The Wisdom of Crowds Design for Sensitive Survey Questions

Roni Lehrer, PhD (University of Mannheim, Collaborative Research Center 884, B6, 30-32, 68131 Mannheim, lehrer@uni-mannheim.de)

Sebastian Juhl (University of Mannheim, Collaborative Research Center 884, B6, 30-32, 68131 Mannheim, sjuhl@mail.uni-mannheim.de)

Thomas Gschwend, PhD (University of Mannheim, School of Social Sciences, A5, 6, 68131 Mannheim, gschwend@uni-mannheim.de)

Keywords:

Social Desirability, Measurement, Sensitive Questions, Wisdom of Crowds, Vote Choice

Abstract:

Survey research on sensitive questions is challenging because respondents often answer untruthfully or completely refuse to answer. Existing survey methods avoid social desirability bias by decreasing estimates' efficiency. We suggest the Wisdom of Crowds survey design that avoids bias without decreasing efficiency. We compare the design's estimate of a right-wing populist party's vote share to the real election results in the 2017 German federal election. We find that the Wisdom of Crowds design performs best in terms of both bias and efficiency while other techniques for sensitive questions perform less well. We conclude that the Wisdom of Crowds design is an important addition to the social scientists' survey methodology toolbox.

Word count: 6900 words

Social desirability bias is a type of measurement error that arises when survey respondents refuse to participate in the survey or give untruthful answers because they do not want to reveal that they have a socially undesirable characteristic, e.g., being involved in criminal activity or holding racist sentiments (Philips and Clancy 1972). It is widely acknowledged that social desirability bias afflicts survey research whenever respondents are asked for sensitive characteristics. Even worse, experimental evidence shows that an explicit assurance of confidentiality does not inevitably lead to a higher willingness to participate in a survey that features sensitive questions (e.g., Singer et al. 1995; Singer and Hippler 1992). Yet, since such characteristics are at the core of many important research questions, the development of survey techniques that mitigate social desirability bias is of outmost importance for several subfields of the social sciences.

Previous approaches to tackle social desirability bias, e.g., different indirect questioning techniques, trade-off measurement error (bias) for increased variability of measurements (efficiency). They do so by adding noise to respondents' answers which prohibits researchers from learning whether an individual respondent has the sensitive characteristic or not. Since researchers know the distribution of noise across all respondents they can still obtain an unbiased measurement of the aggregate share of respondents that hold the sensitive characteristic. By design, the added noise, however, makes the measurements subject to additional variation resulting in wider confidence bounds around the (unbiased) measurement.

We suggest the Wisdom of Crowds survey design (Murr 2011, 2015, 2016, 2017) that avoids this trade-off between anonymity and efficiency. Thereby, it allows researchers to obtain unbiased and more precise estimates of sensitive questions than with previous approaches. The underlying concept is that we suggest asking respondents what they belief the share of the population is that has the sensitive characteristic of interest. Clearly, as this is not a sensitive question, respondents should be more likely to answer truthfully. As Surowiecki (2004, see also Galton 1907; Page 2007) shows, crowds are wise and their mean beliefs are an unbiased estimate of reality when the crowd members are decentralized, independent, diverse thinkers, and have access to an aggregation mechanism.

We find supportive evidence for the Wisdom of Crowds survey design's superiority in comparison to a list experiment, a Randomized Response design, and a sensitive direct questioning technique in the context of the 2017 German Federal Election. We predict the vote share of the right-wing populist party *Alternative für Deutschland* (AfD) whose vote share was systematically underestimated in most public opinion polls (Bergmann and Diermeier 2017; Waubert de Puiseau et al. 2017). Implementing all these survey designs in the German Internet Panel (GIP) which is an online panel study based on a representative sample of the German population three months before the election (Blom et al. 2015), we predict the vote share for the AfD best and most efficiently with the Wisdom of Crowds design. Moreover, we explore under what conditions the Wisdom of Crowds survey design fails to perform well.

The Wisdom of Crowds survey design is used as an approach to election forecasting with considerable success (Graefe 2014). Our contribution is to show that this design belongs in the social scientists' survey methodology toolbox for dealing with sensitive survey questions. It enables researchers to obtain unbiased estimates about sensitive traits without bearing efficiency losses of standard approaches that tackle social desirability bias, e.g., list experiments or randomized response techniques, under fairly mild conditions. The efficiency gains of the Wisdom of Crowds approach vis-à-vis other indirect questioning techniques becomes essential in situations where researchers wish to make accurate and precise predictions. This allows researchers to be more certain about their measurements and to test hypotheses more precisely and reliably.

In the following, we present the Wisdom of Crowds survey design and summarize other more prominent survey techniques for asking sensitive questions. We then turn to our implementation of these techniques in the context of the German 2017 federal elections and present our comparative results. Next, we show the consequences of violating the Wisdom of Crowds assumptions for its results. The last section concludes and provides an outlook on routes for future research.

1. Bias, Efficiency and the Wisdom of Crowds Survey Design

Every measurement technique's purpose is to correctly estimate a specific quantity of interest while minimizing random variations between different measurements. Therefore, a technique's quality is often gauged by its so-called signal-to-noise ratio. This ratio indicates a measurement's informational content, or its signal power, compared to the background noise it captures. Maximizing the signal-to-noise ratio increases both the results' validity and reliability. However, if the instrument is biased, it systematically recovers an incorrect signal and its measures are of little use for inferences about the target population.

In survey research, it has long been recognized that, depending on the survey item, respondents intentionally answer some questions incorrectly, i.e., they send the wrong signals (Philips and Clancy 1972; Coutts and Jann 2011). One reason for survey respondents to obscure their true response is a phenomenon called "social desirability bias" (e.g., Höglinger et al. 2016; Streb et al. 2008; Tourangeau et al. 2000; Tourangeau and Yan 2007). If they perceive the honest response as socially undesirable or conflicting with social norms, respondents are prone to intentionally give incorrect answers or no answers at all. This leads to a systematic distortion of survey responses in which deviant behavior is underrepresented. The more controversial an issue, the greater the potential effect of this bias on survey responses (Streb et al. 2008; Crowne and Marlowe 1960). What is more, research found that no survey mode is completely immune to such bias. Although more pronounced in interviewer-administered questionnaires, social desirability bias also affects self-administered web surveys (Chang and Krosnick 2010; Kreuter et al. 2008). Accordingly, the accurate measurement of sensitive traits, opinions, or attitudes with surveys, constitutes a methodological challenge since it requires researchers to develop instruments that not only maximize the signal-to-noise ratio but also ensure that the instrument itself does not systematically distort the signal.

When estimating the aggregate share of a target population which has a potentially socially undesirable characteristic, research suggests the application of indirect questioning techniques which assure anonymity to individual answers in order to elicit honest responses to a potentially sensitive question (e.g., Waubert de Puiseau et al. 2017). The underlying rationale is that when respondents know that researchers cannot infer from their individual response whether they have the (potentially) undesirable characteristic, respondents will be more willing to answer sensitive questions truthfully. Several studies present empirical evidence for the ability of these techniques to recover a higher share of respondents that hold the sensitive characteristic as compared to the direct questioning (e.g., Höglinger et al. 2016; Jann et al. 2012; Coutts and Jann 2011). Despite this, we note that these techniques come with problems that researchers would like to avoid. More precisely, indirect questioning techniques are by design statistically less efficient than direct questioning. Hence, they are of limited use when researchers need to measure relatively small differences or seek to predict future developments.

Therefore, we suggest an alternative technique that does not suffer from efficiency losses while simultaneously avoiding distortions arising from social desirability bias: the Wisdom of Crowds (e.g., Murr 2011, 2015, 2016, 2017). We suggest asking respondents directly what share of the population has the socially undesirable characteristic. Below we argue that this design is more efficient because it relies on a direct question, yet, at the same time does not require respondents to reveal information about their own behavior and is thus not subject to social desirability bias. Additionally, it is easily understood by respondents and does not impose high requirements on their cognitive abilities. We also explore the conditions under which this approach provides valid estimates and explore the effects of manipulating these conditions on the estimates.

