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, sebastian.juhl@gess.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 indirect questioning techniques address the problem of social desirability bias at the expense of decreasing estimates' efficiency. We suggest the Wisdom of Crowds survey design that does not pose a tradeoff between anonymity and efficiency as an alternative. We outline the conditions necessary for the technique to work and test them empirically. Moreover, we compare the Wisdom of Crowd estimate of a right-wing populist party's vote share to alternative indirect questioning techniques' estimates as well as to the official election result in the 2017 German federal election. Provided its conditions are met, the Wisdom of Crowds design performs best in terms of both bias and efficiency. We conclude that the Wisdom of Crowds design is an important addition to social scientists' survey methodology toolbox.

Social desirability bias is a type of measurement error that arises when survey respondents either 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 activities or holding racist sentiments (Philips and Clancy 1972). In the social science literature, it is widely acknowledged that social desirability bias afflicts survey research whenever respondents are asked about 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, like different indirect questioning techniques, trade-off systematic measurement error (bias) for increased variability of measurements (efficiency). They do so by adding noise to respondents' answers which prohibits researchers from learning whether or not an individual respondent has the sensitive characteristic. Since researchers know the distribution of noise across all respondents they can still obtain an unbiased measurement of the sensitive characteristic's prevalence. By design, the added noise, however, makes the measurements subject to additional variation resulting in wider confidence bounds around the (unbiased) measurement (Höglinger et. al 2016, Höglinger and Diekmann 2017).

In this article, 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 the prevalence of sensitive traits

than alternative approaches. The underlying concept is that we suggest asking respondents what they believe the share of the population is that has the sensitive characteristic of interest. Clearly, as asking about a respondent's believe of the share is not in itself a sensitive question, respondents should be more likely to answer truthfully. As Surowiecki (2004, see also Galton 1907; Page 2007) shows, crowds can be wise and their mean belief constitutes an unbiased estimate of reality when the crowd members are decentralized, independent, diverse thinkers, and have access to an aggregation mechanism.

The Wisdom of Crowds survey design is used as an approach to election forecasting with considerable success (Graefe 2014). Our contribution is to explicitly and empirically investigate its performance once we successively relax the underlying assumptions. Furthermore, we compare its bias and efficiency to a double list experiment, a crosswise-model randomized response technique, and a direct question. Our results imply that researchers who can credibly claim that the Wisdom of Crowd assumptions are met can obtain unbiased and more efficient estimates from the Wisdom of Crowd design than from a list experiments, a randomized response design, or a direct question. Such efficiency gains become 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.

1. Sensitive Survey Questions and the Bias-Efficiency Tradeoff

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 unless the researcher knows the direction and the size of the bias, and corrects the estimates accordingly.

In survey research, it has long been recognized that, depending on the survey item, respondents intentionally answer some questions incorrectly, i.e., they send 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 even affects self-administered web surveys (Chang and Krosnick 2010; Kreuter et al. 2008). Accordingly, measuring the prevalence of sensitive traits, opinions, or attitudes in a target population by surveys constitutes a methodological challenge. Researchers not only need to develop instruments that maximize the signal-to-noise ratio, they also need to ensure that the instrument itself does not systematically distort the signal.

When estimating the aggregate share of a target population that has a potentially socially undesirable characteristic, previous research suggests the application of indirect questioning

techniques which assure anonymity to individual answers in order to elicit honest responses (e.g., Waubert de Puiseau et al. 2017; Thomas et al. 2017; Rosenfeld et al. 2016). 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.

By design, these indirect questioning techniques do not allow researchers to infer individual behavior from survey responses. Only the prevalence of a certain characteristic can be estimated. While the absence of individual-level data certainly constitutes a limitation in some research contexts, the ability to obtain unbiased estimates of the sensitive characteristic's prevalence is of great value in several applied settings. Researchers need to gauge carefully whether the characteristic they aim to study might be affected by social desirability concerns and whether they are willing to forgo the opportunity to obtain possibly biased individual-level data for the opportunity to get an unbiased estimate of the characteristic's prevalence in the target population by employing indirect questioning techniques. Several studies present empirical evidence for the ability of these indirect questioning techniques to recover a higher share of respondents that hold the sensitive characteristic as compared to 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 their own problems that researchers would like to avoid. More precisely, indirect questioning techniques are by design statistically less efficient than direct questions because they trade off bias for efficiency (e.g., Höglinger et al. 2016; Höglinger and Diekmann 2017). 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 design (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 poses a direct question, yet, at the same time, it 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, empirically investigate the effects of manipulating these conditions on the estimates, and compare its performance to other questioning techniques.

2. The Wisdom of Crowds Survey Design and its Conditions

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), anecdotal evidence alone is certainly insufficient for concluding that groups usually outperform individuals. Yet, rigorous research also finds that a crowd is likely to win a guessing contest even against a group of specialists under fairly mild conditions (Sjöberg 2009; Page 2007). This happens because a crowd's estimate becomes more exact the more accurate its members estimate individually *and* the more variable individual estimates are. While the advantage of accuracy is obvious, the advantage of variability lies in the fact that individual errors tend to cancel out.

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:[2]

$$\textit{Collective Error} = \textit{Average Individual Error} - \textit{Estimation Variance},$$

where *Collective Error* is the squared distance between the crowd's mean estimate and the true outcome; *Average Individual Error* is the mean squared distance between each individu-

[2] We change Page's (2007) notation slightly to ensure consistency and ease understanding of the description below. First, Page (2007) states the 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. Second, we substitute "Diversity" for "Variance" because we use "diversity" below in a slightly different context.

al's estimate and the true outcome; and *Estimation Variance* is the mean squared distance between each individual's estimate and the aggregate estimate, i.e., the variance of individual estimates around the crowd's estimate.

