Online Appendix for

Lehrer, Roni, Sebastian Juhl, and Thomas Gschwend.

"The Wisdom of Crowds Design for Sensitive Survey Questions"

Electoral Studies

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 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.[1] Due to the high complexity of social reality the number of factors that co-determine y is large. This

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

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). 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

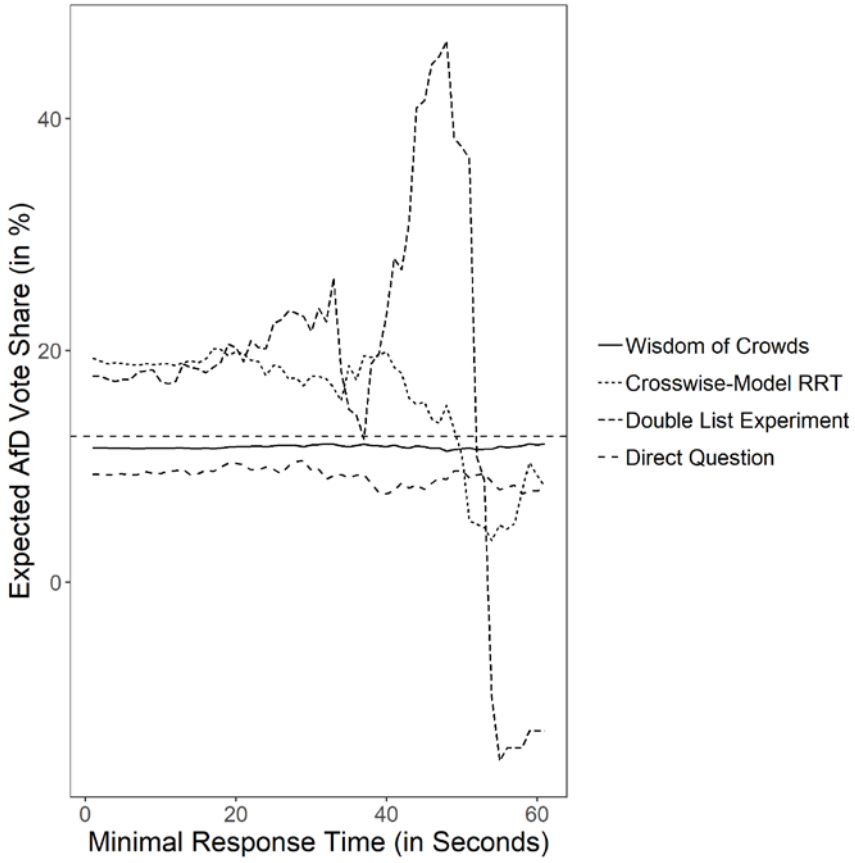
Both techniques, the double list experiment and the crosswise-model RRT, are cognitively more demanding as compared to the direct vote-intention question and the Wisdom of Crowds technique since respondents need to evaluate multiple questions and aggregate individual responses. It is therefore possible 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 3 in the article. 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 dataset 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 technique 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

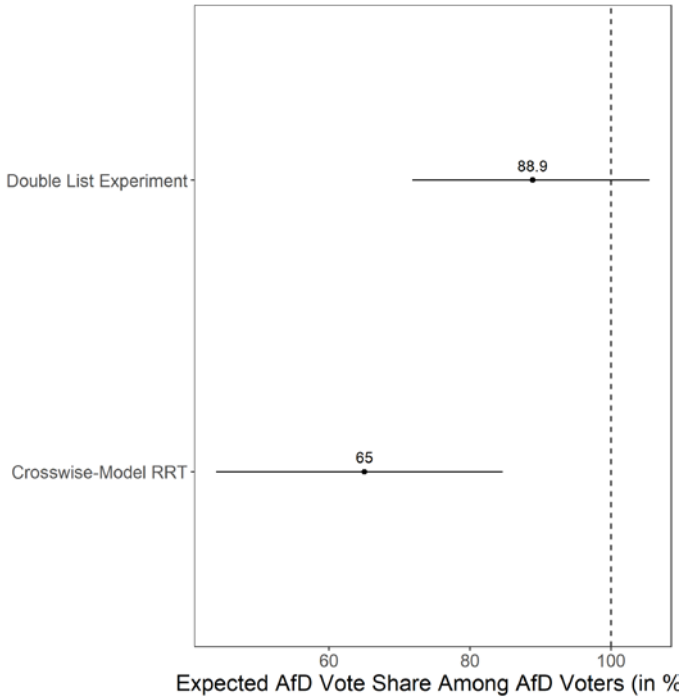
We also assess the validity of the results by only regarding respondents who indicate that they want to vote for the AfD in the direct vote-intention question. Under the assumption that respondents who indicate their willingness to vote for the AfD in the direct vote-intention question will also declare 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%.[2] 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 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 estimate inefficiency, increased cognitive demands on respondents, and problems of noncompliance. Thus, the addition of noise to the signal, although a theoretically appealing approach, does not seem to be particularly useful at

[2] 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.

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.