The predictive power of subjective probabilities: probabilistic and deterministic polling in the Dutch 2017 election

Jochem de Bresser and Arthur van Soest
Tilburg University and Netspar, Tilburg, The Netherlands

[Received November 2017. Revised July 2018]

Summary. The paper evaluates the predictive validity of stated intentions for actual behaviour. In the context of the 2017 Dutch parliamentary election, we compare how well polls based on probabilistic and deterministic questions line up with subsequent votes. Our empirical strategy is built around a randomized experiment in a representative panel. Respondents were either asked which party they will vote for or were asked to allocate probabilities of voting for each party. The results show that probabilities predict individual behaviour better than deterministic statements for a large majority of respondents. There is, however, substantial heterogeneity in the predictive power of subjective probabilities. We find evidence that they work better for those with higher probability numeracy, even though probability numeracy was measured 8 years earlier.

Keywords: Elections; Predictive validity; Probabilistic polling; Subjective probabilities

1. Introduction

This study investigates the value of subjective probability questions for predicting future individual behaviour, analysing a randomized experiment on intended and actual voting behaviour for the Dutch parliamentary 2017 election. It compares the predictive power of subjective probabilities with that of the traditional way of eliciting intentions through deterministic questions. In the experiment, respondents were randomly allocated to either type of questions measuring their expectations concerning the party that they will vote for (so-called choice expectations (Manski, 2004), or intentions). These intentions were elicited approximately 3 months before the election on March 15th, 2017. We compare intentions that were elicited by deterministic items (‘which party will you vote for?’) with probabilistic intentions (‘what is the probability that you will vote for party x?’). Since individuals do not face any restrictions on their actual voting behaviour, this is a clean case in which intentions and outcomes can be compared without the need to model or make assumptions on exogenous events that may influence the actual outcome.

The idea of using subjective probabilities to elicit voting intentions goes back to Meier and Campbell (1979), Meier (1980) and Maas et al. (1990), but none of these studies compared probabilistic and deterministic approaches. Manski (2004) reported on a small pilot study for the 2000 US presidential election and large-scale probabilistic polls have been carried out for the presidential races of 2008, 2012 and 2016. Research on the last two elections focused on the extent to which probabilistic polls anticipated the actual aggregate election outcome. Evidence has been
mixed: it was one of the most accurate polls in 2012 (Gutsche et al., 2014), but it substantially overpredicted the Republicans’ share of the popular vote in 2016. The analysis of the 2008 elections that was reported by Delavande and Manski (2010) is closest to that of the present paper, because it considers the predictive power of verbal and probabilistic polling questions at the level of the individual. Delavande and Manski (2010) showed that combining both types of item improves the prediction of actual votes. However, doing so is costly, since it entails asking two sets of questions to elicit voting intentions. Delavande and Manski (2010) acknowledged that their research design, in which both probabilistic and verbal questions were posed in quick succession to all respondents, did not allow them to evaluate which type of question works best. After all, responses to the verbal questions may be affected by the probabilities that respondents reported previously. Our empirical strategy avoids this problem, since it is based on a large split-sample design.

Three features distinguish the present study from previous work. Firstly, we exploit a randomized experiment in a large, representative household panel that enables us to compare the predictive power of deterministic and probabilistic intentions in a clean way. In contrast with the research that was described above, panel members were exclusively assigned to either type of question. Secondly, whereas previous efforts focused on US presidential races that effectively amount to binary choices, we analyse the more fragmented setting of parliamentary elections in the Netherlands. On March 15th, 2017, the ballot listed 28 parties, 13 of which made it into the Parliament. Such a profusion of options presumably makes probabilities more powerful, since there is more scope for doubt experienced by undecided voters, particularly when the election is still some time off (3 months, in our case). Finally, our data come from a long-standing panel for which much information has been collected in prior surveys. This enables us to relate the predictive power of reported probabilities to relevant background information, such as probability numeracy.

Our results, based on multinomial discrete choice models, indicate that on average and for the large majority of the population probabilistic questions are substantially better predictors of actual votes than deterministic questions. We find that an increase in the reported probability of voting for a party from 0 to 20% increases the likelihood of actually voting for that party by 3–11 percentage points (PPs) in the deterministic sample, compared with an increase of 5–19 PPs in the probabilistic sample. We show that this added power of probabilities can be attributed to the question format and not to systematic differences between samples. Although there is little variation in the predictive power of deterministic intentions, estimates of a random-coefficients discrete choice model reveal substantial heterogeneity in the predictive power of subjective probabilities. They work very well for a large majority of the respondents (84%, according to our estimates) but perform worse than deterministic statements for a small minority (16% of the sample). This heterogeneity is related to probability numeracy: probabilities are better predictors for individuals with higher probability numeracy. This finding is in accordance with earlier studies demonstrating substantial heterogeneity in individuals’ ability to work with probabilities and, in relation to that, the value of their answers to subjective probability questions for predicting actual behaviour (see Armantier et al. (2015) and Binswanger and Salm (2017)).

To choose between a survey design with subjective probabilities or deterministic intentions, a trade-off should be made between benefits and costs. We therefore also briefly consider potential costs, due to a larger burden on the respondents. We find that the survey with 15 subjective probabilities takes significantly longer than the same survey in which a single option is chosen out of 15 alternatives, with a difference at the median of slightly less than 2 min. We find no significant difference between the two designs in the respondents’ evaluations of survey difficulty or attractiveness. We conclude that the costs are dominated by the much larger predictive power
of probabilities at the level of the individual. This study therefore adds to the evidence that election polls in fragmented political systems could be improved by asking probabilities rather than discrete voting intentions.

Our findings are potentially relevant for the usefulness of subjective probabilities on other types of future decisions or events, such as those that feature in intertemporal economic models. The consensus among economists is that beliefs are best measured through probabilities rather than qualitative statements, because they allow one to express uncertainty and are comparable across individuals (Manski, 2004). However, reported probabilities have been found to be affected by non-classical measurement error such as rounding (Manski and Molinari, 2010; Kleinjans and Van Soest, 2014). Many studies have demonstrated that subjective probabilities have empirical validity: they correlate in plausible ways with background variables and help to predict future outcomes and decisions (see, for example, the overview in Hurd (2009)). Until now there has been no direct evidence comparing the predictive power of subjective probabilities with that of the traditional way of eliciting intentions through deterministic questions.

The paper is organized as follows. Section 2 describes the data and presents some descriptive statistics. Sections 3 and 4 contain our results regarding the predictive power of deterministic and probabilistic intentions. We first estimate multinomial choice models that demonstrate that subjective probabilities provide additional power to predict each individual’s actual voting behaviour. Using one of our models, we then construct an index for predictive power and relate it to probability numeracy. Section 5 briefly analyses the additional respondent burden of subjective probability questions. Section 6 concludes.

2. Data

2.1. The ‘Longitudinal Internet studies for the social sciences’ panel

The ‘Longitudinal Internet studies for the social sciences’ panel (which is known as the ‘LISS’ panel) is a large household panel, consisting of approximately 8000 individuals in 5000 households that are broadly representative of the Dutch population (Van der Laan, 2009; De Vos, 2010). (More information on the panel, including code books for all available data and instructions on how to obtain access, can be found at www.lissdata.nl.) Households are selected randomly by Statistics Netherlands from the complete registry of all Dutch non-institutionalized households. Surveys are administered online, and selected households receive a simple computer and an Internet connection if they do not have a computer or Internet access.