Note that the crowd's estimate improves by a unit if either individuals make on average a unit less mistakes in estimating the quantity of interest while holding the variance of their estimates constant *or* the variance of their estimates increases by a unit while individual accuracy is unchanged. Put differently, if a researcher seeks to improve her crowd's estimate by a unit, she can recruit experts to the crowd (or dismiss bad predictors) such that the *Average Individual Error* decreases by a unit holding *Estimation Variance* constant. Since the researcher does not know the true value of interest, she cannot know individual errors, and will have a very hard time including or excluding the right people to improve the *Average Individual Error*. Alternatively, she could decrease the *Collective Error* by a unit by increasing the *Estimation Variance* by one unit holding the *Average Individual Error* constant. Practically, this is a much more feasible solution since it certainly easier to identify strata and views not included in the crowd yet rather than knowing the true outcome when this is exactly what one seeks to learn. Eventually, the greater a crowd's estimation variance at a given level of *Average Individual Error*, the more precise its estimate.

Surowiecki (2004) summarizes many examples of the Wisdom of Crowds and carves out the conditions under which crowds are wise. 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 greater variation in estimates 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 char-

acteristics of group decision making" (Surowiecki 2004, 36). With respect to the Diversity Prediction Theorem, diversity provides estimation variance that decreases the collective error.

Second, 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 only have access to a small subset of information on these factors. Nevertheless, they know something that not everybody else knows as well and hence they have something to contribute to the crowd's estimate. [3] Overall, decentralization allows for estimation variance that reduces collective error.

Third, individuals within the crowd need to be *independent*, i.e., they must not simply state a certain belief because somebody else does. This condition is needed because it ensures that the mistakes individuals make when predicting or estimating a quantity are not positively correlated. If they are, individual errors do *not* cancel out and the group's guess is most likely biased, i.e., the average individual error is high (Surowiecki 2004, 52).

Fourth, there needs to be a mechanism that *aggregates* the different pieces of information individuals within the crowd provide (Surowiecki 2004, 97). If no such mechanism exists, researchers cannot obtain a crowd estimate.

When these four general conditions, diversity, decentralization, independence, and aggregation are combined, crowds provide good estimates. That is, the estimation variance they pro-

[3] There is a more technical treatment of this condition which we briefly summarize in Appendix 1.

vide is likely to be sufficiently high relative to individual estimation errors. Whether this is the case is, ultimately, an empirical question that cannot be answered *a priori* because the true prevalence is unknown. This, however, is not an uncommon situation in survey research. It is also an *a priori* unsolved 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. Similarly, researchers often include knowledge questions in surveys to discriminate between highly and poorly sophisticated respondents. In this context, it is also an *a priori* unsolved empirical puzzle which knowledge questions provide good discrimination between different types of respondents. In all of these situation, researchers turn to contextual knowledge to motivate their decisions. We suggest that researchers use such knowledge as well to justify that the estimation variance in their sample is sufficiently high (or that individual errors are sufficiently low) to render the Wisdom of Crowds design feasible.

To sum up, we argue that, when researchers have contextual knowledge that suggest that their sample at hand is sufficiently diverse, decentralized, and independent, they can obtain unbiased and efficient estimates from the Wisdom of Crowds design. Further, we argue that, when a sample does not have these characteristics, its Wisdom of Crowds estimate is likely biased. Before testing these expectations empirically, we point out that evaluating bias – an absolute quantity – implies comparing the estimate to a true value (benchmark), while judging efficiency – a relative quantity – implies comparing different estimation strategies to one another. We, thus, first turn to the relationship between the Wisdom of Crowds assumptions and estimation bias, before comparing the Wisdom of Crowds design to other questioning techniques for sensitive questions.

3. Wisdom of Crowds Assumptions and Estimation Bias

3.1. Research Design

We apply the Wisdom of Crowd technique in the context of the German federal election in September 2017 and evaluate its performance by comparing its estimate to the *Alternative for Germany's (AfD)* official election result. After the party was formed as an Eurosceptic party in 2013, it increasingly developed into a radical right-wing populist party with strong anti-immigration stances (e.g., Berbuir et al 2015; Schmidt-Beck 2017). After several internal fights that caused major personnel changes as well as numerous scandals, the AfD successively moved significantly closer to the right margin of the ideological spectrum. As a consequence, Bergmann and Diermeier (2017) argue that German voters are reluctant to openly support the party because this is considered to be socially undesirable. In line with this reasoning, German state elections and the election of the European parliament in May 2014 showed that, presumably because of social desirability bias, pre-election polls that rely on the classical direct vote-intention question regularly underestimate AfD vote share (Bergmann and Diermeier 2017; Waubert de Puiseau et al. 2017; Buhl 2018). While we focus on a German case, we note in the context of rising populism (Inglehart and Norris 2016) that we consider one of many cases in which one vote alternative may be perceived as somewhat frowned upon.

We implemented the Wisdom of Crowds survey design in the German Internet Panel (GIP) in July 2017. The GIP is an online panel survey whose respondents were recruited in an offline face-to-face recruitment procedure. The probability-based offline recruitment process ensures that the sample is representative of the German population aged 16-75. In order to transform the offline recruited sample into a representative online sample, one unique feature of the GIP

is that it equips households without a computer or access to the internet with the necessary devices so that they can participate online. Hence, although the GIP is a self-administered online panel, the offline recruiting procedure of both online and offline households as well as the technical support offered by the GIP provides a sample that can be considered representative of the German target population between 16 and 75 year of age who live in private households (Blom et al. 2015).

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 for Germany (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.[4]

As discussed above, 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.

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

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.[5] Similarly, random selection should lead to a fair variation in respondents' media consumption adding to their decentralization. Moreover, it is highly unlikely that respondents know each other or are influenced by common friends simply because the German population counts more than 80 million people of which we survey merely about two and a half thousand (independence). Finally, the Wisdom of Crowds design has been shown to work well in the context of election forecasting (Lewis-Beck and Skalaban 1989; Lewis-Beck and Tien 1999; Lewis-Beck and Stegmaier 2011; Murr 2011, 2016, 2017). Overall, there are many reasons to belief that the Wisdom of Crowds design should perform well in this context.

