estimation with and without selection[4]

Swing voting	Probit model	Probit model with Heckman sample selection
affected by voting recommendation	0.09***	0.08***
	(-0.02)	(-0.02)
rather not surprising	0.04	0.03
	(-0.02)	(-0.02)
rather surprising	0.05*	0.05
	(-0.03)	(-0.03)
very surprising	0.18***	0.15**
	(-0.06)	(-0.07)
multiple vote propensities	0.06**	0.04
	(-0.02)	(-0.03)
party attachment	-0.11***	-0.18***
	(-0.02)	(-0.02)
left	0.05**	0.06**
	(-0.02)	(-0.02)
center	0.14***	0.13***
	(-0.03)	(-0.03)
age	-0.003***	-0.003***
	(-0.001)	(-0.001)
gender		-
education		-
income		-
visited homepage		-
Observations	2,678	5,895
Log likelihood	-1696	-3943
ρ		-0.12 (chi2(1)=2.42, p>0.1)

Average direct marginal effects (discrete change for dummy variables)
Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

[4] For visual reasons, the 14 fixed effects for the number of candidates running per seat are excluded from the table.

In this example, the ρ (rho) of the Heckman sample selection model is non-significant, indicating that we cannot reject that the presumed correlation between the error terms of our specified equations is in fact zero – hence, in this case, if we correctly specified our model, it would be sufficient to retrieve the estimates from the regular probit model. The effects are practically the same in the two models, thus we do not observe any change in the predicted estimates due to sample selection bias. At this point, we could speculate that we have observed and controlled for all important factors that are linked to the self-selection process or that our assumption that the underlying selection process between selection into treatment and selection into the sample is equivalent is not adequate. The answer rests on the suitability of the theory we use for specifying our model.

With regard to the outcome, our model indicates that those affected in their vote choice by the voting recommendation had a significantly higher probability to change their vote choice compared to those who were not affected by the tool. Those users who stated that they were rather surprised or very surprised by the voting recommendation they received compared to those who were not surprised at all also had a significantly higher probability for changing their vote choice. Younger voters and those without a specific party attachment were more likely to change their party choice between the two elections, and voters with a political ideology on the left and in the center, compared to those on the right, were also more likely to switch their vote. Multiple vote propensities significantly increase the chance for swing voting, however, the effect was non-significant in the Heckman sample selection where multiple vote propensities were part of the selection equation.

In specifying this model, we relied on several assumptions in modeling the selection equation. Favorably, we would have a bunch of variables that are only associated with smartvote use and completely unassociated with the outcome of interest. If we find such variables, then selection models are closely related to instrumental variables models, where exogenous variables are used to predict the selection but not the outcome (more on IV estimation in the next section).

Heckman treatment effect model

Heckman sample selection models differ from Heckman treatment effect models (Maddala 1983) mainly in two ways: first in how the sample at hand is structured and second in the estimation procedure. In sample selection models, we only observe the outcome variable of interest for those exposed to the treatment whereas in treatment effect models we observe the outcome variable for both the treated and non-treated participants. This in turn changes the selection bias estimation procedure. In case of treatment effects models, the treatment dummy is, besides the inverse Mills ratio, included in the second stage equation (Guo and Fraser 2010: 97). The main issue in treatment effect models is that the assignment to the treatment condition is non-random (and thus endogenous) and possibly related to the outcome – hence, as in sample selection models, we assume a possible correlation between the error terms in the selection equation and the outcome equation.

The aim of the treatment effect model is to approximate a randomized experiment with observational data. A (mostly) binary treatment condition is specified and the model sets out to overcome the fact that the treatment condition is not randomly assigned by the researcher but rather participants self-select themselves into a treatment condition. In our case, whenever we survey a (random) population of the electorate, voters themselves chose to self-select themselves into our treatment condition of interest, VAA use. In treatment effects model, we measure the outcome for both treated and nontreated groups, the only difference is that we have to account for the selection bias that preceded the treatment condition. In econometric terms, we need to account for the fact that the binary treatment variable is not exogenous but rather endogenous – again, the situation where the treatment variable possibly correlates with the error term of the equation. As in the Heckman sample selection model, the selection equation into treatment is based on a latent model, where we only observe whether individuals have used smartvote or not.