Longitudinal information on a wide range of socio-economic and demographic topics is collected yearly in so-called ‘core’ surveys. In addition, researchers can design their own questionnaires on specific topics. Our analysis combines data from four different surveys. The votes that were cast by respondents in the parliamentary elections of March 15th, 2017, our main outcome variable, were collected in an ‘exit poll’ during the 2 weeks immediately following the election (between March 16th and 30th). Voting intentions for the same election were collected in the core politics and values survey of December 2016, approximately 3 months before the election. (Respondents were invited to take the survey early in December. Those who did not take the survey in December received another invitation in January 2017; only a small minority used this opportunity.) Background variables are obtained from the household box of that month. Finally, in part of our analysis we use a ‘probability numeracy’ variable that was calculated from items included in a one-off disease prevention survey that was designed by Katie Carman and Peter Kooreman and fielded in September 2008 (see Bruine de Bruin and Carman (2012) and Carman and Kooreman (2014)). Unfortunately, a more recent numeracy measure for the LISS respondents is not available.
2.2. Probabilistic poll

Crucially for this study, voting intentions were measured differently in 2016 compared with previous years. Inspired by the probabilistic polls for the 2012 and 2016 US presidential elections, an experiment was set up to compare responses to different types of polling questions. (See Gutsche et al. (2014), www.alpdata.rand.org/?page=election2012 and www.cesrsc.org/election/.) All respondents were asked to report their voting intentions in two steps. First, they were all asked in the same way to indicate the probability that they would vote:

‘If parliamentary elections were held today, what is the percent chance that you will vote? Please fill in a percentage between 0 and 100.’

The intention not to vote is computed as 100% minus the percentage probability that was reported in response to this question and refers to the option not to hand in a ballot.

Second, respondents forecasted which party they would vote for conditionally on voting. One random half of the panel received a single deterministic question:

‘If parliamentary elections were held today, for which party would you vote? [If Pr(vote) = 0: I would not vote], VVD (liberal party), . . . , VNL, Another party, Blank.’

This is the usual way that voting intentions are measured in the LISS. The answer options are the 13 parties represented in Parliament at the time of the survey, any other party and not casting a vote on any of the parties (‘blank’). Respondents who gave a 0% probability of voting at all in the previous question received an additional option ‘I would not vote’.

The other half of the sample were asked to assign probabilities of voting for different parties, voting ‘blank’, or not voting at all:

‘If parliamentary elections were held today, what is the percent chance that you will vote for each of the following parties? Total probability should add up to 100%. [if Pr(vote) = 0: I would not vote], VVD (liberal party), . . . , VNL, Another party, Blank.’

To help respondents to answer these questions in a logically consistent way, all parties were shown on a single screen and the total probability mass that they had already distributed was shown at the bottom. Respondents did not have to assign 15 (or 16) probabilities explicitly: fields left empty were counted as 0s. Moreover, respondents could not proceed to the next question in the survey if they provided probabilities outside the 0–100 interval or if their probabilities did not add up to 100%.

The ‘treatment’ deterministic or probability questions was assigned completely randomly. Balance of covariates by means of balanced sampling (see, for example, Deville and Tillé (2004) and Tillé (2011)) was not ensured. Nonetheless, the two treatment groups are similar in terms of observable characteristics (see the balance tests that are reported in Table 1).

For our analysis, we compute the unconditional probabilities for voting for each party, combining the (unconditional) probability of voting at all with the conditional probabilities of voting for each of the parties given voting:

\[
Pr(\text{vote party } x) = Pr(\text{party } x \mid \text{vote}) Pr(\text{vote}).
\] (1)

Moreover, ‘no vote’ is added as the remaining possible outcome (with probability 1 – Pr(vote) as explained above). As explained above, half of the respondents were given deterministic questions, restricting their probabilities \(Pr(\text{party } x \mid \text{vote})\) to 0 or 1. In contrast, all respondents report their subjective probability of voting, \(Pr(\text{vote})\), as a probability. This implies that the unconditional probabilities can also take values between 0% and 100%. We use the fact that there is no difference in elicitation method for ‘no vote’ as a placebo treatment, since there is no reason to expect any
**Table 1.** Balance tests: descriptive statistics of covariates by question type $^\dagger$

|                  | Overall       | Deterministic sample mean | Probabilistic sample mean | Difference |
|------------------|---------------|---------------------------|---------------------------|------------|
|                  | $N$           | Mean                      | Standard deviation        |            |
| Female           | 3978          | 0.51                      | 0.50                      | 0.50       | $-0.01$ (0.016) |
| Household members| 3978          | 2.4                       | 1.3                       | 2.4        | 2.5          | $0.07$§§ (0.039) |
| Partner          | 3978          | 0.70                      | 0.46                      | 0.69       | 0.72         | $0.03$§§ (0.015) |
| Age              | 3978          | 54                        | 17                        | 54         | 54           | $-0.1$ (0.54)    |
| Net household income | 3882      | 3108                      | 3574                      | 3148       | 3070         | $-78$ (116.0)    |
| Homeowner        | 3938          | 0.73                      | 0.44                      | 0.74       | 0.73         | $-0.01$ (0.014) |
| Probability numeracy $^*$ | 2033 | 0.00                      | 0.84                      | $-0.04$    | 0.04         | $0.07$§§ (0.038) |

**Education**

|                  | Overall       | Deterministic sample mean | Probabilistic sample mean | Difference |
|------------------|---------------|---------------------------|---------------------------|------------|
|                  | $N$           | Mean                      | Standard deviation        |            |
| Primary          | 3974          | 0.06                      | 0.24                      | 0.06       | 0.06         | $-0.01$ (0.007) |
| Intermediate secondary | 3974   | 0.22                      | 0.41                      | 0.23       | 0.21         | $-0.02$§§ (0.013) |
| Higher secondary | 3974          | 0.11                      | 0.31                      | 0.10       | 0.12         | $0.02$§ (0.010)  |
| Intermediate vocational | 3974 | 0.24                      | 0.43                      | 0.25       | 0.23         | $-0.02$ (0.014)  |
| Higher vocational | 3974          | 0.26                      | 0.44                      | 0.25       | 0.26         | $0.02$ (0.014)   |
| University       | 3974          | 0.12                      | 0.32                      | 0.11       | 0.12         | $0.01$ (0.010)   |

**Ethnicity**

|                  | Overall       | Deterministic sample mean | Probabilistic sample mean | Difference |
|------------------|---------------|---------------------------|---------------------------|------------|
|                  | $N$           | Mean                      | Standard deviation        |            |
| Dutch            | 3872          | 0.87                      | 0.34                      | 0.86       | 0.88         | $0.02$§ (0.011)  |
| First-generation Western | 3872  | 0.03                      | 0.16                      | 0.03       | 0.03         | $-0.001$ (0.005) |
| First-generation non-Western | 3872 | 0.03                      | 0.18                      | 0.04       | 0.03         | $-0.002$ (0.006) |
| Second-generation Western | 3872 | 0.05                      | 0.22                      | 0.05       | 0.05         | $-0.007$ (0.007) |
| Second-generation non-Western | 3872 | 0.02                      | 0.14                      | 0.03       | 0.01         | $-0.01$§ (0.004) |

**Urbanization**

|                  | Overall       | Deterministic sample mean | Probabilistic sample mean | Difference |
|------------------|---------------|---------------------------|---------------------------|------------|
|                  | $N$           | Mean                      | Standard deviation        |            |
| Extremely        | 3974          | 0.15                      | 0.35                      | 0.15       | 0.14         | $-0.006$ (0.011) |
| Very             | 3974          | 0.26                      | 0.44                      | 0.26       | 0.26         | $0.004$ (0.014)  |
| Moderately       | 3974          | 0.22                      | 0.42                      | 0.23       | 0.21         | $-0.02$ (0.013)  |
| Slightly         | 3974          | 0.22                      | 0.41                      | 0.21       | 0.22         | $0.003$ (0.013)  |
| Rural            | 3974          | 0.16                      | 0.36                      | 0.14       | 0.17         | $0.02$§§ (0.012) |

$^\dagger$Reported statistical significance is not corrected for multiple comparisons. No single null hypothesis is rejected if we correct the initial $p$-cut-off of 0.05 for multiple comparisons by using the methods proposed by Holland and Copenhaver (1987), by Benjamini and Liu (1999) and Sarkar (2002), or by Simes (1986), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). Only a single null hypothesis is rejected by using these methods for an initial $p$-cut-off of 0.10, corresponding to second-generation non-Western migrants. Balance checks correcting for multiple comparisons are available on request. Standard errors are in parentheses; clustered at household level.