There is, however, a point that raises doubt on our sample's ability to meet the Wisdom of Crowds assumptions. Even if respondents have decentralized information from consuming different media outlets or interacting with different people, German media prominently cover public opinion polls. 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 to predict AfD vote share. To the extent to which

[5] 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.

respondents are influenced by these polling results, three factors, their independence, their diversity, and their decentralization decline which threatens the success of the Wisdom of Crowds technique. Therefore, we consider our empirical test performed here 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 further allows us to compare the estimates to a behavioral benchmark.[6] Unlike other research contexts (but see Rosenfeld et al. 2016) in which researchers do not know the aggregate share of the population that has the sensitive characteristic, the 2017 German federal election reveals this quantity. Therefore, our validation strategy does not rely on the problematic "more-is-better assumption" (Höglinger and Diekmann 2017, 132), i.e., higher estimates of the sensitive characteristic indicate higher validity. Instead, we validate the Wisdom of Crowd design here by comparing its estimate to the official AfD vote share on Election Day.

To probe the Wisdom of Crowd assumptions, we, first, show that the Wisdom of Crowds design performs well when we use the full sample at hand. We, then, limit our analyzes to specific subsamples that, in ways which we specify below, violate the Wisdom of Crowd assumptions.

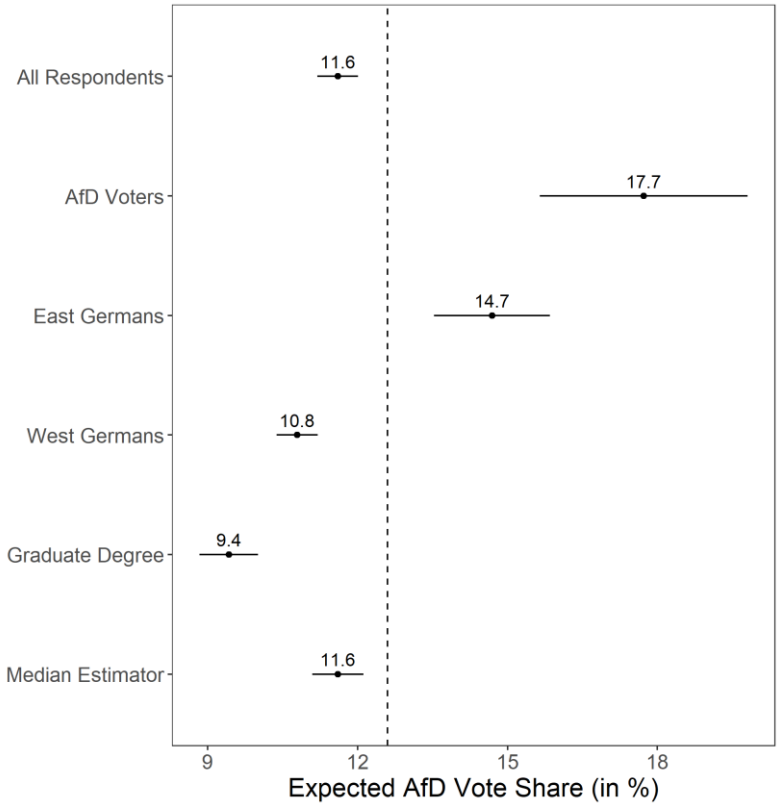
[6] Note that by predicting the future rather than a contemporary quantity, we further add to the task's complexity which makes the test even harder.

3.2 Consequences of Violating the Wisdom of Crowds Assumptions

In this section we probe the Wisdom of Crowd assumptions by comparing the estimates of the full sample with respective estimates of particular subsamples. Figure 1 summarizes the results. Points depict estimated AfD vote shares obtained from different subsamples and horizontal lines represent the corresponding 95% confidence intervals. The vertical line illustrates the actual AfD vote share at the 2017 federal election (12.6%) – our behavioral benchmark.

The first row of Figure 1 indicates that the Wisdom of Crowds design underestimates AfD vote share by only 1 percentage point when all respondents who answer the Wisdom of Crowd question are included in the crowd (bias). Despite the caveats that respondents may be influenced by the same polls or media, their aggregated belief is a very precise prediction of AfD vote share. In fact, when comparing the Wisdom of Crowd design estimate to other techniques' estimates (see below), we show that the Wisdom of Crowds design provides the most accurate and most efficient estimate.

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



Note: Point estimates are depicted by dots and horizontal lines are 95% confidence intervals. The dashed vertical line represents the actual AfD vote share (12.6%). Source: GIP Wave 30.

For a first test of violated assumptions, we limit our sample to respondents with specific characteristics that may render the subsample more likely to accurately predict AfD vote share. As we argue above, individually more accurate estimators can improve a group's estimates. By design, however, these subsamples are also less diverse than the full sample which, as we also argue above, is a significant disadvantage for group estimates. Consider our first subsample composed of respondents who openly self-identify as AfD voters in a direct vote-intention item as example. 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 intention. Hence, the subsample should be better at predicting AfD vote share than the full sample. 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 perform much better than homogenous crowds. Furthermore, partisan supporters' beliefs are likely to be biased by wishful thinking (Meffert et al. 2011; Stiers and Dassonneville 2018) which makes an accurate measurement of prevalence even less likely.

With a similar reasoning, we filter out additional subsamples. Our second and third subsets are East Germans (respondents who live in the former GDR or Berlin) and West Germans, respectively. While West Germany has a tradition of almost 70 years of democratic procedures including free elections under virtually identical institutional circumstances, East Germany's democratic tradition counts less than 30 years and severely fewer Bundestag elections. 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. It would, thus, not be surprising if these factors also affected respondents' ability to correctly predict AfD vote shares (Leiter et al. 2018).

The last subsample we consider are respondents with a graduate degree. University graduates may be experts because they are, on average, more capable of judging and predicting political events. At the same time, previous research has used education as a proxy for social network size. Overall, more educated respondents have a larger social network whose knowledge they can exploit to improve their predictive ability (Leiter et al. 2018).