The problem of self-selection is inherent in observational data on VAA users, therefore the aim should be to model the self-selection into treatment whenever we compare users to non-users. The challenge lies in finding appropriate instruments that do predict participation in the treatment or control group and do explain the selection mechanism. As an example here, we are only interested whether voters who used smartvote prior to the elections have a higher probability to swing vote than non-users. In this case, we view smartvote use as a treatment condition, and we assume that becoming a smartvote users does not happen randomly among the electorate, there is an explicit selection process that causes individuals to use the tool. Therefore, in a treatment effects model, we first specify a selection equation that predicts the treatment condition and then specify an outcome equation where the treatment condition is included as a variable in the model (since we observe the outcome of interest for both treated and non-treated groups).

The dataset we use for the treatment effect model consists of Swiss voters and their electoral behavior and preferences and among which we can distinguish between smartvote users and non-users. Since the outcome of interest is a binary variable indicating swing voting, we again specify a bivariate probit model[5]. In predicting an effect of smartvote use on swing voting, we control for variables such as multiple vote propensities, party attachment, political ideology and age. In the selection equation, we predict smartvote use as a function of age, income, education, gender, multiple vote propensities and having visited political websites prior to the elections. In this model,

[5] Using STATA's command *biprobit* or *biprobittreat*.

the variable smartvote is the binary endogenous treatment variable upon which the selection process is modeled. The following table (see Table B) gives the results for the outcome equation of the ordinary probit model and the Heckman treatment effect model:

Table B. The effect of smartvote use on the probability of swing voting – a bivariate treatment effect model

swing voting	Probit model	Bivariate probit treatment effect model
used smartvote	0.39***	0.64***
	(0.05)	(0.12)
multiple vote propensities	0.28***	0.21***
	(0.05)	(0.06)
party attachment	-0.39***	-0.39***
	(0.04)	(0.04)
left	-0.16***	-0.14***
	(0.05)	(0.05)
right	-0.22***	-0.21***
	(0.06)	(0.06)
age	-0.01***	-0.003*
	(0.001)	(0.002)
education		-
income		-
gender		-
homepage		-
Constant	-0.28***	-0.53***
	(0.09)	(0.15)
Observations	4463	4,067
Log likelihood	-2603	-4089
ρ		-.18** (chi2(1)=5.1 p<.05)

Notes: education and income are coded as dummy categories with a reference group
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficients reported here are the probit coefficients, thus we can only interpret the sign and significance but not the magnitude of the effect. The selection equation is not reported here, but smartvote use was predicted from age, education, income, gender, multiple vote propensities and visiting websites of politicians. Rho (ρ) is the estimated correlation between the error terms of both equations and significant in this model, pointing to a selection bias in our model. A negative rho indicates that the treatment effect is underestimated by an ordinary probit model where the selection bias is not considered. Taking this into account in the treatment effect model, the coefficient on smartvote use is positive and significant, indicating that those voters who used smartvote prior to the elections had a higher probability to switch their vote choice compared to non-users. All confounders in the model are significant and point towards the expected direction. Although this looks great in terms of modeling the inherent selection bias, treatment effect models have one major drawback.

We make strong assumptions when using such models. Specifically, we assume a joint bivariate distribution of the error terms (of the latent selection model and the outcome equation) (Morgan and Winship 2007: 185). In treatment effect models, correct specification of the model is therefore essential for ensuring that the estimates we get from the model are not biased. A way to check for this is to run a Murphy's score test (Murphy 2007, Chiburis 2010) to test the goodness of fit of the bivariate probit model and test whether the model assumptions are indeed satisfied. According to the Murphy's score test, our model is misspecified (chi2(9)=19.9, p<0.05). In other words, the bivariate normal distribution of the error terms in our equations does not hold. It is well known that the Heckman treatment effect model is extremely sensitive to misspecifications (Guo and Fraser 2010: 124) since it makes very strong assumptions about the model properties to begin with. If the model is misspecified, the results can again be biased (ibid.). According to Chiburis et al. (2011), a way around the assumption of jointly normal error terms in bivariate probit models is using two-stage least squares instrumental variable (IV) estimation instead. In IV estimation, we need to find a variable or variables that affect selection into treatment but not the outcome of interest (Guo and Fraser 2010: 99). Instrumental variable estimation is useful if the bivariate probit model is misspecified because this method does not rely on joint normality of the disturbances (Greene 2008: 893). In two-stage least squares, we use the exogenous instruments to predict our endogenous variable in the model and substitute the predicted values for the actual value of the endogenous regressor in the outcome equation (Wooldridge 2002: 484). In doing so, we eliminate the part of the endogenous variable that might be correlated with the error term. Similar to the Heckman models, IV is used when we face problems of omitted or unknown control variables (Angrist and Pischke 2009: 84). Since we have a misspecified bivariate probit model, we will now apply the IV approach to our example.