$^\ddagger p < 0.01$.

$^\S p < 0.05$.

$^§§ p < 0.1$.

$^*$Computed by using a one-parameter logistic item response model; estimates are reported in Table 10 in Section 5.

The difference in the predictive power of intentions across the two treatment groups for the ‘no-vote’ outcome.

These unconditional probabilities will be analysed in relation to the actual voting behaviour that is reported in the exit poll. Hence, behaviour is reported immediately after the election rather than observed. Though the aggregate levels of reported turnout or votes for specific parties are susceptible to reporting bias, there is no reason to expect that this affects our comparison between the two randomized treatments. Furthermore, social desirability bias is weaker when surveys are conducted on line, as ours was, rather than face to face (Duffy et al., 2005).
2.3. Descriptive statistics

2.3.1. Actual vote

The Dutch political landscape around the time of the 2017 parliamentary election was very diverse and voters could choose between 28 parties on the ballot. In addition, voters could not show up at a polling station (no vote) or show up but not cast a (valid) vote on any of the parties (the blank option). At the time of our first survey (December 2016), the definitive list of parties on the ballot was not yet known but the parties that were not yet represented in Parliament were not expected to attract many votes. In the survey we therefore listed only the 13 parties that were already represented in Parliament at that time and an option ‘other party’. In the analysis, we combine the six smallest parties among these 13 with the original ‘other party’ and blank options into a larger ‘other’ category. This leads to a multinomial outcome with nine options: ‘no vote’ (or, to be more precise, no show up), a vote on one of the seven largest political parties and ‘other’ (a vote on another party or a blank vote).

Fig. 1 shows where the seven major parties are in ideological space, following the common two-dimensional party characterization of, for example, Marks et al. (2006) and Van Kersbergen and Krouwel (2008). The horizontal axis labelled left–right reflects the economic dimension, expressing the distinction between egalitarian parties that favour extensive redistribution and regulation (left) versus parties with a more laissez-faire ideology (right). The vertical axis shows a non-economic dimension, with parties that favour cultural liberalism and openness at the top (progressive; often labelled ‘Green, alternative and libertarian’), and parties that favour restrictive immigration policies (conservative or ‘traditional, authoritarian, nationalist’) at the top.

![Fig. 1. Dutch political landscape in March 2017 (source: presentation by Dr André Krouwel, http://prezi.com/g7ks9afcnjqn/?utm_campaign=share&utm_medium=copy&rc=ex0share)](image-url)
Table 2. Descriptive statistics of the outcome variable: actual vote in 2017 election†

| Population (%)‡ | Overall | Deterministic sample mean | Probabilistic sample mean | Difference |
|-----------------|---------|---------------------------|---------------------------|------------|
|                 | Mean    | Standard deviation        |                           |            |
| (a) Missing outcome (non-participation in exit poll) |         |                           |                           |            |
| Vote missing    | 0.09    | 0.28                      | 0.09                      | 0.08       | −0.01 (0.008)     |
| N               | 4349    | 2131                      | 2218                      | 4349       |            |
| (b) Dependent variable: actual vote in 2017 elections (0% or 100%) |         |                           |                           |            |
| VVD (liberal)   | 17.4    | 18.3                      | 38.7                      | 18.5       | 18.1        | −0.4 (1.22)       |
| Other party     | 14.1    | 13.9                      | 34.6                      | 13.8       | 14.1        | 0.3 (1.08)        |
| CDA (Christian) | 10.1    | 13.4                      | 34.1                      | 13.7       | 13.2        | −0.5 (1.09)       |
| D66 (progressive liberal) | 10.0   | 12.2                      | 32.7                      | 11.3       | 13.1        | 1.8§ (1.03)       |
| GL (Green)      | 7.5     | 9.7                       | 29.5                      | 9.3        | 9.9         | 0.6 (0.94)        |
| PVV (populist)  | 10.7    | 9.2                       | 28.9                      | 9.6        | 8.8         | −0.8 (0.90)       |
| SP (socialist)  | 7.4     | 9.1                       | 28.8                      | 9.2        | 9.0         | −0.3 (0.89)       |
| PvdA (labour)   | 4.7     | 7.1                       | 25.7                      | 6.8        | 7.4         | 0.6 (0.81)        |
| No vote         | 18.1    | 7.1                       | 25.7                      | 7.7        | 6.5         | −1.3 (0.82)       |
| N               | 3978    | 1936                      | 2042                      | 3978       |            |

†χ²-test for equality of vote distribution across treatments: χ²(8) = 7.02; p-value 0.54. Standard errors are in parentheses; clustered at household level (3275 clusters for the missing DV model, 3027 clusters for DV).
‡Percentage of the population that was eligible to vote (not a percentage of the vote).
§p < 0.1.

bottom. Hence, the economically liberal yet culturally conservative Volkspartij voor Vrijheid en Democratie, ‘VVD’, can be found in the bottom right-hand corner and the progressive leftists of GroenLinks (Green Left, ‘GL’) at the top left-hand side. The Partij Voor de Vrijheid (Freedom Party, ‘PVV’) led by Geert Wilders was the most conservative party in the progressive–conservative dimension, yet its economic ideas are middle of the road.

Table 2 reports descriptive statistics of actual voting behaviour, reported in the exit poll shortly after the election. The actual votes were elicited in the same way in both subsamples and contain no information about which method of eliciting intentions is more accurate. Panel (a) shows that only 9% of panel members who were eligible to vote and participated in the voting intentions survey in December 2016 did not participate in the exit poll. This fraction is almost identical for both treatment groups. Panel (b) compares voting behaviour reported in the exit poll survey with voting behaviour of the complete population. The liberal VVD received the largest share of the vote, 18.3% in the sample and 17.4% in the population. The category ‘other party’ received approximately 14% of the votes, both in the sample and in the population. This category is comprised of many small parties and the blank option.

The Christian Democrat ‘CDA’ is the second largest party in the sample at 13%, followed by the progressively liberal ‘D66’ party at just over 12% (with corresponding population figures around 10% for both). The Greens GL and the Socialist Party ‘SP’ received between 9% and 10% of the votes in the panel, as did the populist PVV. Whereas the Greens and socialists did better in the panel than in the population, the opposite is true for the PVV, which became the second largest party with the support of 10.7% of the eligible voting population. The smallest individual party in our analysis is the labour party ‘PvdA’, with 7% of the votes in the sample and less than 5% in the population.
Table 3. Descriptive statistics of voting intentions (3 months before the election; 0–100%)

| Party          | Deterministic sample | Probabilistic sample |
|----------------|----------------------|----------------------|
|                | Mean | Fraction equal to | Mean | Fraction equal to |
|                | 0    | 1–99 | 100  | 0    | 1–99 | 100  |
| VVD (liberal)  | 13.7 | 0.85 | 0.04 | 0.10 | 13.7 | 0.67 | 0.29 | 0.04 |
| Other party    | 18.1 | 0.78 | 0.09 | 0.12 | 14.6 | 0.60 | 0.36 | 0.04 |
| CDA (Christian)| 7.9  | 0.92 | 0.02 | 0.06 | 9.4  | 0.72 | 0.25 | 0.03 |
| D66 (progressive liberal) | 9.1  | 0.90 | 0.03 | 0.07 | 11.1 | 0.65 | 0.34 | 0.02 |
| GL (Green)     | 7.1  | 0.92 | 0.03 | 0.05 | 8.0  | 0.73 | 0.26 | 0.02 |
| PVV (populist) | 13.9 | 0.84 | 0.06 | 0.10 | 11.5 | 0.75 | 0.20 | 0.05 |
| SP (socialist) | 6.5  | 0.93 | 0.03 | 0.04 | 7.4  | 0.75 | 0.23 | 0.02 |
| PvdA (labour)  | 6.4  | 0.93 | 0.03 | 0.04 | 8.3  | 0.72 | 0.26 | 0.02 |
| No vote        | 17.2 | 0.59 | 0.33 | 0.07 | 16.1 | 0.58 | 0.34 | 0.07 |
| N              | 1936 |     |     |      | 2042 |     |     |      |

With the exception of the PVV, the ranking of parties is the same in the sample as in the population. In contrast, there is a large and salient difference between sample and population when it comes to the proportion that did not vote at all: 18% in the population and only 7% in the sample. To put this discrepancy of 11 PPs in perspective, Delavande and Manski (2010) found that, for the presidential elections in 2008, turnout in the American life panel was 28 PPs higher than in the population. We cannot say whether the high reported turnout in the LISS is due to selection, an effect of panel participation on the likelihood of voting or simply misreporting. For our analysis this is not really relevant, since we compare two randomized treatment groups within the LISS sample. A $\chi^2$-test does not reject the null hypothesis that the voting patterns (columns deterministic and probabilistic in Table 2) are the same for the two groups with different treatments ($p$-value 0.54).