1.1. Wisdom of Crowds Survey Design

The Wisdom of Crowds design directly asks respondents to state their beliefs about the share of the target population that has a potentially socially undesirable characteristic. The mean of these reported beliefs is the measurement or prediction of interest.[1] The Wisdom of Crowds design is based on the famous observation by Francis Galton (1907) that a crowd's average belief is surprisingly accurate. The data Galton used to observe this fact had been collected in a guessing competition at a cattle fair in 1906 in Plymouth, England. Contestants were asked to guess the amount of meat that could be obtained from a displayed live ox once slaughtered

[1] Originally, Galton (1907) uses the median prediction as estimate. However, social choice theory suggests that the mean prediction has the desired properties (Page 2007). In our empirical application, there is no difference between the mean and the median prediction as we outline below.

and dressed. The crowd’s mean estimate was a mere pound off – better than any individual guess (Galton 1907; Surowiecki 2004).

Even though there are many impressive instances of the accuracy of the Wisdom of Crowds (as do of the Folly of Crowds) we cannot conclude that groups outperform individuals from anecdotal evidence alone. Rigorous research suggests, yet, that a crowd is likely to win a guessing contest even against a group of experts under fairly mild conditions (Sjöberg 2009; Page 2007). This happens because a crowd’s estimates become more exact the more accurate its members estimate individually *and* the more diverse its estimates are. The advantage of accuracy is obvious, the advantage of diversity lies in the fact that errors tend to cancel out.

Surowiecki (2004) summarizes many examples of the Wisdom of Crowds and carves out the conditions under which crowds are wise.[2] He argues that a wise crowd requires, first, *diversity* in opinions and thinking among its members (see also Graefe 2014; Sjöberg 2009). Eventually, diversity in opinions and thinking leads to variation in predictions or guesses. More importantly, however, diversity “helps because it actually adds perspectives that would otherwise be absent and because it takes away, or at least weakens, some of the destructive characteristics of group decision making” (Surowiecki 2004, 36).

Second, individuals within the crowd need to be *independent*, i.e., they must not simply state a certain belief because anybody else does. This characteristic is needed because it ensures that the mistakes individuals make when predicting or estimating a quantity are not positively

[2] There are more technical treatments of these conditions as well. We summarize them in the appendix.

correlated. If they are, individual errors do *not* cancel out and the group's guess is most likely biased (Surowiecki 2004, 52).

Third, individual knowledge should be *decentralized*. To Surowiecki (2004, 88-89) decentralized knowledge is both specialized and local. That is, individuals may not be able to judge all factors that affect the quantity they seek to predict or they have access to a small subset of data for these factors only. Nevertheless, they know something that not everybody else knows as well and hence they have something to contribute to the crowd's prediction.

Fourth, there needs to be a mechanism that aggregates the different pieces of information the individuals within the crowd provide (Surowiecki 2004, 97).

When these four general characteristics, diversity, independence, decentralization and aggregation are combined, crowds are likely to make good predictions. Hence, for the Wisdom of Crowds design to work, these conditions have to be met. Whether these assumptions hold in a given setting is an empirical question – just like it is an empirical question whether a certain item in a questionnaire is a sensitive question and whether respondents are both able to understand and willing to implement the potentially demanding task other sensitive questioning techniques impose.

Below, we show that the Wisdom of Crowds design performs well when a random sample of the German population predicts the vote share of a right-wing populist party whose vote share had been underestimated by polls at the time. Besides the Wisdom of Crowds design, we also test a list experiment and a randomized response technique (RRT) of vote intention. We compare their estimates to both a behavioral benchmark as well as the potentially sensitive vote-intention question. We find that the Wisdom of Crowds design performs best both in terms of accuracy and efficiency.

2. Research Design

We apply the Wisdom of Crowd technique in the context of the German federal election in September 2017 and compare its performance to the RRT, a list experiment, and the potentially sensitive direct vote-intention question. In particular, we seek to test the methods' forecasting ability with respect to the *Alternative for Germany's (AfD)* election result at the 2017 German federal election. Political scientists classify the AfD as a right-wing populist party (e.g., Berbuir et al 2015; Schmidt-Beck 2017). German state elections and the 2013 German federal election showed that, presumably because of social desirability bias, pre-election polls regularly underestimate AfD vote share (Bergmann and Diermeier 2017; Waubert de Puiseau et al. 2017; Buhl 2017).

For the Wisdom of Crowds design to work, the crowd, that is our survey sample, needs to be diverse, decentralized and independent so that the aggregation of individual responses constitutes a valuable estimate of the quantity of interest, i.e., AfD vote share.

At first sight, the use of a random sample of the German population seems to ensure that the crowd meets all criteria. The random sample composition provides respondents with different levels of political sophistication and cognitive abilities which makes it likely that they utilize different sources of private information when guessing the aggregate AfD vote share (diversity). Furthermore, decentralization seems to be guaranteed by respondents' scattering across the country including states with high as well as states with low AfD vote shares.[3] Similarly, random selection should lead to a fair variation in respondents' media consumption adding to their decentralization. Finally, it is highly unlikely that respondents know each other or are influenced by common friends simply because the German population counts about 80 million people of which we survey merely two thousand (independence).

However, even if respondents have decentralized information from consuming different media outlets or interacting with different people, public opinion polls are prominently covered by the German media. While there is some variation in polling companies' results and the exposure different respondents have to them, they potentially provide a very strong signal respondents can rely on when being asked about AfD vote share. To the extent to which respondents are influenced by these polling results, all three factors, their independence, their diversity, and their decentralization decline which threatens the success of the Wisdom of Crowds technique. We consider our empirical test performed here, therefore, as a hard case for the Wisdom of Crowds design because some cards are clearly stacked against its success.

In contrast to previous research on the measurement of sensitive traits, our research design also allows us to compare the methods' predictions[4] to a behavioral benchmark. Even though respondents could not know the share of AfD voters before the election took place, we can

[3] The following website documents current and past opinion polls for all German states and expected state election vote shares: http://www.wahlrecht.de/umfragen/landtage/index.htm.

[4] Please note that by predicting the future rather than a contemporary quantity, we further add to the task's complexity which makes the test even harder.

evaluate a method's performance after the election. Hence, outperforming other methods in predicting AfD vote share would indicate superiority of the Wisdom of Crowds technique.

Below, we detail the survey designs' implementations. All questions were asked in the German Internet Panel (GIP) in July 2017. The GIP is a representative online survey based on an offline sample of the German population (Blom et al. 2015).

2.1 Wisdom of Crowds

The implementation of the Wisdom of Crowds design is straightforward. We simply ask respondents:

> "*What do you think: What percentage share of second votes will the Alternative für Deutschland (AfD) receive in the next Bundestag election this September? The second vote is the vote for a political party.*"

We ask respondents to answer this question by providing an estimate of AfD vote share in the interval from 0 to 100. The mean is the estimate for the final vote share.[5]

[5] Hence, the estimator's variability is the common variability of the mean.

2.2 Direct Question

We also implement the standard question German pollsters ask to learn about respondents' vote intentions. The marginal distribution of this question is used by German polling institutes when preparing press releases for the media. We use this potentially sensitive question as a baseline instrument. Our translation of the standard vote-intention question reads as follows:

> *"If Bundestag elections were held on Sunday, what party would you vote for with your second vote? The second vote is the vote for a political party."*

Respondents could choose one of thirteen answer options that include the set of parties that were represented in the sitting Bundestag (CDU/CSU, SPD, Left Party, the Greens) and parties that gained parliamentary representation in several state parliaments in the years prior to the election (e.g., FDP, AfD, and the Pirates). Additionally, respondents could write down in a separate text field their vote intention for a party not in this choice set. Finally, they could also indicate that they would not vote, were not eligible to vote (e.g., because they are not old enough or do not hold German citizenship), that they do not know which party to vote for, or that they refuse to answer this question.

2.3 Randomized Response Technique

The first indirect questioning technique we consider here is the RRT. We apply the "cross-wise-model" RRT developed by Yu, Tian, and Tang (2008) in order to increase the efficiency of the estimates. Moreover, several studies provide evidence for the validity of this design (e.g., Waubert de Puiseau et al. 2017; Höglinger et al. 2016). This design simultaneously confronts the respondents with two questions with dichotomous answers – one sensitive item of central interest to the researcher and one unrelated, non-sensitive item for which the probability distribution within the population is known. The respondents are asked to evaluate both questions simultaneously and indicate whether or not their responses to both questions are identical (both affirmative or both negative). Since the respondents only indicate whether or not the responses to these questions differ and not the answer to any of them, researchers can credibly assure that the individual answers to the sensitive question will remain confidential and only known to the individual respondent.