In our example, we know that education, income and gender do not affect the likelihood for swing voting but are significant predictors of smartvote use. Hence, we will run a two-stage least squares IV analysis on our model, where education, income and gender will serve as exogenous instruments (model 2). At the same time, as a comparison we are running a regular OLS regression where smartvote use is not instrumented (model 1). Here are the results (see Table C.):

Table C. The effect of smartvote use on the probability for swing voting – linear versus two-stage least squares regression

	Model 1 OLS swing	Model 2 IV swing
used smartvote	0.13***	0.21***
	(0.02)	(0.04)
multiple vote propensities	0.09***	0.07***
	(0.02)	(0.02)
party attachment	-0.13***	-0.12***
	(0.01)	(0.01)
left	-0.06***	-0.07***
	(0.02)	(0.02)
right	-0.07***	-0.08***
	(0.02)	(0.02)
age	-0.00***	-0.00**
	(0.00)	(0.00)
constant	0.40***	0.33***
	(0.03)	(0.05)
Observations	4,463	4,147
R-squared	0.08	0.07

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Instrumenting smartvote use yields a significant effect of using the tool on swing voting. The nice thing about two-stage least squares estimation is that we can test the validity of the instruments used. In our model, the Hansen test for overidentification restrictions is non-significant (chi2(9)=6.4 p>0.1), indicating that we used valid instruments in the model. The Hausman endogeneity test reveals that the retrieved IV estimates differ significantly from the regular OLS estimates (chi2(1)=5.6, p<0.05), thus we should follow the IV estimates since they are more consistent. If we compare the IV results to a regular OLS estimation, we find that the "true" coefficient on smartvote use is 59% bigger[6].

Although linear instrumental variables are an alternative to the bivariate probit model if the latter is misspecified, the effects we measure are somewhat different from the ones we measure in treatment effect models. IV estimation consistently estimates what Imbens and Angrist (1994) called the Local Average Treatment Effect (LATE), while bivariate probit models produce the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT) (Chiburis 2011: 3). ATE gives us the causal effect

[6] In using this approach, we disregard the binary structure of our dependent variables, a common approach for estimating causal effects that are misspecified in bivariate probit models (Chiburis et al. 2011: 2).

of the treatment from a randomly drawn individual in the population whereas ATT gives us the causal effect of the treatment on a randomly drawn individual who actually received the treatment (Morton and Williams 2010: 95). LATE measures the causal effect of the treatment for those in the sample that *comply* with the treatment, for those "whose treatment status can be changed by the instrument" (Angrist and Pischke 2009: 114). Therefore, the estimators from both models likely differ since they are not measuring the same treatment effect (Angrist 1991). In our example, the Average Treatment Effect in the bivariate probit treatment effect model is .20 and the Average Treatment Effect for the Treated .21, thus similar to the IV estimate of .21.

The main challenges

The issue of selection bias in observational data on VAA research can be overcome by either testing the results in experimental settings or by applying adequate statistical techniques that take the selection bias explicitly into account. Although labeled as the golden standard, experimental settings also often suffer from a lack of representativity with regard to the actual population of interest (cf. Morton and Williams 2010). Ideally, the experiment would be conducted among a representative sample of the electorate, since we are interested in the effects of the VAA on voters. More often than not, however, students are recruited for experimental research, limiting the explanatory power of the results to the sample at hand. Hence, selection biases also occur in experiments, either induced through the researcher or through those willing to participate in the experiment (self-selection into the sample) (Guo and Fraser 2010: 90). The main advantage of experiments is, however, that random treatment assignment satisfies the ignorability of treatment assumption and confident causal claims can be made – bounded to the selected sample. Outside of the experimental setting, the statistical means to correct for selection bias basically try to create a situation where we also end up with ignorability of treatment assignment, in spite of the fact that we have no control over who ends up in our treatment condition or sample and who does not.