2.3.2. Intentions

Table 3 contains descriptive statistics of voting intentions. It presents means and other summary statistics for the subsamples that received probabilistic and deterministic questions. In the subsample that faced a deterministic choice between parties, only 2–9% of the probabilities were not equal to either 0% or 100%. This is because all respondents in this treatment who reported a 0% or 100% probability of not voting at all automatically received probability 100 or 0 for each party. In contrast, the probabilistic subsample exhibits substantial variation across political parties in the fraction of intermediate probabilities. Only 20% doubt between voting for the populist PVV and some other option, whereas 34% consider voting for the progressive liberals of the D66 party but are not certain yet. On-line appendix OAA presents additional descriptives of the marginal distributions of voting intentions.

To compare voting intentions with actual votes in the aggregate, Fig. 2 combines the overall actual and intended vote shares for all alternatives. The intended vote shares are the sample averages of intentions for both subsamples as reported in Table 3. Actual vote shares are
Fig. 2. Aggregate intentions and actual votes (●, probabilistic intentions; ○, deterministic intentions; no overlap with lines, difference is significant at 10%): ●, no vote; ○, VVD; ●, PVV; ●, CDA; ●, D66; ●, GL; ●, SP; ●, PvdA; ○, other party

averages of indicators for reported votes calculated across the entire sample: the means under ‘overall sample’ in Table 2. Actual votes are on the vertical axis and intentions on the horizontal axis. If predicted and actual vote shares were exactly equal, all circles would lie on the 45° line. This is apparently not so. The probabilistic expectations (full circles) are closer to the diagonal than the deterministic expectations (empty circles) in seven out of nine cases. Moreover, the differences are significant at the 10% level for four of the seven options for which probabilities outperform deterministic statements. (The remaining cases are the VVD party for which probabilistic and deterministic aggregates are almost the same, and the PvdA party where both are on different sides of the 45° line, with the deterministic aggregate somewhat closer to it than the probabilistic forecast.) This suggests that the probabilities give better predictors of aggregate behaviour than the deterministic answers. Aggregating across the nine options, the sum of squared forecast errors is 201.04 for deterministic intentions and 144.48 (28% lower) for probabilistic intentions.

The main focus of the paper is the predictive power of subjective probabilities at the individual rather than the aggregate level. On-line appendix OAB measures these relationships by means of kernel regressions of votes on intentions. We find no difference in predictive performance between elicitation methods for the ‘no-vote’ outcome, which is reassuring since the intention not to vote was elicited in the same way in both samples. Whereas the predictive power for the ‘no-vote’ outcome is weak in both samples, the intention to vote for a party predicts behaviour more strongly. Moreover, probabilities outperform deterministic statements for the parties, especially at high levels of stated intentions. Respondents who report a high probability of voting for a certain party tend to vote for that party more often than those who select that party from the list. The next section formalizes these insights through discrete choice models.
Table 4. Multinomial choice models of actual vote: fixed coefficients †

| Baseline: ‘no vote’ | Results for the following alternatives: |
|---------------------|----------------------------------------|
|                     | VVD | PVV | CDA | D66 | GL | SP | PvdA | Other |
| Intention (%)       | 0.0297‡ | (0.00251) | 0.00456 | 0.00540 | 0.00356 | -0.00100 | 0.00146 | -0.00483 | 0.00798§ | -0.0118‡ |
| Intention × prob. questions | -0.00264 | (0.00366) | | | | | | | | |
| Intention × party   | 0.0186‡ | 0.00833§ | 0.0196† | 0.0184‡ | 0.0203‡ | 0.0177‡ | 0.0198‡ | 0.0155‡ | | |
| Prob. questions     | -0.192 | 0.181 | -0.273 | 0.0383 | -0.0164 | -0.178 | -0.0969 | -0.0268 | | |
| Constant            | 1.286‡ | -0.0441 | 1.308‡ | 0.982‡ | 0.830‡ | 1.046‡ | 0.197 | 1.182‡ | | |
| Observations        | 3978 | | | | | | | | | |
| Log-likelihood      | -5273.95 | | | | | | | | | |

†The dependent variable distinguishes between VVD, PVV, CDA, D66, GL, SP, PvdA and ‘other’. Robust standard errors are in parentheses, clustered at household level (3027 clusters).
‡p < 0.01.
§p < 0.05.
§§p < 0.1.
Table 5. Multinomial choice models of actual vote: mixed logit—indpendent normal mixing distributions†

|                          | Baseline: VVD | Results for the following alternatives: |
|--------------------------|---------------|----------------------------------------|
|                          | No vote       | PVV | CDA | D66 | GL | SP | PvdA | Other |
| **Means of parameters**  |               |     |     |     |    |    |      |       |
| Intention (%)            | 0.0303‡       |     |     |     |    |    |      |       |
|                          | (0.00181)     |     |     |     |    |    |      |       |
| Intention × prob. questions | 0.0266‡     |     |     |     |    |    |      |       |
|                          | (0.00316)     |     |     |     |    |    |      |       |
| Intention × no vote      | −0.00136      |     |     |     |    |    |      |       |
|                          | (0.00367)     |     |     |     |    |    |      |       |
| Intention × prob. quest. × no vote | −0.0345‡   |     |     |     |    |    |      |       |
|                          | (0.00660)     |     |     |     |    |    |      |       |
| Prob. quest. × no vote   | 0.247         |     |     |     |    |    |      |       |
|                          | (0.276)       |     |     |     |    |    |      |       |
| Constant                 | −1.354‡       | −1.235‡ | −0.0540 | −0.334‡ | −0.448‡ | −0.449‡ | −0.949‡ | −0.651‡ |
|                          | (0.195)       | (0.129) | (0.0802) | (0.0849) | (0.0853) | (0.0929) | (0.102) | (0.102) |
| **Standard deviations of parameters** | |     |     |     |    |    |      |       |
| Intention (%)            | 0.00856       |     |     |     |    |    |      |       |
|                          | (0.00705)     |     |     |     |    |    |      |       |
| Intention × prob. questions | 0.0266‡     |     |     |     |    |    |      |       |
|                          | (0.00401)     |     |     |     |    |    |      |       |
| Observations             | 3978          |     |     |     |    |    |      |       |
| Log-likelihood           | −5309.04      |     |     |     |    |    |      |       |

†The dependent variable distinguishes between ‘no vote’, PVV, CDA, D66, GL, SP, PvdA and ‘other’. Robust standard errors are in parentheses, clustered at household level (3027 clusters).
‡p < 0.01.
Multinomial choice models

Multinomial choice models provide a natural modelling approach for actual voting behaviour, with a choice of one alternative out of a set of nine options. Tables 4 and 5 contain estimates of two models. The first (Table 4) is a standard multinomial logit model with fixed coefficients. It contains alternative-specific constants and their interactions with a dummy for the probabilistic sample. Moreover, interactions with deterministic and probabilistic intentions are added, allowing the predictive power of deterministic intentions and the added value of probabilistic intentions to vary across parties. The estimates indicate two main points. Firstly, the intention not to vote significantly predicts not voting and carries the same predictive power in both subsamples. This was expected, as those intentions are elicited by means of the same question in both samples (a placebo test). Secondly and more importantly, for all political parties except one (the PVV party), intentions collected by means of probabilistic questions have significantly larger predictive power than the deterministic questions: the coefficients on the interactions between reported intentions and the probabilistic treatment dummy are always positive, and significant at 1% in all cases except for the PVV party. As expected in light of random assignment to question types, we obtain virtually identical estimates when control variables are added to the model (estimates are available on request).