In order to implement the crosswise-model in the GIP online questionnaire, we first show respondents the following introduction that should prevent any misunderstandings and increase their willingness to comply:

> "*Sometimes, we test new methods in our study. In the following, we will show you two questions simultaneously. Please indicate whether or not your responses to these questions are identical.*
>
> *To begin with, we would like to ask you to think of a friend or a relative whose house number is known to you. Please memorize the house number's first digit and then click on the 'Continue' button."*

In our application, we utilize the house number of a friend or a relative as randomizing device because the "Benford illusion" increases the estimate's efficiency (Diekmann 2012). Note that German house numbers first digits follow the Benford distribution and hence the probability that a respondent memorizes a 1, 2, 3, or a 4 is about 70 %.

After clicking the ‘Continue’ button, respondents are prompted to the RRT question. Once again, we emphasize that they do not need to answer each question individually, but rather to indicate whether or not the answers to both questions are identical. The question text is as follows:

> *“Please indicate whether both answers are identical, either both ‘yes’ or both ‘no,’ or whether they differ, one answer is ‘yes’ while the other is ‘no.’*
>
> - *Are you going to vote for the AfD in the next federal election with your second vote?*
> - *We asked you to think of a friend or relative whose house number is known to you. Is that house number’s first digit a 1, 2, 3 or a 4?”*

Respondents can choose between the two answers: “identical (both answers either ‘yes’ or ‘no’)” or “different (one answer is ‘yes’ while the other is ‘no’)”. In order to avoid order effects, we randomize the order of these two answer categories.

As Yu, Tian, and Tang (2008) show, the proportion of respondents who intend to vote for the AfD can be calculated by:

$$\hat{\pi} = \left(\frac{r}{n} + p - 1\right)/(2p - 1),$$

where $\hat{\pi}$ is the estimated share of AfD voters, *r* represents the number of respondents who report that both answers are identical, *n* is the total number of respondents answering the question, and *p* is the probability that the house number’s first digit is a 1, 2, 3, or a 4 ($p = 0.7$).[6]

[6] While Yu et al. (2008) also specify a formula to evaluate the variability of the estimator, we rely on bootstrapping to obtain estimates for the variability (see below).

Despite the advantages discussed here, the crosswise-model RRT has also some limitations that can hamper its applicability in different contexts. Since this technique differs from the standard questionnaire design, the respondents need to understand the procedure and realize that it preserves the anonymity of their individual responses. Furthermore, they need to comply with the instructions given by the researcher (Waubert de Puiseau et al. 2017; Höglinger et al. 2016; Jann et al. 2012; Krumpal 2012; Coutts and Jann 2011; Yu et al. 2008). Against this background, Coutts and Jann (2011) conclude that the list experiment, to which we turn next, is a superior alternative to the RRT.

2.4 List Experiment

The classical version of the list experiment, also known as the "Item Count Technique" (ICT) (Holbrook and Krosnick 2010a) or the "Unmatched Count Technique" (UCT) (Coutts and Jann 2011), randomly splits the respondents into two experimental groups. The respondents in both groups receive a list consisting of different non-sensitive elements. The only difference between the two groups is that one group, the treatment group, receives an additional, sensitive list element. Hence, the number of elements in both lists differs. In the following, respondents are asked to indicate *how many* (and not *which*) elements on the list apply to them. For the respondents, it is easy to see that the researcher cannot infer about individual responses based on the answers. Hence, they should be more willing to report socially undesirable characteristics as compared to a direct question (Coutts and Jann 2011; Streb et al. 2008). Another advantage of this technique is that the analysis of the results is fairly simple. The estimate for the share of respondents having the sensitive characteristic is simply the difference in means reported by the two experimental groups (e.g., Rosenfeld et al. 2016; Glynn 2013).

Here, we implement a double list experiment in our representative online survey to address efficiency limitations (Glynn 2013; Droitcour et al. 1991). We randomly assign respondents to two different experimental groups. In contrast to the classical list experiment, however, we develop two different lists, List A and List B, containing four non-sensitive items each. The respondents in both groups are then asked to consecutively evaluate both lists and indicate how many elements apply to them. While the first group receives List A without the sensitive item, the second group evaluates List A with the additional sensitive item. In the following, the first group gets List B with the sensitive item while the second group evaluates List B without the additional idem. Consequently, both groups simultaneously serve as control and treatment group. This procedure increases the estimator's efficiency since, as opposed to the classical implementation of the list experiment, all respondents receive the treatment (Coutts and Jann 2011; Droitcour et al. 1991). Notwithstanding this gain in efficiency, the noise resulting from the aggregation of multiple responses remains.

Since we conduct the double list experiment approximately three months before the federal election in Germany, we ask the respondents how many items on each list they probably will do within the next three months.[7] Table 1 shows the different elements of Lists A and B. n each of these lists, the fourth element represents the sensitive characteristic, which is of interest here. In order to further increase the estimate's efficiency, we design the lists in a way that the correlation between the lists is positive, whereas the correlation of some items within a list is negative (Glynn 2013). We also address possible ceiling and floor effects by including

[7] The precise wording of the question is: *"How many of the following things are you probably going to do within the next three months?"*

items that almost everybody answers affirmatively and items that almost no one answers affirmatively (e.g., Johann et al. 2016; Rosenfeld et al. 2016; Glynn 2013; Kuklinski et al. 1997).

Table 1: Lists of Items for the Double List Experiment

	List A	List B
1	look at an election poster more closely	talk to friends or relatives about politics
2	watch the weather forecast on TV	watch the news on TV
3	participate in a protest march	engage in voluntary work
4	*vote for the Alternative für Deutschland (AfD) in the next federal election*	*vote for the Alternative für Deutschland (AfD) in the next federal election*
5	read the party manifestos of all parties represented in the German Bundestag	run as a candidate for political office

Following Droitcour et al. (1991, 189), the AfD vote share in the population can be calculated by

$$\hat{\pi} = \frac{(\bar{X}_{A5} - \bar{X}_{A4}) + (\bar{X}_{B5} - \bar{X}_{B4})}{2},$$

where $\bar{X}_{A5}$ is the mean response to List A with five items.[8]

[8] Droitcour et al. (1991) also derives the variability of the estimate analytically. In order to ensure the comparability of the methods tested here, we also derive bootstrapped standard errors (see below).

3. Empirical Comparison of the Techniques

3.1 Methodological Notes

We implement the four techniques outlined above in Wave 30 (July 2017) of the GIP and, hence, slightly less than three months before the German federal election took place on September 24, 2017. In order to avoid any distortion of the results caused by repeatedly asking respondents about the same sensitive question, we randomly split the respondents into three equally sized groups (see Table 2).[9] While all respondents receive the potentially sensitive vote-intention question and the Wisdom of Crowds question, only one third of the respondents were allocated to the RRT design. We assign the other two groups to the list experiment where one group receives List A with the treatment and List B without the treatment, and the other List A without treatment and List B with the treatment. Our analytical strategy takes this random allocation of respondents into account.

[9] Table 2 and subsequent analyses include respondents only if they were asked the direct vote-intention question, the Wisdom of Crowds question, and either the RRT question or the list experiment question. 17 respondents answered the direct vote-intention question but dropped out in other, unrelated questions that were asked prior to the sensitive questions discussed here. All results reported remain substantially identical when these 17 respondents are included in the analyses.

Table 2: Summary of the Experimental Groups

Question Type	Respondents	Share of Respondents	Missings	Item Non-Response Rate
Wisdom of Crowds	2597	100.0	32	1.2
Vote Intention	2597	100.0	481	18.5
RRT	867	33.4	94	10.9
List Experiment	1730	66.6	7	0.4
List A	865	33.3	3	0.3
List B	865	33.3	4	0.5

Our first observation concerns the degree of item non-response across the different approaches. The direct question strategy using the standard vote-intention item generates the highest share of missing values. 481 respondents in our study, i.e., almost every fifth respondent who participate in the survey, does not report a valid response to this question. This is consistent with the expectation that direct questioning of potentially sensitive items might lead to a higher item non-response rate (Tourangeau and Yan 2007). Recall that our data stems from a self-administered online survey and hence refused answers are likely to be an even more severe problem in face-to-face surveys (Kreuter et al. 2008). Both the Wisdom of Crowds as well as the list experiment yield low item non-response rates. The RRT provides a non-response rate that lies in between the non-response rates of direct questioning and the two other designs, Wisdom of Crowds and the list experiment. We attribute the RRT's higher item non-response rate to the unfamiliar question format and the increased cognitive burden it places on the respondents (e.g., Holbrook and Krosnick 2010b; Jann et al. 2012).