In establishing this scenario, the challenge lies in modeling a theoretically sound selection process and naturally a theoretically sound outcome equation of interest. Only then can we be confident that our improvement techniques actually did improve our results. Since more often than not we have to rely on assumptions about the causes of the selection process, difficulties remain. If the selection process remains a mystery, an adequate way around the endogeneity problem would be to find suitable instruments that highly predict VAA use but are unrelated to any subsequent behavior induced by the tool.

Not only is adequately modeling the selection process a difficult task, but the model itself also has some obstacles in store. In order to make use of the information in the error terms of our equations of interest, strong assumptions are made about the nature of the joint distribution of those error terms. Hence, selection and treatment models are sensitive to model misspecifications, which need to be tested after the estimation procedure (Greene 2008: 891, Guo and Fraser 2010:124).

The models used in this paper give a first impression of how selection biases need to be taken into account in VAA research, but still give much room for improvement. The effect of smartvote use, for example, might very likely vary depending on personal characteristic of users. Thus, in the next step, the conditionality of effects need to be tackled. Since we simultaneously measure two equations with common variables in Heckman models, the interpretation of the effects are more challenging than in ordinary regression or logit/probit models. As soon as there is a selection bias ($\rho \neq 0$) and a predictor appears in both equations, then we get both a direct effect and an indirect effect of the predictor, where both have to be taken into account when calculating the marginal effects in bivariate probit models (Greene 2008: 885). Adding interactions into the equations thus makes the interpretation of the effects even more difficult. For now, we have ample reason to assume that we need to control for the selection mechanism when we analyze effects of VAA use, since the consistent IV estimate of smartvote use yields significantly different results than we would obtain without controlling for the selection bias.

Conclusion

Choices are always at the outset of modeling any kind of social behavior. What we usually end up observing is whether an individual decided for either A or B, but how this decision came about generally remains hidden. It is exactly this process of arriving at the final observed decision that is essential if we are to compare groups of individuals based on the decision they have taken. If we are aware of the underlying process leading up to a decision, we can make sure to observe it and control for it whenever we make the comparison. But if that underlying process is obscure, other steps need to be taken if we want to make valid claims.

In VAA research, we are generally interested in whether the information provided by the tool has any influence on people's political decisions or attitudes and if so to what extent. The data gathered through observational studies faces two challenges for making any statements about these questions: first, even if we conduct a survey among a random population sample, individuals in the sample who have made use of VAA services have decided to do so out of several reasons which likely distinguish them from non-users. Second, surveys conducted among VAA users are generally non-representative since we do not know the population. In the first case, we are dealing with selection into treatment whereas in the second case we are dealing with both selection into treatment and selection into the sample. As soon as we have reason to believe that these selection processes are linked to an outcome of interest, the task is to take these selection biases into account. If we do not do this, standard techniques such as OLS or logit/probit will give us inaccurate estimates from which we draw inaccurate conclusions (Sartori 2003).

In this paper, we tried to illustrate this problem on a concrete example from Swiss VAA data. We were interested to see whether smartvote users who were affected by the voting recommendation in their vote choice were more inclined to change their vote choice at the polls and whether smartvote users in general, compared to non-users, had a higher probability for swing voting in the last Swiss federal elections. In trying to take the selection biases into account in our models, our results indicate that smartvote use does indeed have an effect on users vote choice.

As the example has demonstrated, modeling the selection bias is a daunting task. Not only do we have to make reasonable assumptions about a process of which we do not actually know much about but we are also faced with strong model assumptions that can leave us strayed. Especially in models where not only the selection but also the outcome of interest is binary such as in our examples, arriving at a jointly normal distribution of the error terms is hard to satisfy. Essentially, the challenge for future research is to find better explanations for why individuals end up choosing to use VAAs or, even better, find adequate instruments that highly predict VAA use but are unassociated with political choices and attitudes. For now, awareness of the outlined techniques and applying them to the analysis of observational data should become a standard task in research on VAA's and their impact on electoral behavior.