Panel b. of Table OAC1 (in the on-line appendix) presents estimates for the same model but with probabilistic intentions transformed to match intentions that were elicited in the deterministic treatment. We did not adjust intentions for the ‘no-vote’ alternative, since both samples received the same question on their intention to vote at all. For the parties, we replace the conditional probabilities for the probabilistic sample by 100% for a unique mode, splitting probability mass evenly in case of multiple modes (which occurs for 15% of the observations), and 0% for the other options. The results show that this discretization of probabilistic intentions removes all added value for prediction relative to the deterministic questions. The differences in slopes (i.e. the coefficients on the interactions) are reduced to close to 0 and insignificant for all parties individually as well as jointly (the p-value of the joint test is 0.40, compared with less than 0.0001 for the non-discretized probabilities). This demonstrates that the additional predictive power of subjective probabilities is completely due to the more detailed information that these probabilities provide. If this information is largely removed (largely because, in the multiple-modes case, the transformed probabilities are still more informative than the deterministic intentions), the additional predictive power is lost.

One way to increase the flexibility of the multinomial logit model and to allow for heterogeneous treatment effects is to model key parameters as random coefficients. To keep the number of random coefficients manageable, we assume that the effects of intentions are the same for all parties in both subsamples (but not for ‘no vote’, for which the treatment is a placebo treatment). We estimate a random-coefficients version of the multinomial logit model (which is often called the mixed logit model) with two random coefficients: the coefficient on intentions and the coefficient on the interaction of intentions with the subjective probabilities treatment. Through this specification, we allow the predictive power of intentions in both subsamples to vary across respondents in a parsimonious way. (We also experimented with models with more random coefficients but did not find substantial heterogeneity in other coefficients; estimates are available on request.)

Table 5 presents the estimates for independent normal distributions of the two random coefficients. The mean predictive power of deterministic intentions is similar to that estimated in the fixed coefficients model of Table 4 and the associated standard deviation is small in size and not significantly different from 0. Hence, the model does not indicate substantial heterogeneity
### Table 6. Sample average marginal effect of a 20-PP increase in intent on probability of voting for the party concerned†

| 20-PP increase in intent to vote... | Results for model 3a | | Results for model 3b | |
|-----------------------------------|----------------------|----------------------|----------------------|----------------------|
|                                   | Deterministic | Probabilistic | Difference (PPs)‡ | Difference (%)§ | Deterministic | Probabilistic | Difference (PP)‡ | Differences (%)§ |
| VVD (liberal)                     | 0.11          | 0.15            | 4                   | 36               | 0.10          | 0.19            | 9                   | 90               |
| Other party                       | 0.05          | 0.09            | 4                   | 80               | 0.06          | 0.11            | 5                   | 83               |
| CDA (Christian)                   | 0.11          | 0.14            | 3                   | 27               | 0.10          | 0.18            | 8                   | 80               |
| D66 (progressive liberal)         | 0.07          | 0.12            | 5                   | 71               | 0.08          | 0.14            | 6                   | 75               |
| GL (Green)                        | 0.07          | 0.11            | 4                   | 57               | 0.07          | 0.13            | 6                   | 86               |
| PVV (populist)                    | 0.03          | 0.05            | 2                   | 67               | 0.03          | 0.07            | 4                   | 133              |
| SP (socialist)                    | 0.06          | 0.09            | 3                   | 50               | 0.07          | 0.13            | 6                   | 86               |
| PvdA (labour)                     | 0.05          | 0.07            | 2                   | 40               | 0.04          | 0.08            | 4                   | 100              |
| No vote                           | 0.03          | 0.03            | 0                   | 0                | 0.03          | 0.02            | −1                  | −33              |

‡Difference probabilistic – deterministic.
§Percentage difference (probabilistic – deterministic)/deterministic × 100.

†Example: for the first alternative intentions change from \((0, \frac{1}{8}, \frac{1}{8}, \ldots, \frac{1}{8})\) to \((0.2, 0.1, 0.1, \ldots, 0.1)\).
in the predictive power of deterministic intentions. The positive and significant mean coefficient on the interaction \( \text{intention} \times \text{prob. questions} \) shows that, on average, probabilistic intentions outperform deterministic intentions. Moreover, the estimated standard deviation is significant as well and virtually equal to the mean, implying that, for about 84\% (\( \Phi(0.0266/0.0266) \)) of all respondents, the probability questions indeed provide additional power for predicting whether someone voted for a specific party. In the next subsection, we shall analyse how this heterogeneity in the additional predictive power of subjective probabilities relates to individual characteristics. Probabilistic intentions have no additional predictive power for the ‘no-vote’ option, the placebo: the coefficient on the interaction \( \text{intention} \times \text{prob. quest.} \times \text{no vote} \) is negative and significant, cancelling the difference that was found for the other alternatives.

To facilitate interpretation of the magnitudes of the coefficients in Tables 4 and 5, we report average marginal effects in Table 6. These are calculated as the average increase in the probability of voting for a given party that results from a 20-PP increase in the intention to vote for that party. For each party we compare the situation in which the individual assigns a probability of 0 to vote for this party and \( \frac{1}{3} \) to each other option with that in which (s)he assigns 20\% to this party and 10\% to the other options. Table 6, part (a), uses the multinomial logit estimates of Table 4, allowing the predictive power of deterministic intentions and probabilities to vary across the nine alternatives. The marginal effect of deterministic intentions is weakest for the populist PVV party (3 PP) and strongest for the liberal VVD and Christian CDA parties (11 PP). Probabilities add between 2 PP and 5 PP, which is large relative to the effect of deterministic intentions (effects of probabilistic intentions are 30–80\% larger than those of deterministic intentions). The average marginal effects according to the mixed logit model in Table 5 are qualitatively similar and lead to the same overall conclusion, though the magnitudes of the differences between the two subsamples are sometimes rather different. In particular, probabilities outperform deterministic statements more strongly in absolute (4–9 PP) and relative (75–133\%) terms.

Summarizing, we find clear evidence that subjective probabilities are much better in predicting individual behaviour than are deterministic intentions. This added value is a consequence of the finer response scale which provides additional information, and disappears when probabilities are transformed into modes. There is significant variation in the predictive power of probabilistic intentions, which we analyse in the next section.

4. Heterogeneity in the predictive power of probabilities

The mixed logit model with normal mixing distributions that was presented in Table 5 can be used to back out estimates of the two individual-specific parameters for each respondent. These individual-specific estimates are the posterior means of the random coefficients, conditional on the individual’s reported intentions and actual voting outcome. We are especially interested in the individual-specific parameter on the interaction \( \text{intention} \times \text{prob. quest.} \) for the sample that received probabilistic questions, since this parameter provides a measure of the predictive power of subjective probabilities at the level of the individual.

The posterior means are proxies of the individual-specific parameters. There are two reasons why they are not identical to them. Firstly, the posterior means are calculated from the estimates that are reported in Table 5, and the estimation uncertainty of the mixed logit carries through in subsequent analysis. On-line appendix OAD analyses this source of estimation uncertainty in the individual-level parameters and explains how we account for it. Estimation uncertainty of the mixed logit would disappear if the number of respondents tends to \( \infty \). However, this still leaves the second issue: we observe only a single decision (the actual vote in the election) for each respondent. For any given individual the estimated posterior mean would be a consistent
Fig. 3. Kernel densities of random coefficients in mixed logit model 5b: (a) distribution of posterior means (intention x prob. quest.; intention); (b) by party for which individual voted (VVD, other, CD, D66, GL, PVV, SP, PvdA, no vote).
estimate of their parameter only if the number of observed choices would tend to \( \infty \). In the analysis below, we just use the proxies (the posterior means) at face value and do not try to analyse their deviations from the individual-specific parameters.