Below, we aim to compare the three sensitive techniques to the direct vote-intention question in order to investigate whether these methods are well suited to address concerns about social desirability bias. In contrast to most studies that compare the performance of these methods (e.g., Höglinger et al. 2016; Diekmann 2012; Jann et al. 2012; Krumpal 2012; Coutts and Jann 2011), the German federal election provides an actual behavioral benchmark against which the performance of the different techniques can be evaluated. By doing so, we not only investigate how close the estimate derived by each of these techniques comes to the actual AfD vote share on Election Day but also how dispersed the estimates are. Given that the estimates' dispersion decreases with an increase in the sample size, simply comparing the standard errors and the confidence intervals across different techniques is insufficient because the number of respondents differs between the question types.

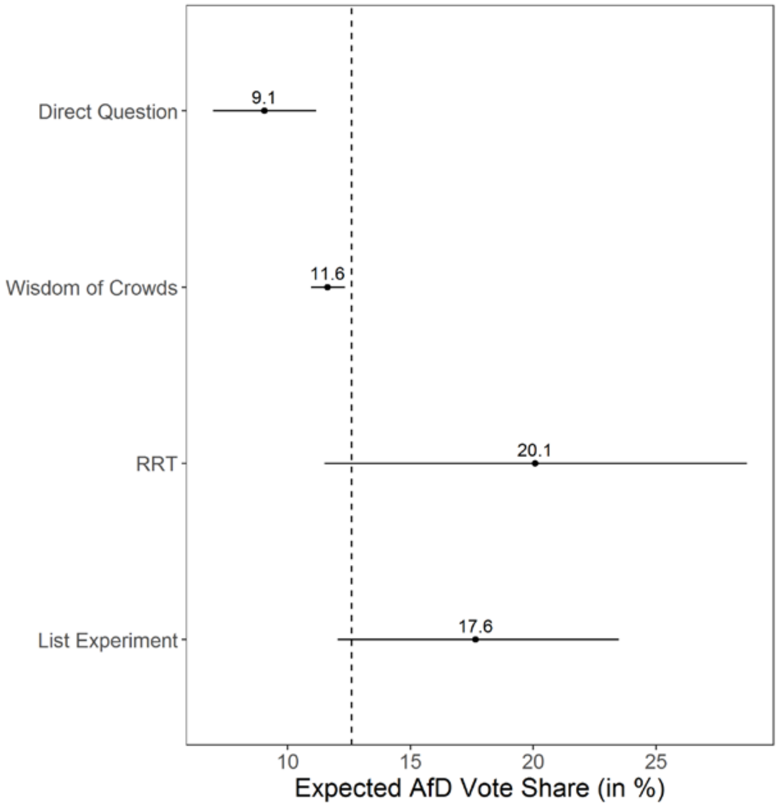
In the following, we therefore apply a bootstrapping algorithm that artificially decreases the number of data points for the direct vote-intention question and the Wisdom of Crowds item. This facilitates a comparison of the estimators' efficiency across the different techniques despite the unequal number of respondents allocated to each of the techniques. To this end, we randomly sample subsets of 865 respondents with replacement from the 2597 respondents and calculate this subsample's expected share of AfD voters. Since we are also interested in studying how non-response and non-compliance affects efficiency, we also include respondents who refused to answer or who are not eligible to vote. We repeat this resampling procedure 1000 times and record the estimated AfD vote share at each iteration. This procedure allows us to compare the techniques with respect to their efficiency despite the varying number of respondents.

3.2 Results

Figure 1 summarizes the results. Points depict estimated AfD vote shares obtained by the different techniques and horizontal lines represent the corresponding bootstrapped 95 % confidence intervals. The vertical line illustrates the actual AfD vote share at the 2017 federal election (12.6 percent) – our behavioral benchmark. Strikingly, the estimates vary across the different techniques. As expected, the direct vote-intention question underestimates AfD vote share by roughly 3.5 percentage points. Put differently, more than one in four AfD voters is not detected by the direct question. The confidence interval covers a range of 4.2 percentage points. Compared to this result, other German polling institutes and forecasting models like *Zweitstimme.org* (Munzert et al. 2017) report an AfD vote share between 7 percent (Allensbach) and 9 percent (Infratest Dimap) at the time.[10]

[10] The website *wahlrecht.de* collects published polls over time (http://www.wahlrecht.de/umfragen/).

Figure 1: Expected AfD Vote Share at the 2017 German Federal Election



Note: Point estimates are depicted by dots and horizontal lines are 95 % bootstrapped confidence intervals. The dashed vertical line represents the actual AfD vote share. Source: GIP Wave 30.

In essence, three reasons might account for the discrepancy between the pre-election polls and the actual election outcome. First, respondents might change their vote choice between the pre-election polling and the actual election. In principle, this can happen in one of two ways: they either change their vote choice to the AfD or from the AfD to another party (or abstain). As the GIP is a panel study, we were able to ask respondents in both September 2017, i.e., in the month the election took place, as well as in November 2017 whether they had voted AfD

(Blom et al. 2018a; Blom et al. 2018b).[11] According to both the September and the November Wave, for about one in twenty respondents the vote intention stated in September does not match their behavior at the election with respect to the AfD. However, the net effect on AfD vote share is virtually balanced in both survey waves indicating a net vote share gain compared to July of mere .55 percentage points in September and .4 percentage points in November. Hence, we are confident that swing voters do not drive the underestimation of AfD vote share by the direct question.

A second reason for the bad performance of the direct questioning is that respondents might perceive it to be socially undesirable to declare the intention to vote for the AfD and therefore do not reveal their true intention. Third, sampling effects might affect the estimated AfD vote share if AfD voters are systematically underrepresented in the sample. Unfortunately, underrepresentation of AfD voters in the sample is untestable unless we can rule out social desirability bias, which, of course, we cannot.

For the implementation of the Wisdom of Crowds technique it does not matter why biases inflict the direct question because it exploits respondents' beliefs rather than their actions or intentions. The second row of Figure 1 indicates that this is an effective strategy as the Wisdom of Crowds design underestimates AfD vote share by only 1 percentage point. At the same time, the narrow 95 % confidence interval covers a range of 1.4 percentage points and is

[11] Respondents who answered the survey in September 2017 before polling stations closed, were asked whether they had voted already. If so, they were asked whom they voted for, otherwise they were asked whom they intended to vote for. We use the available information for each respondent.

significantly smaller than all other techniques' confidence bounds. Yet, it does not cover the actual AfD vote share on Election Day.

The RRT overestimates AfD vote share by 7.5 percentage points, yet, the wide confidence intervals (more than 17 percentage points) cover the true AfD vote share. This indicates an issue with false positives that are triggered by the survey design (Höglinger and Diekmann 2017). Rather similar results are obtained by the double list experiment. It overestimates AfD vote share somewhat less than the RRT (5.4 percentage points overestimation) and has tighter confidence bounds (11.5 percentage points).