References

Angrist, Joshua D. (1991). Instrumental Variable Estimation of Average Treatment Effects in Econometrics and Epidemology, NBER Technical Working Paper no. 0115.

Angrist, Joshua D. and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton and Oxford: Princeton University Press.

Antonakis, John, Samuel Bendahan, Philippe Jacquart and Rafael Lalive (2010). On Making Causal Claims: A Review and Recommendations. *The Leadership Quarterly* 21, 1086-1120.

Antonakis John and Lalive Rafael (2011). Counterfactuals and causal inference: Methods and principles for social research. Review of S. L. Morgan and C. Winship. *Structural Equation Modeling*, *18*, 152-159.

Chiburis, Richard C., Jishnu Das and Michael Lokshin (2011). A Practical Comparison of the Bivariate Probit and Linear IV Estimators. World Bank Policy Research Working Paper 5601.

Chiburis, Richard C. (2010). Score Tests of Normality in Bivariate Probit Models: Comment. Working paper, University of Texas at Austin.

Greene, William H. (2008). *Econometric Analysis* (6th ed.) Upper Saddle River: Pearson.

Guo, Shenyang and Mark W. Fraser (2010). *Propensity Score Analysis. Statistical Methods and Applications*. Los Angeles: SAGE Publications.

Heckman, James J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System, *Econometrica*, 46(6): 931.959.

Heckman, James J. (1979). Sample selection bias as a specification error, *Econometrica*, 47: 153-161.

Imbens, G. and Joshua D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects, *Econometrica* 62(2): 467-475.

Kennedy, Peter (2007). *A Guide to Econometrics*. 5th Edition. Malden: Blackwell Publishing.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, MA: Cambridge University Press.

Morgan, Stephan L. and Christopher Winship (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Morton, Rebecca B. and Kenneth C. Williams (2010). *Experimental Political Science and the Study of Causality. From Nature to the Lab*. Cambridge: Cambridge University Press.

Murphy, A. (2007). Score Tests of Normality in Bivariate Probit Models, *Economics Letters*, 95(3): 374-379.

Mutz, Diana C. (2007). Political Psychology and Choice, in: Dalton, Russell J. and Hans-Dieter Klingemann (eds.), *The Oxford Handbook of Political Behavior*. Oxford: Oxford University Press, 80-99.

Rosenbaum, P. R. and Donald R. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.

Ruusuvirta, Outi (2011). Mirror, Mirror on the Wall, Who is the Most Suitable of Them All? The Effects of Online Voting Advice Application Use on Citizen's Electoral Behavior. Paper presented at the 69th Midwest Political Science Association Annual Conference. March 31st-April 3rd 2011, Chicago, USA.

Sartori, Anna (2003). An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions, *Political Analysis* 11, 111-138.

Vassil, Kristjan (2011a). The Impact of VAAs on Vote Choice. Evidence from Switzerland. Paper presented at the 69th Midwest Political Science Association Annual Conference. March 31st-April 3rd 2011, Chicago, USA.

Vassil, Kristjan (2011b). Voting Smarter? The Impact of Voting Advice Applications on Political Behavior. Forthcoming doctoral thesis. EUI Florenz.

Appendix

A. Variable Descriptives

	Heckman selection model	Heckman treatment effect model (bivariate probit)	IV model	
Observations	5895/2481	4067	4147	
Variable	mean	mean	mean	Scale
Swing voting	0.32	0.32	0.32	Dummy (1= swing voter, 0=stable voter)
Smartvote use	0.42	0.63	0.64	Dummy (1=user, 0=non-user)
Multiple vote propensities	0.52	0.65	0.66	Dummy (1=multiple, 0=non-multiple)
Party attachment	0.49	0.58	0.58	Dummy (1=attached, 0=not attached)
Left	0.35	0.43	0.43	Dummy (1=left of the political spectrum, 0=reference group center)
Right	0.38	0.25	0.25	Dummy (1=right of the political spectrum, =reference group center)
Age	48	47.5	47.5	Continuous (18-96)
Education	2.9	3.2	3.2	Dummy (5 categories into 4 dummies)
Income	4.1	4.4	4.4	Dummy (6 categories into 5 dummies)
Gender	0.42	0.33	0.33	Dummy (1=female, 0=male)
Homepage visit	0.21	0.28		Dummy (1=yes, 0=no)