Fig. 3 plots the densities of the posterior means for all respondents. Fig. 3(a) shows the distribution of the posterior means of the main effect of intentions and of the interaction of intentions and the probabilistic treatment dummy. We limit the sample to the relevant subsample in both cases: respondents who received the deterministic questions for the main effect and respondents who received the probabilistic treatment for the interaction. As was evident from the estimates in Table 5, there is little heterogeneity in the effect of deterministic intentions but substantial variation in the interaction term. Roughly, the higher posterior means are obtained for those whose actual vote is in line with their intentions and the lower posterior means for those who deviated. The distribution is spread out, however, since not everyone had clear intentions (probabilities are often unequal to 0 or 1) and since the weights of the prior information and the information that is provided by the actual vote depend on party size. The added value of probabilities thus varies across respondents: the density peaks for a coefficient just under 0.04 and it has a heavy left-hand tail. The variation in the posterior means indicates that the combination of intentions with a single vote already provides substantial information on the individual-specific coefficients beyond the marginal information in the mixing distribution.

Fig. 3(b) displays densities for the interaction term separately by the political party that individuals voted for. It shows that the predominant feature of the overall density, the heavy left-hand tail, is evident for the constituents of each party. Though the general shape of the distribution is similar for voters of all parties, the locations differ. For instance, Fig. 3(b) shows that the distribution for respondents who did not vote lies slightly to the right of the others. Importantly, such variation in the posterior means of voters who choose different alternatives reflects features of the amount of information that is carried by the actual vote and the reported intentions in addition to variation in the extent to which intentions are consistent with actual decisions. For a given set of stated intentions, such as a 100% probability on the party that the individual actually voted for, there is significant variation in posterior means across parties. Therefore, we control for the party that someone voted for in some specifications of the model that is discussed below, where the posterior means are analysed in relation to probability numeracy.

On-line appendix OAD analyses estimation error in the posterior means. The results that are reported there indicate that sampling error in coefficients that results from using estimates of the mixed logit is substantial relative to the cross-sectional variation in the point estimates of posterior means. The estimates (and standard errors) of the models below that explain individual means take this uncertainty into account.

4.1. **Probability numeracy and the predictive power of subjective probabilities**

We construct a measure of probability numeracy from a nine-item scale that was administered as part of the 2008 *disease prevention* survey in the LISS panel. Table 7 contains a list of these items as well as the estimates of the item–response model that was used to aggregate them into a single measure for each individual. Unfortunately, we can only construct a numeracy score for half of our sample because of panel refreshments between 2008 and 2017. This loss of data motivates a two-step procedure, in which all available information is first used to generate individual-specific measures of the predictiveness of intentions and numeracy, after which the two are related to each other in a second step. As discussed above, the mixed logit estimates from Table 5 are used to construct posterior means of the predictive power of intentions for each respondent. This uses information from all reported intentions and decisions to estimate
Probabilistic and Deterministic Polling

Table 7. Estimates of item response model used to predict probability numeracy†

| Item | Question                                                                                                                                                                                                 |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (a) Items ranked by increasing difficulty                                                                                                                                                                 |
| Question 1 | If the chance of getting a disease is 10%, how many people out of 100 would be expected to get the disease?                                                                                               |
| Question 2 | If the chance of getting a disease is 10%, how many people out of 1000 would be expected to get the disease?                                                                                               |
| Question 3 | Which of the following represents the biggest risk of getting a disease? 1%; 10%; 5%                                                                                                                |
| Question 4 | Which of the following numbers represents the biggest risk of getting a disease? 1 in 100; 1 in 1000; 1 in 10                                                                                             |
| Question 5 | In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS? |
| Question 6 | If the chance of getting a disease is 20 out of 100, this would be the same as having a ... % chance of getting the disease                                                                                     |
| Question 7 | Imagine that we roll a fair, six-sided die 1000 times. Out of 1000 rolls, how many times do you think the die would come up even (2, 4, or 6)?                                                                 |
| Question 8 | The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?                                                                                |
| Question 9 | In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car? . . . % of the tickets                                              |

Item-specific parameters

| Easy Question | Difficult Question |
|---------------|-------------------|
| 1             | 2                 | 3               | 4               | 5               | 6               | 7               | 8               | 9               |
| Item          | Discrimination    | Difficulty      | Individuals    | Log-likelihood |
|               | 1.742‡ (0.0501)   | −2.079‡ (0.0788) | −1.922‡ (0.0723) | −1.371‡ (0.0535) | −1.142‡ (0.0501) | −1.066‡ (0.0485) | −0.941‡ (0.0481) | −0.571‡ (0.0427) | −0.460‡ (0.0418) | −0.214‡ (0.0409) | 2045 | −8111.44 |

†Robust standard errors are in parentheses, clustered at household level (1589 clusters).
‡p < 0.01.

the prior distribution of random parameters. Similarly, we use all available information on individuals in the 2017 sample to estimate the measurement model for probability numeracy, regardless of treatment assignment. In the second step we relate predictiveness to numeracy using only those observations for which we observe both.

Table 8 displays estimates of linear models that relate probability numeracy to background characteristics and actual voting behaviour. Probability numeracy varies significantly with voting behaviour: respondents who abstain from voting have the lowest average numeracy, followed by the socialists SP and populists PVV. The other constituencies all have higher numeracy. Though education also clearly matters, with the higher educated displaying better numeracy, significant and substantial differences across parties remain if education is controlled for (see the third column of Table 8).

Fig. 4 shows kernel regressions of the posterior means that measure the predictive power of subjective probabilities, the coefficients on intention × prob. questions multiplied by 100, on probability numeracy. We find that the two are positively related and that this association
Table 8. Ordinary least squares regression models of probability numeracy†

| Actual vote (baseline: ‘no vote’) | Probability numeracy |
|----------------------------------|----------------------|
| VVD                              | 0.782‡ (0.122)       |
| Other party                      | 0.642‡ (0.124)       |
| CDA                              | 0.638‡ (0.127)       |
| D66                              | 0.808‡ (0.130)       |
| GL                               | 0.854‡ (0.138)       |
| SP                               | 0.266§§ (0.142)      |
| PVV                              | 0.353§ (0.138)       |
| PvdA                             | 0.783‡ (0.137)       |

| Education (baseline: primary)    | Probability numeracy |
|----------------------------------|----------------------|
| Intermediate secondary           | 0.316§ (0.126)       |
| Higher secondary                 | 0.684‡ (0.142)       |
| Intermediate vocational          | 0.485‡ (0.130)       |
| Higher vocational                | 0.750‡ (0.132)       |
| University                       | 0.929‡ (0.138)       |

| Controls                         | No (1040)            |
| Observations                     | Yes* (1002)          |
| $R^2$                            | 0.082 (0.317)        |

†Cluster robust standard errors are in parentheses (921 and 885 households).
‡$p < 0.01$.
§$p < 0.05$.
§§$p < 0.1$.
*Specification controls for gender, age, net household income, household type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, first-generation Western, first-generation non-Western, second-generation Western, second-generation non-Western).

becomes more pronounced when we control for demographics and for the party that an individual voted for. (Controlling for other covariates is achieved by first regressing probability numeracy on the other covariates and then performing the kernel regression of the posterior mean on the residual of the first regression rather than probability numeracy itself.) The association is non-linear: at low levels of numeracy, increases are associated with a tighter link between choice expectations and actual behaviour whereas the relationship flattens out for middle levels of numeracy and picks up again for those at the top end of the numeracy distribution.