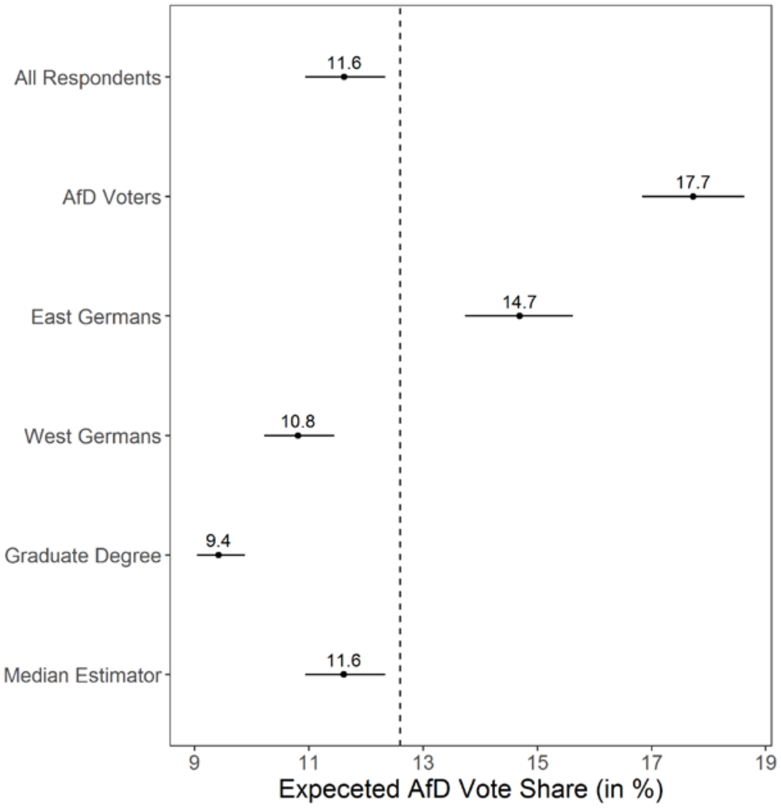
In comparison, all three techniques for sensitive questions outperform the potentially sensitive direct question because they either decrease the amount of bias (Wisdom of Crowds design) or give rise to confidence bounds that cover the true AfD vote share (RRT and double list experiment). However, the RRT's and the double list experiment's confidence bounds are not only their advantage, they are also their disadvantage. On one hand, their high inefficiency that is caused by added noise to ensure privacy allows these techniques to cover the true AfD vote share. On the other hand, due to their inefficiency and the resulting high levels of uncertainty, both techniques are of limited practical utility. Based on the results presented here, we would not be able to discern whether the AfD competes with the rather small parties Greens, the Left, and the FDP for the third position in parliament, whether the AfD competes with the SPD for the second place, or whether it comes close to attacking CDU/CSU's largest party status. Hence, the additional noise induced by the RRT and the list experiment renders their estimates almost useless for inferences on parties' electoral performance. A problem that is likely to inflict many other research subjects as well. Since the Wisdom of Crowds design does not suffer from increased inefficiency, it is a superior alternative if precise estimates are required, given its assumptions are met. We now turn to empirically assessing to what extent its assumptions outlined above can be relaxed.

4. Probing the Wisdom of Crowds Assumptions

Above, we argue that, for the Wisdom of Crowds design to work, individuals within the crowd have to (i) respond independently, (ii) have access to local knowledge, and (iii) think diversely. Finally, an aggregation mechanism has to exist in order to obtain the group belief (Surowiecki 2004). In the following, we alter our research design in different ways to demonstrate what the consequences of relaxing these assumptions are.

First, we limit our sample to respondents that openly state an AfD vote intention in the direct question. On the one hand, we may think of these respondents as experts for AfD vote share because AfD voters hidden to researchers may not be hidden to this subsample that is more outspoken about their vote intentions. On the other hand, AfD voters are not simply a random subsample of the German population as they are more homogenous in terms of many characteristics (e.g., world views, education, media consumption, and so on). This additional homogeneity may prove harmful to their predictive ability because, as argued above, diverse crowds predict often much better than homogenous crowds. Very similar arguments apply to the other subsamples that we filter out. These subsamples include respondents from East Germany (former GDR and Berlin) or West Germany as well as respondents with a graduate degree. The geographic subsamples may be experts because of their local knowledge. East Germans live in those German regions that are most supportive of the AfD. West Germans, on the other hand, significantly outnumber East Germans and may have a better understanding of right-wing populist parties' rise and fall in German parliaments after World War II. University graduates may be experts because they are, on average, more capable of judging and predicting political events. Yet, again, all of these subsamples are also likely to be less diverse than the full sample of respondents.

Figure 2: Expected AfD Vote Share at the 2017 German Federal Election by Subsamples



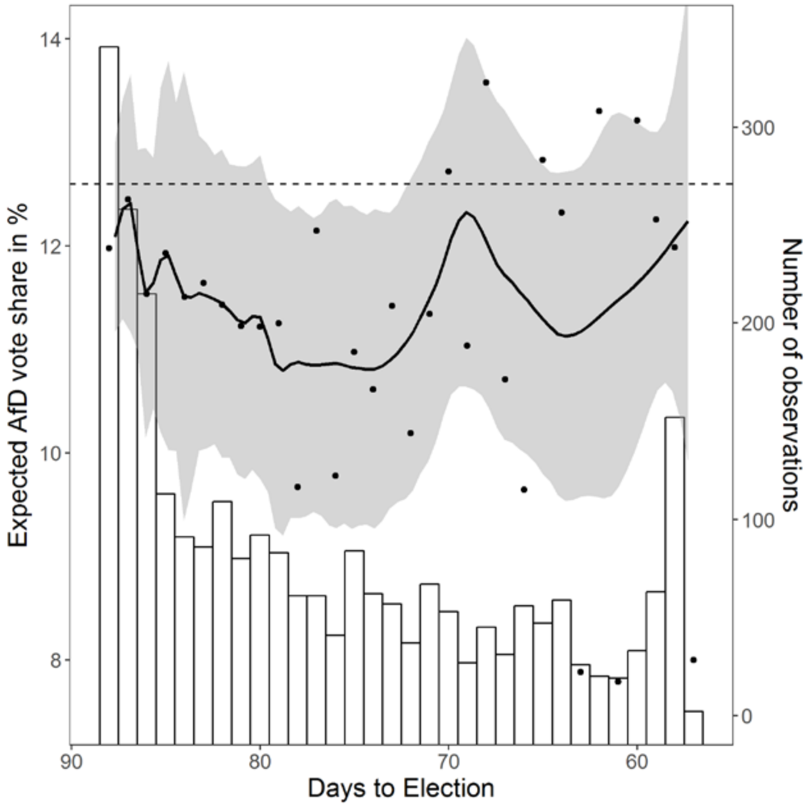
Note: Point estimates are depicted by dots and horizontal lines are 95 % bootstrapped confidence intervals. The dashed vertical line represents the actual AfD vote share. Source: GIP Wave 30.

The results of these subsamples are depicted in Figure 2. To ease comparison, the dashed line indicates the true AfD result at the election and the top estimate replicates all respondents' estimate. All subsample that we consider, AfD voters, East Germans, West Germans and university graduates, make predictions that are further off from the true value than the full sample. This clearly indicates that the full sample's accuracy is not due to potential experts and random errors, however, it rather suggests that the diversity provided by the full sample drives the crowd's wisdom.

Second, we aggregate group beliefs by computing their median rather than their mean. This follows the approach originally developed by Galton (1907). As the bottom row in Figure 2 indicates, there is no difference to the mean estimate. This suggests that other estimators may be superior in some circumstances. Yet, it is beyond this study to further evaluate these conditions.

Third, respondents that state their beliefs about AfD vote share at the same time may also be a more homogenous group because they are primed with similar news coverage. Thus, the more respondents answer on a single day, the less informative and diverse each response might be (Murr 2015). As Figure 3 shows, there is high variance in the mean estimates between days (points). However, the observed variance is mostly due to the varying number of respondents. A more meaningful evaluation of time effects is possible with the Loess estimates indicated by the line because they are based on 15 % of all observations and hence depict the result one would obtain from surveying 390 respondents over a few days. While they cover a range of about 2 percentage points (expected vote shares are roughly between 11 and 12.5 percent), their confidence intervals are always significantly larger than 1.5 percentage points implying that results are statistically indistinguishable from one another. Overall, these results suggest that given a reasonable sample size, the Wisdom of Crowds design works well.[12]

Figure 3: Expected AfD Vote Share at the 2017 German Federal Election by Response Date



Note: The line indicates the loess estimate, the grey area its 95 % confidence interval. Points are means for the corresponding day. Bars depict the number of respondents per day. Source: GIP Wave 30.

[12] In Appendix 2 we further show that differences in methods' performances are not driven by quick responses that are likely uninformative. In Appendix 3, we limit the sample to AfD voters only and show that neither the RRT nor the list experiment recover the "true" AfD vote share in this subsample (100 %) precisely.

5. Conclusion

Survey research is essential to many subfields of the social sciences. However, researchers are ill advised to directly ask respondents sensitive questions because respondents tend to not give any answer at all or untruthful answers (Philips and Clancy 1972; Singer et al. 1995; Singer and Hippler 1992).

To mitigate the problems caused by social desirability bias, we introduce the Wisdom of Crowds design and suggest it as a superior alternative to indirect questioning techniques, especially in situations in which exact predictions and estimates are the aim of inference. As research on the Wisdom of Crowds suggests (Surowiecki 2004; Galton 1907; Lewis-Beck and Skalaban 1989; Lewis-Beck and Tien 1999; Lewis-Beck and Stegmaier 2011; Murr 2011, 2016, 2017), crowds can make good estimates of complicated relationship when their members make independent choices, think diversely, are organized in a decentralized way, and are able to aggregate their information. In a nutshell, the Wisdom of Crowds works because diversity contributes new information to the crowd's information pool and mistakes cancel out across individual beliefs. Based on this finding, we develop the Wisdom of Crowds survey design to learn what share of a target population has a sensitive trait. We suggest to simply asking survey respondents what share of the population has the trait of interest. The crowd's mean belief is the Wisdom of Crowd estimate.