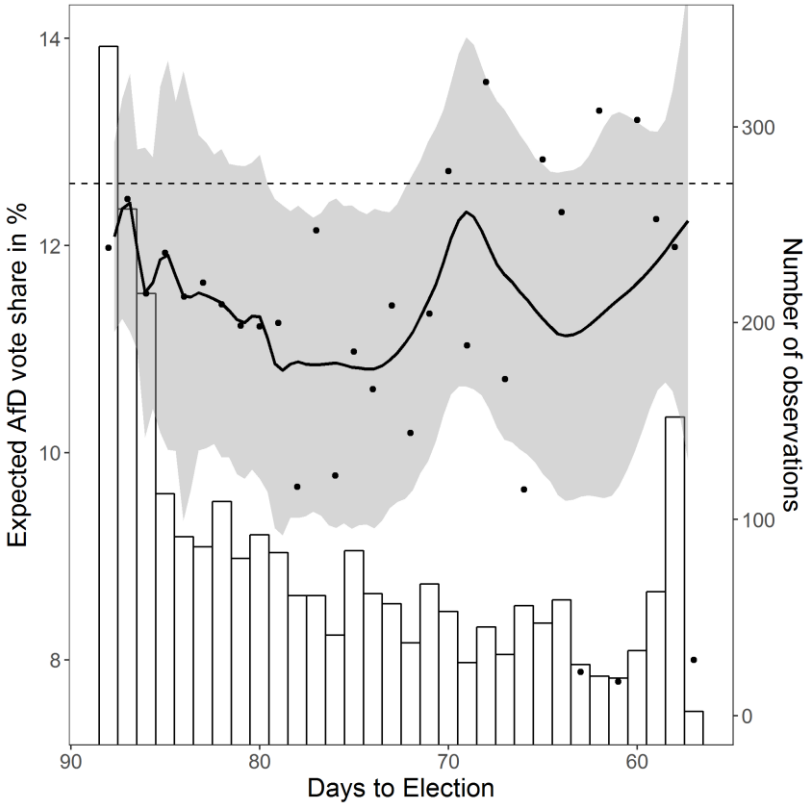
Despite these arguments why certain subsamples of the population should be better at predicting AfD vote share, recall that these subsamples are filtered out on a specific demographic making it likely that the resulting group estimate is a product of less diverse individual estimates. If this is true, all of these subsamples should perform worse than the full sample when predicting AfD vote share.

The results of the subsamples' predictions are depicted in Figure 1. All subsamples 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. It rather suggests that the diversity provided by the full sample drives the crowd's wisdom.

We now turn to a more fine-grained disaggregation of our sample that also allows us to evaluate how the number of respondents affects a group's predicative ability. 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 2 shows, there is high variance in the mean estimates between days (points) the survey had been in the field. 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 within a few days. While they cover a range of about 2 percentage points (expected vote shares are roughly between 11 and 12.5%), their confidence intervals are always significantly larger than 1.5 percentage points implying that results are statistically indistinguishable from one another. Over-

all, these results suggest that given a reasonable sample size, the Wisdom of Crowds design works well.

Figure 2: 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. The dashed line is the AfD election result (12.6%). Source: GIP Wave 30.

Finally, we relax a different assumption of the Wisdom of Crowds design. We aggregate group beliefs by computing their median rather than their mean. This follows the approach originally developed by Galton (1907). Besides the historical relevance, median estimates are less susceptible to outliers than mean estimates. Particularly, if the true prevalence is fairly high or low, say 1%, an outlying estimate, say 20%, is counterbalanced in the mean estimate context by 20 estimates of 0%. In the median estimate context, a single 0% estimate suffices.

However, the Wisdom of Crowds' mathematical underpinnings work only when relying on the mean estimate.

As the bottom row in Figure 1 indicates, there is no difference to the mean estimate. Future research will need to fully investigate how differences in median and mean estimates are linked to a crowd's diversity and its predictive ability. Nevertheless, we note that when researchers suspect that a crowd's predictions are skewed, for instance because of correlated beliefs, relying on the median estimator may proof fruitful. At the very least, researchers are able to obtain an upper and lower bound of their prevalence of interest.

We conclude that there is a systematic relationship between the Wisdom of Crowds assumptions and the design's estimation error. The Wisdom of Crowds design works when its assumptions are met, and it fails when its assumptions are violated. Yet, how well does it perform when its assumptions are met in comparison to other questioning techniques?

4. Comparing the Wisdom of Crowds Design to other Questioning Techniques

We compare the Wisdom of Crowds design to the direct vote-intention question, the crosswise-model randomized response technique (RRT), and the double list experiment. We now present their implementations in the GIP as well as some methodological aspects before presenting evidence that the Wisdom of Crowds design can outperform other methods.

4.1. Implementation of Additional Question Techniques

4.1.1. Direct Vote-Intention Question

We 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. 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 can choose one of thirteen answer options that include the set of parties that were represented in the Bundestag at the time (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.

4.1.2. *Crosswise-Model 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; Rosenfeld et al. 2016). It simultaneously confronts 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 to 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 RRT in the GIP online questionnaire, respondents first read the following introduction that should prevent any misunderstandings and increase their willingness to comply:

> "*Sometimes, we test new methods in our survey. 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 shown 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 response-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).[7]

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, 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.

4.1.3. Double 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 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. The 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

[7] 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).

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).

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.[8] Table 1 shows the different elements of Lists A and B. In 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 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; Thomas et al. 2017).

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 for Germany (AfD) in the next federal election*	*vote for the Alternative for Germany (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

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

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

where $\bar{X}_{A5}$ is the mean response to List A with five items.[9]

4.2. Methodological Notes

We implement all 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).[10] 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 crosswise-model RRT design. We assign the other two groups to the double 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] 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).

[10] 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 crosswise-model RRT question or the double 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 in%	Missings	Item Non-Response Rate in%
Wisdom of Crowds	2597	100.0	26	1.0
Direct Vote-Intention	2597	100.0	481	18.5
Crosswise-Model RRT	867	33.4	94	10.9
Double List Experiment	1730	66.6	7	0.4
List A	865	33.3	3	0.3
List B	865	33.3	4	0.5

Given that we implement the question in a self-administered online survey, we expect social desirability to be relatively low, yet, not undetectable (e.g., Chang and Krosnick 2010; Kreuter et al. 2008). We, thus, consider the comparative test here a hard case because the direct question should perform better than in, say, interviewer-administered face-to-face surveys.