Table 9 contains estimates of linear regression models of individual level posterior means on numeracy, education and other controls (which are not reported). Numeracy enters the model
Fig. 4. Kernel regressions of the predictive power of probabilistic intentions on probability numeracy (Islamic, 95% confidence bands): (a) raw data; (b) controlling for covariates; (c) controlling for covariates and vote
Table 9. Models of mixed logit interaction coefficients†

| Dependent variable: mixed logit coefficient × 100 |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                                  | (1)                             | (2)                             | (3)                             | (4)                             |
| Probability numeracy            | 0.0752                          | 0.124§                          | 0.124§                          | 0.154‡                          |
|                                | (−0.0277; 0.199)                | (2.89 × 10⁻⁴; 0.284)            | (0.0197; 0.255)                 | (0.0295; 0.313)                 |
| Education (baseline: primary)   |                                 |                                 |                                 |                                 |
| Intermediate secondary          | −0.0355                         | −0.00129                        |                                 |                                 |
|                                | (−0.444; 0.354)                 | (−0.378; 0.378)                 |                                 |                                 |
| Higher secondary                | −0.0540                         | 0.0646                          |                                 |                                 |
|                                | (−0.544; 0.408)                 | (−0.388; 0.525)                 |                                 |                                 |
| Intermediate vocational         | −0.0464                         | 0.0320                          |                                 |                                 |
|                                | (−0.462; 0.361)                 | (−0.348; 0.441)                 |                                 |                                 |
| Higher vocational               | 0.0207                          | 0.103                           |                                 |                                 |
|                                | (−0.390; 0.440)                 | (−0.283; 0.528)                 |                                 |                                 |
| University                      | 0.210                           | 0.276                           |                                 |                                 |
|                                | (−0.262; 0.725)                 | (−0.162; 0.798)                 |                                 |                                 |
| Actual vote (baseline: ‘no vote’) |                                 |                                 |                                 |                                 |
| VVD                             |                                 | −1.284‡                         | −1.269‡                         |                                 |
|                                |                                 | (−2.160; −0.521)                | (−2.177; −0.488)                |                                 |
| Other party                     | −1.037‡                         | −1.142‡                         |                                 |                                 |
|                                | (−1.827; −0.359)                | (−2.043; −0.399)                |                                 |                                 |
| CDA                             | −1.423‡                         | −1.473‡                         |                                 |                                 |
|                                | (−2.352; −0.618)                | (−2.488; −0.609)                |                                 |                                 |
| D66                             | −1.271‡                         | 1.500‡                          |                                 |                                 |
|                                | (−2.274; −0.507)                | (2.904; 0.560)                  |                                 |                                 |
| GL                              | −1.294‡                         | −1.395‡                         |                                 |                                 |
|                                | (−2.207; −0.506)                | (−2.419; −0.560)                |                                 |                                 |
| SP                              | −1.539‡                         | −1.572‡                         |                                 |                                 |
|                                | (−2.543; −0.683)                | (−2.653; −0.584)                |                                 |                                 |
| PVV                             | −0.830‡                         | −0.875‡                         |                                 |                                 |
|                                | (−1.611; −0.185)                | (−1.717; −0.189)                |                                 |                                 |
| PvdA                            | −0.960‡                         | 0.846                           |                                 |                                 |
|                                | (−1.769; −0.286)                | (2.636; −0.189)                 |                                 |                                 |
| Controls                        | No                              | Yes§§                          | No                              | Yes§§                          |
| Observations                    | 1002                            | 1002                            | 1002                            | 1002                            |
| $R^2$                           | 0.0036                          | 0.063                           | 0.085                           | 0.14                            |

†95% confidence intervals are in parentheses. Confidence intervals and $p$-values take into account estimation error in the dependent variable (see the on-line appendix OAD).
‡$p < 0.01$.
§$p < 0.05$.
§§Specification controls for gender, age, net household income, household type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, first-generation Western, first-generation non-Western, second-generation Western, second-generation non-Western).

The estimates provide evidence that probability numeracy correlates positively with the extent to which probabilities predict voting. The association between numeracy and the predictive power of probabilities becomes stronger when controlling for education and the party one voted for. This link between numeracy and the predictive power of probabilities is especially striking in light of the 8-year period between the elicitation of numeracy and the collection of voting data. Several robustness checks corroborate the significantly positive correlation by using
both normal and log-normal mixing distributions and models with more random coefficients (on intention × no vote and intention × prob. quest. × no vote; the results are available on request). Although we expect nine items to yield a reliable measure of numeracy, some remaining measurement error cannot be ruled out. If that error is classical, uncorrelated with the true measure, it would bias our estimate for the relationship between predictiveness and numeracy towards 0.

5. Completion time and survey evaluation

To judge whether the additional predictive power of subjective probability questions makes incorporating them in the survey worthwhile, it seems useful also to consider the costs in terms of higher respondent burden. Each data set that is collected in the LISS panel includes variables that measure the time that a respondent spent answering the questions, which can be seen as an approximation of the effort that panel members put into their answers. In Table 10, we compare the time that is taken to complete the survey across the samples that received probabilistic and deterministic questions. Panel (a) presents some percentiles (which are less sensitive to outliers than means and standard deviations), showing that a large majority of the respondents spent between 10 and 45 min on completing the survey, with a median slightly over 15 min. Interestingly, the percentiles are higher for the probabilistic than for the deterministic sample. Using quantile regressions, panel (b) of Table 10 confirms that these differences are statistically significant and do not change much if we control for a wide range of demographics (as expected due to randomly assigned treatment). The probabilistic questions caused respondents to take 1 min longer at the first quartile, 2 min at the median and 3 min at the third quartile. Apparently, respondents tend to put in more effort to report probabilities than they do to select a single party.

| Table 10. Descriptive statistics and regression models of the time to complete the survey† |
| --- |
| **N** | **Percentiles (min)** |
|  | p10 | p25 | p50 | p75 | p90 |
| (a) Descriptive statistics of time to complete survey | | | | | |
| Deterministic | 1933 | 9.7 | 12.2 | 16.2 | 22.8 | 38.2 |
| Probabilistic | 2037 | 10.3 | 13.3 | 17.6 | 25.1 | 46.4 |
| Quantile regressions | | | | | |
| Probabilistic questions | 0.72§ | 1.05§ | 1.81§ | 2.70§ | 7.31§§ |
| (0.214) | (0.218) | (0.271) | (0.495) | (4.243) |
| Controls | Yes | Yes | Yes | Yes | Yes |
| **N** | 3725 | 3725 | 3725 | 3725 | 3725 |

†Standard errors are in parentheses; clustered at household level (2867 clusters).
‡Specification controls for gender, age, education, net household income, household type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, first-generation Western, first-generation non-Western, second-generation Western, second-generation non-Western).
§p < 0.01.
§§p < 0.1.
Table 11. Question difficulty†

| Questions were... | ... difficult | ... clear | ... thought provoking | ... interesting | ... enjoyable |
|-------------------|--------------|----------|----------------------|----------------|-------------|
| (a) Descriptive statistics of survey evaluation |             |          |                      |                |             |
| 1, certainly not  | 0.47         | 0.01     | 0.10                 | 0.03           | 0.02        |
| 2                 | 0.19         | 0.02     | 0.11                 | 0.04           | 0.04        |
| 3                 | 0.17         | 0.14     | 0.36                 | 0.29           | 0.32        |
| 4                 | 0.12         | 0.35     | 0.27                 | 0.36           | 0.32        |
| 5, certainly yes | 0.05         | 0.48     | 0.17                 | 0.28           | 0.30        |
| N                 | 3970         | 3970     | 3970                 | 3970           | 3970        |

(b) Ordered logit models of survey evaluation‡

| Probabilistic questions | 0.046 | −0.072 | −0.075 | −0.108§ | −0.072 |
|-------------------------|-------|--------|--------|---------|--------|
| (0.062)                 | (0.063)| (0.059)| (0.061)| (0.0613)|        |

| Controls                | Yes   | Yes    | Yes    | Yes     | Yes    |
|-------------------------|-------|--------|--------|---------|--------|
| N                       | 3725  | 3725   | 3725   | 3725    | 3725   |

| Log-likelihood          | −5044.67 | −4138.71 | −5499.29 | −4818.51 | −4740.34 |

†Standard errors are in parentheses; clustered at household level (2867 clusters).
‡Specification controls for gender, age, education, net household income, household type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, first-generation Western, first-generation non-Western, second-generation Western, second-generation non-Western).
§p < 0.1.