We also compare the Wisdom of Crowds design to both a behavioral benchmark and other survey techniques for sensitive questions (i.e., Randomized Response Technique and list experiment) in the context of the 2017 German federal election with data from the German Internet Panel (Blom et al. 2015). In particular, we predict vote shares for the right-wing populist party *Alternative für Deutschland* (AfD) whose vote share was constantly underestimated by public opinion polls (Bergmann and Diermeier 2017; Waubert de Puiseau et al. 2017). We find that each of the three techniques for sensitive questions outperform the direct questioning either by reducing bias (Wisdom of Crowds) or by giving rise to confidence bounds that cover the true AfD vote share (RRT and list experiment). Yet, the high uncertainty associated with the RRT and the list experiment renders these methods almost useless for an accurate prediction of the vote share. In contrast, the Wisdom of Crowds design performs remarkably well in terms of bias and efficiency which makes this technique a valuable extension to the standard direct vote-intention question in pre-election polls.

The Wisdom of Crowds design's superiority is due to its approach to respondent privacy. While other survey techniques for sensitive questions, i.e., RRT and list experiment, add noise to respondents' answers to ensure privacy, the Wisdom of Crowds design does not pose a personally sensitive question. It rather asks respondents for their belief about the sensitive trait in the population which is not a sensitive question and, hence, does not suffer from social desirability bias. Furthermore, as compared to the RRT and the list experiment, the Wisdom of Crowd design does not impose high cognitive costs on respondents. Since it is easy to analyze and to include in classical surveys as well as cheap in implementation, the results presented here support the suggestion of Graefe (2014) and Murr (2016) that it should be used more frequently in academic and commercial opinion research.

With our additional analyses, we are able to show that violations of the Wisdom of Crowds assumptions, e.g., focusing on subsamples and thereby making the crowd less diverse, renders worse Wisdom of Crowds estimates than the full survey sample. We conclude that the Wisdom of Crowds survey design is a superior survey technique when its assumptions are met.

The results presented here have important implications for research on sensitive traits. Most importantly, the results suggest an additional and – when its assumptions are met – superior survey technique to learn about sensitive traits. Moreover, it is likely that future research will find ways to relax some of the Wisdom of Crowds assumptions. For instance, Gaissmaier and Marewski (2011) show that the Wisdom of Crowds approach works even in non-random samples. Additional research is needed to establish what amount and type of self-selection into the crowd is admissible without endangering a crowd’s wisdom. Moreover, the limits of crowd knowledge should be explored. While we argue that the availability of public opinion polls renders the case at hand a hard case to test the Wisdom of Crowds design, it also ensures that most respondents have some meaningful belief about AfD vote shares. However, we would not expect the same sample to be good at predicting, say, local elections in New Zealand simply because beliefs are most likely to be uninformed. Therefore, future research should seek to determine which topics crowds are competent to judge and which they are not. Finally, the Wisdom of Crowds design may be suited to study populations that are reluctant to participate in surveys or marginal populations if the general population has informative and diverse beliefs about them.

References

Berbuir, Nicole, Marcel Lewandowsky, and Jasmin Siri. 2015. “The AfD and Its Sympathisers: Finally a Right-Wing Populist Movement in Germany?” *German Politics* 24 (2). Routledge:154–78. https://doi.org/10.1080/09644008.2014.982546.

Bergmann, Knut, and Matthias Diermeier. 2017. “Die AfD: Eine Unterschätzte Partei. Soziale Erwünschtheit Als Erklärung Für Fehlerhafte Prognosen.” No. 7/2017. IW-Report.

Blom, Annelies G., Barbara Felderer, Jessica Herzing, Ulrich Krieger, Tobias Rettig, and SFB 884 Political Economy of Reforms, Universität Mannheim. 2019. *German Internet Panel, Wave 31 (September 2017). GESIS Data Archive, Cologne. Forthcoming.*

Blom, Annelies G., Barbara Felderer, Jessica Herzing, Ulrich Krieger, Tobias Rettig, and SFB 884 Political Economy of Reforms, Universität Mannheim. 2018. *German Internet Panel, Wave 30 (July 2017). GESIS Data Archive, Cologne. ZA6904 Data File Version 1.0.0*. https://doi.org/10.4232/1.12977.

Blom, Annelies G., Christina Gathmann, and Ulrich Krieger. 2015. “Setting Up an Online Panel Representative of the General Population: The German Internet Panel.” *Field Methods* 27 (4):391–408. https://doi.org/10.1177/1525822X15574494.

Buhl, Yannick. 2018. “Die Unterschätzten Rechtspopulisten. Wird Die AfD Bei Der Bundestagswahl Stärker, Als Es Umfragen Und Prognosen Vorhersagen?” *CORRELAID.ORG BLOG*, September 23, 2018. https://correlaid.org/blog/posts/ist-die-afd-unterschaetzt.

Chang, Linchiat, and Jon A. Krosnick. 2010. “Comparing Oral Interviewing with Self-Administered Computerized Questionnaires: An Experiment.” *Public Opinion Quarterly* 74 (1):154–67. https://doi.org/10.1093/poq/nfp090.

Coutts, Elisabeth, and Ben Jann. 2011. “Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT).” *Sociological Methods & Research* 40 (1):169–93. https://doi.org/10.1177/0049124110390768.

Crowne, Douglas P., and David Marlowe. 1960. “A New Scale of Social Desirability Independent of Psychopathology.” *Journal of Consulting Psychology* 24 (4):349–54. https://doi.org/10.1037/h0047358.

Diekmann, Andreas. 2012. “Making Use of ‘Benford’s Law’ for the Randomized Response Technique.” *Sociological Methods & Research* 41 (2):325–34. https://doi.org/10.1177/0049124112452525.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 1991. “The Item-Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application.” In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 185–210. Hoboken, NJ, USA: John Wiley & Sons Inc.

Gaissmaier, Wolfgang, and Julian N. Marewski. 2011. “Forecasting Elections with Mere Recognition from Small, Lousy Samples: A Comparison of Collective Recognition, Wisdom of Crowds, and Representative Polls.” *Judgement and Decision Making* 6 (1):73–88.

Galton, Francis. 1907. “Vox Populi.” *Nature* 75:450–51.

Glynn, Adam N. 2013. “What Can We Learn with Statistical Truth Serum?” *Public Opinion Quarterly* 77 (S1):159–72. https://doi.org/10.1093/poq/nfs070.

Graefe, Andreas. 2014. “Accuracy of Vote Expectation Surveys in Forecasting Elections.” *Public Opinion Quarterly* 78 (S1):204–32. https://doi.org/10.1093/poq/nfu008.

Höglinger, Marc, and Andreas Diekmann. 2017. “Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT.” *Political Analysis* 25 (1):131–37. https://doi.org/10.1017/pan.2016.5.

Höglinger, Marc, Ben Jann, and Andreas Diekmann. 2016. “Sensitive Questions in Online Surveys: An Experimental Comparison of the Randomized Response Technique and the Crosswise Model.” *Survey Research Methods* 10 (3):171–87. https://doi.org/10.18148/srm/2016.v10i3.6703.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. “Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling Into Question the Method’s Validity.” *Public Opinion Quarterly* 74 (2):328–43. https://doi.org/10.1093/poq/nfq012.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. “Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique.” *Public Opinion Quarterly* 74 (1):37–67. https://doi.org/10.1093/poq/nfp065.

Hong, Lu, and Scott Page. 2009. “Interpreted and Generated Signals.” *Journal of Economic Theory* 144 (5):2174–96. https://doi.org/10.1016/j.jet.2009.01.006.

Jann, Ben, Julia Jerke, and Ivar Krumpal. 2012. “Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism.” *Public Opinion Quarterly* 76 (1):32–49. https://doi.org/10.1093/poq/nfr036.