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 participates in the survey, do 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 double list experiment yield low item non-response rates. The crosswise-model RRT provides a non-response rate that lies in between the rates of direct questioning and the two other designs, Wisdom of Crowds and the double list experiment. We attribute the crosswise-model

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; Rosenfeld et al. 2016; 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 sample size, simply comparing standard errors and confidence intervals across different techniques is insufficient because the number of respondents differs between question types (Table 2).

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 affect 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.[11] For comparability, we also apply this procedure to the indirect survey techniques.

4.3. Comparative Performance of Different Questioning Techniques

In this section we compare the performance of various indirect questioning techniques to recover the quantity of interest, namely AfD vote share on Election Day. Figure 3 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%) which serves as our behavioral benchmark.

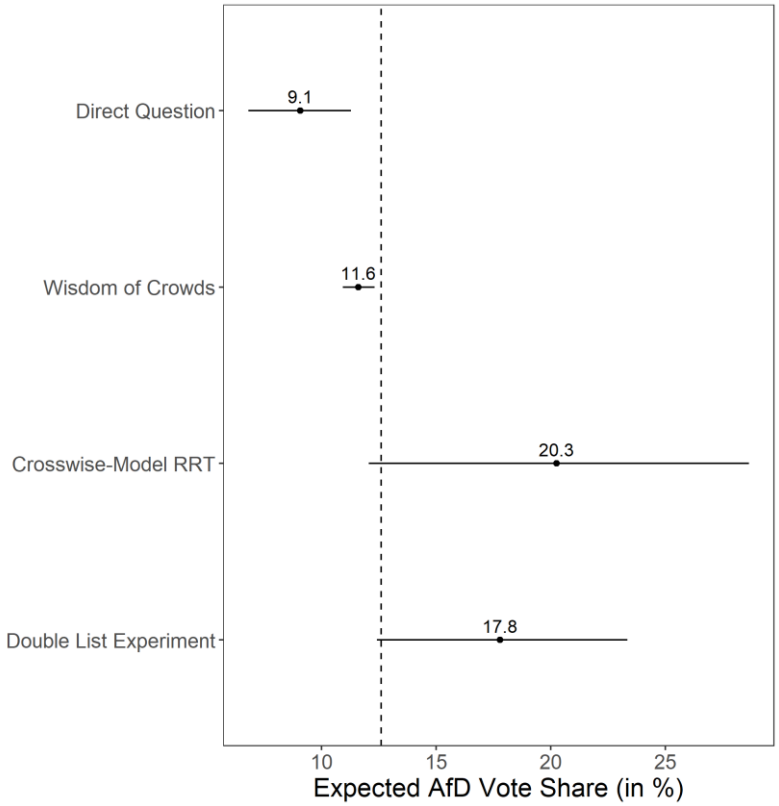
To ease comparison, the Wisdom of Crowds estimate is in Figure 3's top row, along with its bootstrapped confidence intervals. As shown above, it underestimates AfD vote share by only 1 percentage point. At the same time, the narrow bootstrapped 95% confidence interval covers a range of 1.4 percentage points and is significantly smaller than any other techniques' confidence bounds. Yet, it does not cover the actual AfD vote share on Election Day.

In comparison to other techniques, we note that there are striking differences between estimates. 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.

[11] We also present results for bootstraps with the full sample size.

Compared to this result, other German polling institutes and forecasting models like *Zweitstimme.org* (Munzert et al. 2017, Stoetzer et al. 2019) report an AfD vote share between 7% (Allensbach) and 9% (Infratest Dimap) at the time our survey was in the field.[12]

Figure 3: 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 (12.6%). Source: GIP Wave 30.

[12] The website *wahlrecht.de* collects published polls over time (http://www.wahlrecht.de/umfragen/). Our estimates are rather similar to those we report in Figure 1 when we set the size of the bootstrapped sample to the size of the full dataset. The 2597 available responses (including 481 missing responses and 152 respondents who state that they would not vote or are not eligible to vote) result in the same expected AfD vote share of 9.1%. As expected when using more than twice as many data points, the 95% confidence interval is tighter and covers the area [7.87%; 10.30%].

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 the following 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 both in 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. 2018; Blom et al. 2019).[13] 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

[13] 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.

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.

Although the crosswise-model RRT overestimates AfD vote share by 7.5 percentage points, the wide confidence intervals (more than 17 percentage points) cover the true AfD vote share. The surprisingly high point estimate indicates an issue with false positives that are triggered by the survey design (Höglinger and Diekmann 2017). It also supports the finding reported in numerous previous studies that this design introduces several other problems like noncompliance or misunderstandings concerning the instructions (Waubert de Puiseau et al. 2017; Höglinger et al. 2016; Jann et al. 2012; Krumpal 2012; Coutts and Jann 2011; Yu et al. 2008). Rather similar results are obtained by the double list experiment. It overestimates AfD vote share somewhat less than the crosswise-model RRT (5.4 percentage points overestimation) and has tighter confidence bounds (11.5 percentage points). Still, as already reported by Coutts and Jann (2011, 183), we find that the estimate derived from the double list experiment comes with high levels of uncertainty which makes the substantive interpretation of the results difficult.

In comparison, all three techniques for sensitive questions outperform the 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 (crosswise-model RRT and double list experiment). However, the crosswise-model RRT's and the double list experiment's

confidence bounds are not only their advantage, they are also a notable drawback.[14] 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 crosswise-model RRT design and the double list experiment renders their estimates almost useless for inferences on parties' electoral performance. This problem is likely to inflict many other research subjects as well. Since the Wisdom of Crowds design does not imply a tradeoff between efficiency and anonymity, it does not suffer from increased inefficiency. Hence, this design is a superior alternative if precise estimates are required, given that its assumptions are met.

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

[14] 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 crosswise-model RRT nor the double list experiment recovers the "true" AfD vote share in this subsample (100%) precisely.

any answer at all or untruthful answers if they believe that a honest response would run counter to social norms (Philips and Clancy 1972; Singer et al. 1995; Singer et al. 1992).