At the end of each survey, the LISS routinely asks some diagnostic questions about the perceived difficulty and the extent to which respondents enjoyed filling out the questionnaire. We compared the answers across treatments but did not find any significant differences between the two treatments; see Table 11 for details. (This result might be due to the order of the questions; voting intentions (the only difference between the treatments) were located 55th and 56th among 170 items.) We can therefore conclude that, even though the subjective probabilities required some additional effort, the respondents did not find the probabilistic survey substantially more difficult, less interesting or less enjoyable than the deterministic survey.

6. Conclusion

This paper looks at the predictive power of subjective probability questions in surveys that elicit how survey respondents intend to behave in the future. In particular, we use the context of parliamentary elections in the Netherlands to relate individuals’ intentions to vote for different parties, elicited 3 months before the election, to how they actually voted. We exploit experimental variation in the question format that was used to measure intentions and to compare deterministic items, in which respondents choose a single party as their best prediction, with probabilistic questions that allow individuals to express uncertainty and doubt. Such a probabilistic approach to polling has been applied to US presidential elections since 2008, but we are the first to compare the predictive power of probabilistic and deterministic intentions by using a large split-sample design. Moreover, whereas US elections are contests between two parties, the Dutch elections ask voters to choose between many more options. Our outcome variable distinguishes between nine possibilities: the seven major parties, the option to vote for any other party or to vote blank, and the option not to vote at all. The multiparty nature of the Dutch political system creates scope for using probabilities to express undecidedness and, indeed, 70% of the respondents assign positive probabilities to more than one party. Such information is not
contained in deterministic voting intentions, and our main research question is whether (and, if so, for which respondents) it has added value for predicting individual behaviour.

Comparing average intentions and actual votes at the aggregate level, we find that probabilities are closer to realized behaviour for seven out of nine options and the difference between question formats is significant for four of these. Our main finding is that, at the level of the individual, subjective probabilities are substantially better predictors of actual voting than deterministic intentions. This follows from non-parametric regressions as well as multinomial choice models. In multinomial and mixed logit models, the estimated average marginal effect of a probabilistic prediction is between 27% and 133% higher than that of a deterministic question for the same party. These benefits of probabilities over deterministic answers are largely due to the additional information that is contained in the probabilities: if probabilities are first discretized to resemble the deterministic answers (the party with the largest probability), their predictive performance is very similar to (and not significantly different from) that of the deterministic questions.

We use the estimates of a mixed logit model to approximate the predictive power of probabilistic intentions for each individual in the subsample that received those questions. Using normal mixing distributions we find that subjective probabilities provide additional predictive power for actual behaviour for a large majority (84%) of the respondents. There is a significantly positive association between the extent to which probabilities add to the predictive power and probability numeracy as measured by using a nine-item scale. This must be a very persistent relationship, since the measure of probability numeracy is constructed based on data from 2008 whereas the poll data were collected more than 8 years later. The result is in line with earlier studies demonstrating heterogeneity in the ability to work with probabilities and their predictive value. In contrast, we find that the predictive power of deterministic intentions hardly varies across the sample.

Probabilistic questions may require more effort from respondents: at the median those who answered them took almost 2 min longer to complete the survey than respondents who were asked to choose a single party did. Nonetheless, both formats yield similar rates of item non-response and there are no significant differences between the two treatments in the evaluations of the difficulty or attractiveness of the survey.

Comparing benefits and costs, our main overall conclusion is that subjective probabilities should be preferred over deterministic questions when one aims to predict behaviour at the level of the individual. They work well when predicting choices from a discrete menu of options. Future research can investigate whether the benefits that are observed for the application of voting extend to other domains.

Acknowledgements

This work is part of the research programme ‘Innovational research incentives scheme veni’ with project number 451-15-018, which is (partly) financed by the Netherlands Organisation for Scientific Research. We thank the Associate Editor and a referee, Rob Alessie, Tabea Bucher-Koenen, Joachim Winter, the audience at the research workshop ‘Empirical economics’ at the Munich Center for the Economies of Aging and seminar participants at the University of Aarhus for useful comments and helpful discussions.

References

Armantier, O., Bruine de Bruin, W., Topa, G., Van der Klaauw, W. and Zafar, B. (2015) Inflation expectations and behavior: do survey respondents act on their beliefs? Int. Econ. Rev., 56, 505–536.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, 57, 289–300.

Benjamini, Y. and Liu, W. (1999) A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Planing Inf.*, 82, 163–170.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29, 1165–1188.

Binswanger, J. and Salm, M. (2017) Does everyone use probabilities?: The role of cognitive skills. *Eur. Econ. Rev.*, 98, 73–85.

Bruine de Bruin, W. and Carman, K. G. (2012) Measuring risk perceptions: what does the excessive use of 50% mean? *Med. Decsn Makng*, 32, 232–236.

Carman, K. G. and Kooreman, P. (2014) Probability perceptions and preventive health care. *J. Risk Uncertainty*, 49, 43–71.

Delavande, A. and Manski, C. F. (2010) Probabilistic polling and voting in the 2008 presidential election evidence from the American life panel. *Publ. Opin. Q.*, 74, 433–459.

Deville, J.-C. and Tillé, Y. (2004) Efficient balanced sampling: the cube method. *Biometrika*, 91, 893–912.

De V os, K. (2010) Representativeness of the LISS-panel 2008, 2009, 2010. *Mimeo. Center for Economic Research*, Tilburg.

Duffy, B., Smith, K., Terhanian, G. and Bremer, J. (2005) Comparing data from online and face-to-face surveys. *Int. J. Markt Res.*, 47, 615–639.

Gutsche, T. L., Kapteyn, A., Meijer, E. and Weerman, B. (2014) The Rand continuous 2012 presidential election poll. *Publ. Opin. Q.*, 78, spec! iss. 1, 233–254.

Holland, B. S. and Copenhaver, M. D. (1987) An improved sequentially rejective Bonferroni test procedure. *Biometrics*, 43, 417–423.

Hurd, M. (2009) Subjective probabilities in household surveys. *A. Econ. Rev.*, 1, 543–562.

Kleinjans, C. and Van Soest, A. (2014) Non-response and focal point answers to subjective probability questions. *J. Appl. Econometr.*, 29, 567–585.

Maas, K., Steenbergen, M. and Saris, W. (1990) Vote probabilities. *Elect. Stud.*, 9, no. 2, 91–107.

Manski, C. F. (2004) Measuring expectations. *Econometrika*, 75, 1329–1376.

Manski, C. and Molinari, F. (2010) Rounding probabilistic expectations in surveys. *J. Bus. Econ. Statist.*, 28, 219–231.

Marks, G., Hooghe, L., Nelson, M. and Edwards, E. (2006) Party competition and European integration in the East and West. *Compari. Polit. Stud.*, 39, 155–175.

Meier, K. J. (1980) Rationality and voting: a Downsian analysis of the 1972 election. *Westrn Polit. Q.*, 33, 38–49.

Meier, K. J. and Campbell, J. E. (1979) Issue voting: an empirical examination of individually necessary and jointly sufficient conditions. *Am. Polit. Q.*, 7, 21–50.

Sarkar, S. K. (2002) Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30, 239–257.

Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73, 751–754.

Tillé, Y. (2011) Ten years of balanced sampling with the cube method: an appraisal. *Sure. Methodol.*, 37, 215–226.

Van Kersbergen, K. and Krouwel, A. (2008) A double-edged sword!: The Dutch centre-right and the ‘foreigners issue’. *J. Eur. Publ. Poly*, 15, 398–414.

Van der Laan, J. (2009) Representativity of the LISS panel. *Discussion Paper 09041*. Statistics Netherlands, Voorburg.