Johann, David, Kathrin Thomas, Thorsten Faas, and Sebastian Fietkau. 2016. “Alternative Messverfahren Rechtspopulistischen Wählens Im Vergleich: Empirische Erkenntnisse Aus Deutschland Und Österreich.” In *Wahlen Und Wähler: Analysen Aus Anlass Der Bundestagswahl 2013*, edited by Harald Schoen and Bernhard Weßels, 447–70. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-11206-6_20.

Krause, Stefan, Richard James, Jolyon J. Faria, Graeme D. Ruxton, and Jens Krause. 2011. “Swarm Intelligence in Humans: Diversity Can Trump Ability.” *Animal Behaviour* 81 (5):941–48. https://doi.org/10.1016/j.anbehav.2010.12.018.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” *Public Opinion Quarterly* 72 (5):847–65. https://doi.org/10.1093/poq/nfn063.

Krogh, Anders, and Jesper Vedelsby. 1994. “Neural Network Ensembles, Cross Validation and Active Learning.” In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, 231–38. NIPS’94. Cambridge, MA, USA: MIT Press. https://doi.org/10.1.1.37.8876.

Krosnick, Jon A. 1999. “Survey Research.” *Annual Review of Psychology* 50 (1):537–67. https://doi.org/10.1146/annurev.psych.50.1.537.

Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3):213–36. https://doi.org/10.1002/acp.2350050305.

Krumpal, Ivar. 2012. “Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning.” *Social Science Research* 41 (6):1387–1403. https://doi.org/10.1016/j.ssresearch.2012.05.015.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. “Racial Prejudice and Attitudes Toward Affirmative Action.” *American Journal of Political Science*. https://doi.org/10.2307/2111770.

Lewis-Beck, Michael S., and Andrew Skalaban. 1989. “Citizen Forecasting: Can Voters See into the Future?” *British Journal of Political Science* 19 (1):146–53. https://doi.org/10.1017/S000712340000538X.

Lewis-Beck, Michael S., and Mary Stegmaier. 2011. “Citizen Forecasting: Can UK Voters See the Future?” *Electoral Studies* 30 (2):264–68. https://doi.org/10.1016/j.electstud.2010.09.012.

Lewis-Beck, Michael S., and Charles Tien. 1999. “Voters as Forecasters: A Micromodel of Election Prediction.” *International Journal of Forecasting* 15 (2):175–84. https://doi.org/10.1016/S0169-2070(98)00063-6.

Munzert, Simon, Lukas Stötzer, Thomas Gschwend, Marcel Neunhoeffer, and Sebastian Sternberg. 2017. “Zweitstimme.org. Ein Strukturell-Dynamisches Vorhersagemodell Für

Bundestagswahlen." *Politische Vierteljahresschrift* 58 (3):418–41. https://doi.org/10.5771/0032-3470-2017-3-418.

Murr, Andreas E. 2017. "Wisdom of Crowds." In *Handbook of Political Behavior,* edited by Kai Arzheimer, Jocelyn Evans, and Michael S. Lewis-Beck, 835–60. Los Angeles: Sage.

Murr, Andreas E. 2016. "The Wisdom of Crowds: What Do Citizens Forecast for the 2015 British General Election?" *Electoral Studies* 41:283–88. https://doi.org/10.1016/j.electstud.2015.11.018.

Murr, Andreas E. 2015. "The Wisdom of Crowds: Applying Condorcet's Jury Theorem to Forecasting US Presidential Elections." *International Journal of Forecasting* 31 (3):916–29. https://doi.org/10.1016/j.ijforecast.2014.12.002.

Murr, Andreas E. 2011. "'Wisdom of Crowds' ? A Decentralised Election Forecasting Model That Uses Citizens' Local Expectations." *Electoral Studies* 30 (4):771–83. https://doi.org/10.1016/j.electstud.2011.07.005.

Page, Scott E. 2014. "Where Diversity Comes from and Why It Matters?" *European Journal of Social Psychology* 44 (4):267–79. https://doi.org/10.1002/ejsp.2016.

Page, Scott E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ: Princeton University Press.

Phillips, Derek L., and Kevin J. Clancy. 1972. "Some Effects of 'Social Desirability' in Survey Studies." *American Journal of Sociology* 77 (5):921–40. https://doi.org/10.1086/225231.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2016. “An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions.” *American Journal of Political Science* 60 (3):783–802. https://doi.org/10.1111/ajps.12205.

Schmitt-Beck, Rüdiger. 2017. “The ‘Alternative Für Deutschland in the Electorate’: Between Single-Issue and Right-Wing Populist Party.” *German Politics* 26 (1):124–48. https://doi.org/10.1080/09644008.2016.1184650.

Singer, Eleanor, Hans Jürgen Hippler, and Norbert Schwarz. 1992. “Confidentiality Assurances in Surveys: Reassurance or Threat?” *International Journal of Public Opinion Research* 4 (3):256–68. https://doi.org/10.1093/ijpor/4.3.256.

Singer, Eleanor, Dawn R. von Thurn, and Esther R. Miller. 1995. “Confidentiality Assurances and Response Experimental Literature.” *Public Opinion Quarterly* 59 (February):66–77.

Sjöberg, Lennart. 2009. “Are All Crowds Equally Wise? A Comparison of Political Election Forecasts by Experts and the Public.” *Journal of Forecasting* 28 (1):1–18. https://doi.org/10.1002/for.1083.

Streb, Matthew J., Barbara Burrell, Brian Frederick, and Michael A. Genovese. 2008. “Social Desirability Effects and Support for a Female American President.” *Public Opinion Quarterly* 72 (1):76–89. https://doi.org/10.1093/poq/nfm035.

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday.

Tourangeau, Roger, Lance J. Rips, and Kenneth A. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133 (5):859–83. https://doi.org/10.1037/0033-2909.133.5.859.

Waubert de Puiseau, Berenike, Adrian Hoffmann, and Jochen Musch. 2017. “How Indirect Questioning Techniques May Promote Democracy: A Preelection Polling Experiment.” *Basic and Applied Social Psychology* 39 (4). Routledge:209–17. https://doi.org/10.1080/01973533.2017.1331351.

Appendix 1: Prediction Accuracy, Prediction Diversity and the Wisdom of Crowds

Under what conditions do crowds appear wise? As social choice theory shows, the wisdom of crowds works because diversity is a substitute for expertise (Page 2007). Even more, Sjöberg (2009) finds empirically that the aggregate prediction of non-experts outperforms experts' aggregate prediction although the experts were more accurate than less informed and less interested non-experts. Graefe (2014, 213) explains this surprising finding by the groups' heterogeneity. While the expert group varied less in their demographics, the non-expert group exhibits a high diversity among its members. Consequently, it is likely that the members of the expert group were biased in the same direction. Since the individual answers are highly correlated, their biases do not cancel out each other when aggregated.

Consider the following example. A researcher seeks to understand a phenomenon (e.g., the share of a population that holds a particular attitude) that is determined by an array of factors such that $y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n$, where y is the phenomenon of interest, x_i is the i^th factor that co-determines y, β_i is the effect a unit change in x_i has on y, and α is a constant effect.[13] Due to the high complexity of social reality the number of factors that co-determine y is large. This complexity makes social scientists adopt theoretical and statistical models that are simplifications of the world and that willfully ignore certain factors. Put differently, even the most sophisticated experts are most unlikely to be able to know of all determining factors, leaving alone having access to sufficient data to gauge the values of all β_i's (Page 2007).

[13] Of course, interactions of determining factors are captured in this framework as well.

By information aggregation, however, a crowd of laypeople can easily give rise to a far more sophisticated model of reality than experts use – provided the crowd has diverse views on reality (Page 2007). For this mechanism to work, laypeople in our formal example have to consider different determining factors, even if every individual layperson would consider one or two factors only. When each layperson then states their estimate of y, it is highly likely that these are negatively correlated to each other, that is when some laypeople overestimates the influence of a particular determining factor (and hence, say, overestimate y) others systematically underestimate it (Hong and Page 2009). By information aggregation, i.e., aggregation of individual estimates, these errors cancel out and the crowd's model is more accurate than most individual estimates (e.g. Graefe 2014).

In fact, the "Diversity Prediction Theorem" (Page 2007, see also Krogh and Vedelsby 1994) shows that estimation accuracy and diversity are equally important. It reads:[14]

$$\text{Collective Error} = \text{Average Individual Error} - \text{Estimation Diversity}$$

where Collective Error is the squared distance between the crowd's mean estimation and the true outcome; Average Individual Error is the mean squared distance between each individual's estimation and the true outcome; and Estimation Diversity is the mean squared distance between each individual's estimate and the mean prediction, i.e., the variance of individual estimates around the crowd's prediction.