To mitigate the problems caused by social desirability bias, we introduce the Wisdom of Crowds design. 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 provide good estimates of complex phenomena when their members make independent choices, think diversely, are organized in a decentralized way, and an appropriate aggregation mechanism combines individual information. In a nutshell, the Wisdom of Crowds works because diversity contributes new information to the crowd's information pool and independent mistakes cancel out across individual beliefs. Based on this, we develop the Wisdom of Crowds survey design to learn what share of a target population has a sensitive trait.

Following Surowiecki (2004), we argue that crowds need diversity, independence, decentralization, and an aggregation mechanism to be wise. We investigate the effect of violations of these assumptions, e.g., focusing on subsamples and thereby making the crowd less diverse, independent, and decentralized. Empirically, we focus on the 2017 German federal election with data from the German Internet Panel (Blom et al. 2015). In particular, we predict the vote share of the right-wing populist party *Alternative for Germany* (AfD) which was constantly underestimated by public opinion polls (Bergmann and Diermeier 2017; Waubert de Puiseau et al. 2017). Using these data, we show that the Wisdom of Crowds assumptions are integral preconditions for crowds to be wise, and that crowds perform poorly when they are too homogenous. We conclude that the Wisdom of Crowds survey design is a valid survey tech-

nique for sensitive question if its assumptions are met, i.e., when there is sufficient estimation variance to counterbalance individual errors.

Further, we compare the Wisdom of Crowds design to other survey techniques for sensitive questions (i.e., crosswise-model Randomized Response Technique and double list experiment) in the context of the 2017 German federal election. 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 (crosswise-model RRT and double list experiment). Yet, the high uncertainty associated with both indirect questioning techniques renders these methods only useful if very large samples can be realized for an accurate prediction of vote shares. In contrast, the Wisdom of Crowds design performs remarkably well in terms of bias and efficiency.

The Wisdom of Crowds design's good performance is due to its approach to respondent privacy. While other survey techniques for sensitive questions 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. Therefore, it does not constitute a tradeoff between respondents' anonymity and estimate efficiency. Furthermore, as compared to the crosswise-model RRT and the double list experiment, the Wisdom of Crowd design does not impose high cognitive demands 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.

The results presented here have important implications for research on sensitive characteristics in general. Most importantly, the results suggest that the Wisdom of Crowds design is an additional and – when its assumptions are met – seemingly superior survey technique to learn about the prevalence of sensitive traits in a target population. 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.

Additionally, the limits of crowd knowledge should be explored more systematically. 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. If the crowd is completely uninformed, the diversity condition is violated since individuals in the crowd have the same level of information. As we show here, this violation most likely renders the Wisdom of Crowds survey design useless in such contexts. Future research should seek to determine which topics crowds are competent to judge and which they are not.

The Wisdom of Crowds design may also 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. Finally, the Wisdom of Crowds design may constitute a way how researchers cannot only learn about the prevalence of a sensitive characteristic, but also about which respondents have it. If, for a given characteristic, there is a systematic link to respond-

ents' beliefs about the characteristic, say drug users overestimate the prevalence of drug users, then the Wisdom of Crowds design can help to identify probable drug users. Further research is needed to establish under what conditions this strategy is fruitful.

We show that the generally good performance of the Wisdom of Crowds design in the context of election forecasting rests on the validity of four conditions. If these conditions are met, the Wisdom of Crowds design constitutes a promising alternative to the inefficient indirect questioning techniques in other research contexts where social desirability bias is a concern.

References

Berbuir, N., Lewandowsky, M., Siri, J., 2015. The AfD and its Sympathisers: Finally a Right-Wing Populist Movement in Germany? Ger. Polit. 24, 154–178. crosswise-model

Bergmann, K., Diermeier, M., 2017. Die AfD: Eine unterschätzte Partei. Soziale Erwünschtheit als Erklärung für fehlerhafte Prognosen (No. 7/2017), IW-Report.

Blom, A.G., Felderer, B., Herzing, J., Krieger, U., Rettig, T., Political Economy of Reforms, Universität Mannheim, S. 884, 2019. German Internet Panel, Wave 31 (September 2017). GESIS Data Archive, Cologne. Forthcoming.

Blom, A.G., Felderer, B., Herzing, J., Krieger, U., Rettig, T., Political Economy of Reforms, Universität Mannheim, S. 884, 2018. German Internet Panel, Wave 30 (July 2017), GESIS Data Archive, Cologne. ZA6904 Data file Version 1.0.0.

Blom, A.G., Gathmann, C., Krieger, U., 2015. Setting Up an Online Panel Representative of the General Population: The German Internet Panel. Field Methods 27, 391–408.

Blom, A.G., Herzing, J.M.E., Cornesse, C., Sakshaug, J.W., Krieger, U., Bossert, D., 2017. Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. Soc. Sci. Comput. Rev. 35, 498–520.

Buhl, Y., 2017. Die unterschätzten Rechtspopulisten. Wird die AfD bei der Bundestagswahl stärker, als es Umfragen und Prognosen vorhersagen? CORRELAID.ORG BLOG.

Chang, L., Krosnick, J.A., 2010. Comparing oral interviewing with self-administered computerized questionnaires: An experiment. Public Opin. Q. 74, 154–167.

Coutts, E., Jann, B., 2011. Sensitive questions in online surveys: Experimental results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). Sociol. Methods Res. 40, 169–193.

Crowne, D.P., Marlowe, D., 1960. A new scale of social desirability independent of psychopathology. J. Consult. Psychol. 24, 349–354.

Diekmann, A., 2012. Making Use of "Benford's Law" for the Randomized Response Technique. Sociol. Methods Res. 41, 325–334.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W., Ezzati, T.M., 1991. The Item-Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application, in: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (Eds.), Measurement Errors in Surveys. John Wiley & Sons Inc, Hoboken, NJ, USA, pp. 185–210.

Gaissmaier, W., Marewski, J.N., 2011. Forecasting elections with mere recognition from small, lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls. Judgement Decis. Mak. 6, 73–88.