[14] Page (2007) states this theorem in terms of crowd predictions instead of estimates. These are, however, equivalent since both deal with the expression of beliefs about an unknown quantity.

Note that the crowd's estimate becomes a unit better if either individuals make on average a unit less errors in estimation (and the variance of their estimates remains unchanged) *or* the variance of their estimates increases by a unit while individual accuracy is unchanged. Put differently, even though many laypeople may give responses that are, from a researcher's point of view, very inaccurate, they can yield a highly precise estimate when aggregated provided the group is sufficiently diverse such that their underlying implicit forecasting models differ sufficiently (Page 2014). When crowds are diverse, they are likely to be wise too.

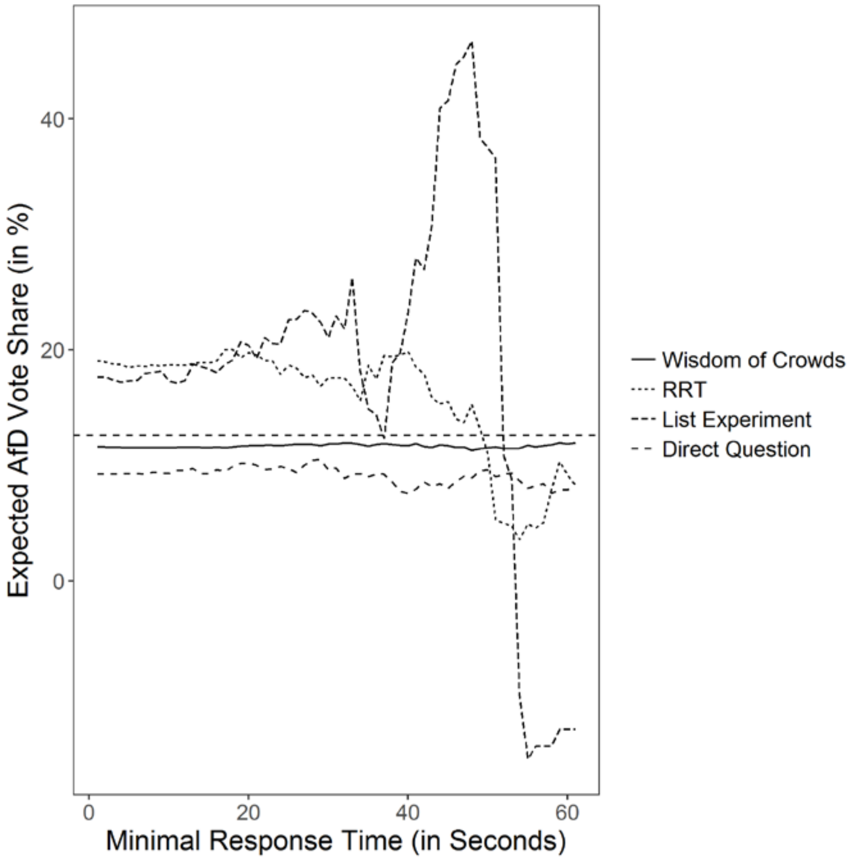
Appendix 2: Cognitive Demand on Respondents

Both techniques, the list experiment and the RRT, are cognitively more demanding as compared to the direct vote intention question and the Wisdom of Crowds technique since the respondents need to evaluate multiple questions and aggregate the individual responses. It is therefore possible that respondents do not understand or comply with the instructions and simply try to quickly finish the questionnaire (e.g., Krosnick 1991, 1999). Especially the RRT design seems to be prone to problems of noncompliance (e.g., Coutts and Jann 2011). In this context, it is possible that quick responses do not carry the same amount of information as slower responses. In order to address potential problems of noncompliance or misunderstandings, we test the sensitivity of the results to the sequential exclusion of quick responses.

Figure A1 shows how the estimate obtained by the different techniques tested here change if we exclude responses that are below a certain threshold depicted on the horizontal axis. At the left point of the graph, no observation is excluded and the estimates resemble the ones shown in Figure 1. By moving further to the right on the x-axis, more and more observations drop out of the estimation.

It is easy to see that considerable variation in the estimates occurs once one consecutively excludes more and more observations. This especially holds for the estimate derived by the list experiment and the RRT. The estimates obtained by the direct vote-intention question and the Wisdom of Crowds technique, however, are very robust to the sequential removal of responses. Notable changes only occur after about 30 seconds for the direct question since the size of the data set decreases to less than one fourth of its original size. A similar decrease in sample size can be observed for the Wisdom of Crowds design. Yet, despite this decrease, the estimate remains very stable and comes closer to the actual election outcome than any other techniques. From this we conclude that the differences between the techniques do not change once quick responses are removed from the dataset.

Figure A1: Expected AfD Vote Share at the 2017 German Federal Election per Response Time



Note: The black lines represent the different techniques' point estimates and the gray line shows the actual AfD vote share. Source: GIP Wave 30.

Appendix 3: Validation Based on Self-Identified AfD Voters

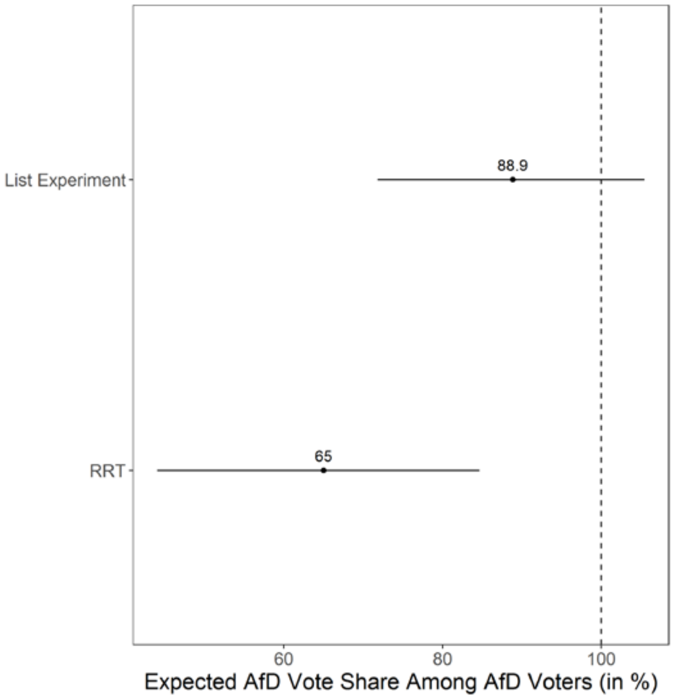
We also assess the validity of the results by only regarding respondents who indicate that they want to vote for the AfD in the direct vote-intention question. Under the assumption that respondents who indicate their willingness to vote for the AfD in the direct vote-intention question will also declare to vote for the AfD in the list experiment and the RRT design where the anonymity of their responses is assured, we expect the estimates for this subsample to 100 %.[15] Except for the fact that we select the cases based on the dependent variable, the analysis remains unchanged. Figure A2 presents the results.

Although the double list experiment underestimates the AfD vote share by 11 percentage points, its wide confidence interval, which covers a range of 34 percentage points, also includes the theoretically expected AfD vote share of 100 % among self-identified AfD voters. The point estimate of the RRT design is 35 percentage points below the 100 % benchmark and the upper bound of the associated confidence interval is 15 percentage points below the expected value of 100 %. Hence, our analysis confirms prior findings that the RRT design is vulnerable to – intentional or unintentional – compliance problems on the side of the respondents which raise concerns about the findings' validity (e.g., Höglinger et al. 2016; Holbrook and Krosnick 2010b). These results illustrate that although both techniques adjust for social desirability bias in pre-election polling they come with numerous problems and difficulties like estimates inefficiency, increased cognitive demands on respondents, and problems of noncompliance. Thus, the addition of noise to the signal, although a theoretically appealing

[15] The analysis assumes that we can measure vote intention without measurement error. Yet, it is also reasonable that some respondents are unsure and therefore give inconsistent answers about their vote intention.

approach, does not seem to be particularly useful for its practical application in learning about sensitive traits.

Figure A2: Expected AfD Vote Share at the 2017 German Federal Election of Self-Identified AfD Voters



Note: Point estimates are depicted by the dots while the horizontal lines are the 95 % bootstrapped confidence intervals. The dashed vertical line represents the theoretically expected AfD vote share. Source: GIP Wave 30.