Galton, F., 1907. Vox populi. Nature 75, 450–451.

Glynn, A.N., 2013. What can we learn with statistical truth serum? Public Opin. Q. 77, 159–172.

Graefe, A., 2014. Accuracy of vote expectation surveys in forecasting elections. Public Opin. Q. 78, 204–232.

Höglinger, M., Diekmann, A., 2017. Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. Polit. Anal. 25, 131–137.

Höglinger, M., Jann, B., Diekmann, A., 2016. Sensitive Questions in Online Surveys: An Experimental Comparison of the Randomized Response Technique and the Crosswise Model. Surv. Res. Methods 10, 171–187.

Holbrook, A.L., Krosnick, J.A., 2010. Measuring Voter Turnout by using the Randomized Response Technique: Evidence Calling Into Question the Method's Validity. Public Opin. Q. 74, 328–343.

Holbrook, A.L., Krosnick, J.A., 2010. Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opin. Q. 74, 37–67.

Hong, L., Page, S., 2009. Interpreted and generated signals. J. Econ. Theory 144, 2174–2196.

Inglehart, R., Norris, P., 2016. Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash, HKS Working Paper No. RWP16-026.

Jann, B., Jerke, J., Krumpal, I., 2012. Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. Public Opin. Q. 76, 32–49.

Johann, D., Thomas, K., Faas, T., Fietkau, S., 2016. Alternative Messverfahren rechtspopulistischen Wählens im Vergleich: Empirische Erkenntnisse aus Deutschland und Österreich, in: Schoen, H., Weßels, B. (Eds.), Wahlen Und Wähler: Analysen Aus Anlass Der Bundestagswahl 2013. Springer, Wiesbaden, pp. 447–470.

Kreuter, F., Presser, S., Tourangeau, R., 2008. Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. Public Opin. Q. 72, 847–865.

Krogh, A., Vedelsby, J., 1994. Neural Network Ensembles, Cross Validation and Active Learning, in: Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94. MIT Press, Cambridge, MA, USA, pp. 231–238.

Krosnick, J.A., 1999. Survey Research. Annu. Rev. Psychol. 50, 537–567.

Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl. Cogn. Psychol. 5, 213–236.

Krumpal, I., 2012. Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. Soc. Sci. Res. 41, 1387–1403.

Kuklinski, J.H., Sniderman, P.M., Knight, K., Piazza, T., Tetlock, P.E., Lawrence, G.R., Mellers, B., 1997. Racial Prejudice and Attitudes Toward Affirmative Action. Am. J. Pol. Sci. 41, 402–419.

Leiter, D., Murr, A., Rascón Ramírez, E., Stegmaier, M., 2018. Social networks and citizen election forecasting: The more friends the better. Int. J. Forecast. 34, 235–248.

Lewis-Beck, M.S., Skalaban, A., 1989. Citizen Forecasting: Can Voters See into the Future? Br. J. Polit. Sci. 19, 146–153.

Lewis-Beck, M.S., Stegmaier, M., 2011. Citizen forecasting: Can UK voters see the future? Elect. Stud. 30, 264–268.

Lewis-Beck, M.S., Tien, C., 1999. Voters as forecasters: A micromodel of election prediction. Int. J. Forecast. 15, 175–184.

Meffert, M.F., Huber, S., Gschwend, T., Pappi, F.U., 2011. More than wishful thinking: Causes and consequences of voters' electoral expectations about parties and coalitions. Elect. Stud. 30, 804–815.

Munzert, S., Stötzer, L., Gschwend, T., Neunhoeffer, M., Sternberg, S., 2017. Zweitstimme.org. Ein strukturell-dynamisches Vorhersagemodell für Bundestagswahlen. Polit. Vierteljahresschr. 58, 418–441.

Murr, A.E., 2017. Wisdom of Crowds, in: Arzheimer, K., Evans, J., Lewis-Beck, M.S. (Eds.), Handbook of Political Behavior. Sage, Los Angeles, pp. 835–860.

Murr, A.E., 2016. The wisdom of crowds: What do citizens forecast for the 2015 British General Election? Elect. Stud. 41, 283–288.

Murr, A.E., 2015. The wisdom of crowds: Applying Condorcet's jury theorem to forecasting US presidential elections. Int. J. Forecast. 31, 916–929.

Murr, A.E., 2011. "Wisdom of crowds" ? A decentralised election forecasting model that uses citizens' local expectations. Elect. Stud. 30, 771–783.

Page, S.E., 2014. Where diversity comes from and why it matters? Eur. J. Soc. Psychol. 44, 267–279.

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

Phillips, D.L., Clancy, K.J., 1972. Some Effects of "Social Desirability" in Survey Studies. Am. J. Sociol. 77, 921–940.

Rosenfeld, B., Imai, K., Shapiro, J.N., 2016. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. Am. J. Pol. Sci. 60, 783–802.

Schmitt-Beck, R., 2017. The 'Alternative für Deutschland in the Electorate': Between Single-Issue and Right-Wing Populist Party. Ger. Polit. 26, 124–148.

Singer, E., Hippler, H.J., Schwarz, N., 1992. Confidentiality assurances in surveys: Reassurance or threat? Int. J. Public Opin. Res. 4, 256–268.

Singer, E., von Thurn, D.R., Miller, E.R., 1995. Confidentiality Assurances and Response Experimental Literature. Public Opin. Q. 59, 66–77.

Sjöberg, L., 2009. Are all crowds equally wise? A comparison of political election forecasts by experts and the public. J. Forecast. 28, 1–18.

Stiers, D., Dassonneville, R., 2018. Affect versus cognition: Wishful thinking on election day: An analysis using exit poll data from Belgium. Int. J. Forecast. 34, 199–215.

Stoetzer, L.F., Gschwend, T., Neunhoeffer, M., Munzert, S., Sternberg, S., 2019. Forecasting Elections in Multi-Party Systems: A Bayesian Approach Combining Polls and Fundamentals. Polit. Anal. forthcoming.

Streb, M.J., Burrell, B., Frederick, B., Genovese, M.A., 2008. Social desirability effects and support for a female American president. Public Opin. Q. 72, 76–89.

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

Thomas, K., Johann, D., Kritzinger, S., Plescia, C., Zeglovits, E., 2017. Estimating Sensitive Behavior: The ICT and High-Incidence Electoral Behavior. Int. J. Public Opin. Res. 29,

Tourangeau, R., Rips, L.J., Rasinski, K.A., 2000. The psychology of survey response. Cambridge University Press.

Tourangeau, R., Yan, T., 2007. Sensitive Questions in Surveys. Psychol. Bull. 133, 859–883.

Waubert de Puiseau, B., Hoffmann, A., Musch, J., 2017. How Indirect Questioning Techniques May Promote Democracy: A Preelection Polling Experiment. Basic Appl. Soc. Psych. 39, 209–217.

Yu, J.-W., Tian, G.-L., Tang, M.-L., 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika 67, 251–263.

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 prevalence of interest, x_i is the ith 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.[15] 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 all determining factors, leaving alone having access to sufficient data to gauge the values of all β_i's (Page 2007).

[15] 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).

Appendix 2: Cognitive Demand on Respondents

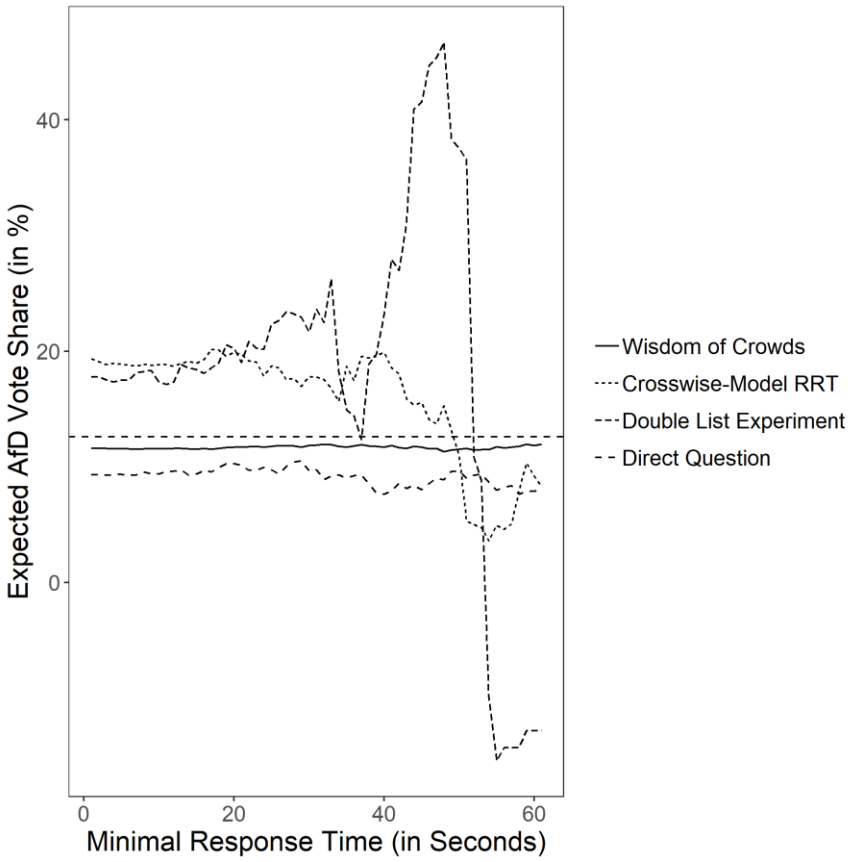
Both techniques, the double list experiment and the crosswise-model RRT, are cognitively more demanding as compared to the direct vote-intention question and the Wisdom of Crowds technique since respondents need to evaluate multiple questions and aggregate individual responses. It is therefore possible 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 crosswise-model 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 end 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 double list experiment and the crosswise-model 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 (12.6%). Source: GIP Wave 30.

Appendix 3: Validation Based on Self-Identified AfD Voters

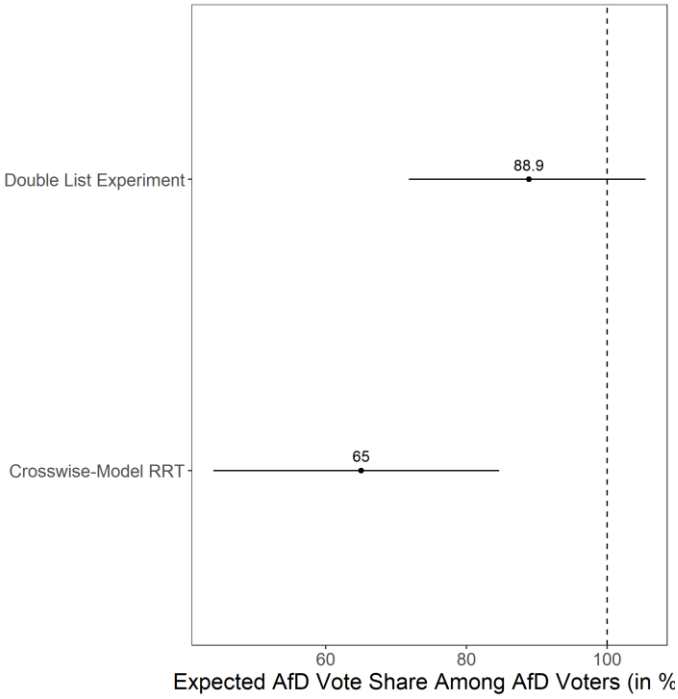
We also assess the validity of the results by only regarding respondents who indicate that they want to vote for the AfD in the direct vote-intention question. Under the assumption that respondents who indicate their willingness to vote for the AfD in the direct vote-intention question will also declare to vote for the AfD in the list experiment and the crosswise-model RRT design where the anonymity of their responses is assured, we expect the estimates for this subsample to 100%.[16] 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 crosswise-model 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 crosswise-model 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

[16] 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.

signal, although a theoretically appealing approach, does not seem to be particularly useful at least in the implementation we chose 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 (100%). Source: GIP Wave 30.