Supplementary Material

We present all additional figures and tables for the document “Estimating the Recommendation Certainty in Candidate-based Voting Advice Applications” in the order in which they are referenced.

Table 3. Party abbreviations, translations, political positions.

Abbreviation	Full German Name	Official English Translation	Political Position	Seats
SP	Sozialdemokratische Partei der Schweiz	Social Democratic Party	Left	8
Greens	Grüne Partei der Schweiz	Green Party	Left	4
GLP	Grünliberale Partei der Schweiz	Green Liberal Party	Left-Liberal	4
EVP	Evangelische Volkspartei der Schweiz	Evangelical People's Party	Centrist	1
Centre	Die Mitte	The Centre	Centrist	3
FDP	Freisinnig-Demokratische Partei	The Liberals	Liberal-Right	5
EDU	Eidgenössisch-Demokratische Union	Federal Democratic Union	Right-Conservative	1
SVP	Schweizerische Volkspartei	Swiss People's Party	Right	10

Notes: The selected parties represent the canton of Zurich in the Swiss National Council. Abbreviations, party names and English translations follow commonly used terminology. Political positions reflect each party’s median placement on the Smartmap, which situates parties on a data-driven ideological space based on policy responses. Seat counts refer to the 2023–2027 legislative period and are drawn from the Swiss Federal Statistical Office (2023).

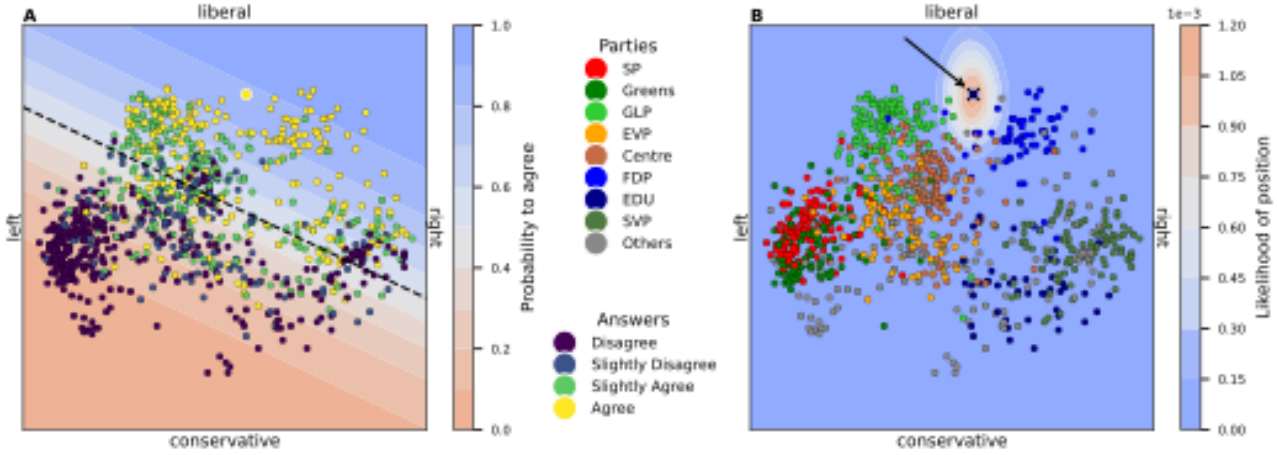

Figure 9. Statistical model trained with the candidates’ data. Notes: (A) The left panel shows all National Council candidates from the canton of Zurich positioned in the two-dimensional ideological space learned by the statistical model. Each point corresponds to one candidate, colored by their recorded stance on increasing the retirement age. The background shading represents the model-estimated probability of agreeing with the proposal across the ideological space. (B) The right panel shows the same candidate positions, but colored by party affiliation. The axes are labelled a posteriori to reflect the dominant structure of the learned latent space. The background contour visualizes the posterior uncertainty of the inferred ideological position for one specific FDP candidate. Warmer areas indicate regions of higher posterior density.

Table 4. Fluctuation, error, and duration for different estimation algorithms and k=10.

Algorithm	Method	Fluctuation (%)	Error (%)	Duration (ms)	Recommendation
CRA	Ground truth	4.37 ± 0.03			Only a-posteriori
Historic	User-agnostic	1.30%	11.44 ± 0.13	<1ms	Requires user data
Linear	User-agnostic	1.33%	11.63 ± 0.13	<1ms	Linear progress bar
One-shot	Maximum-likelihood	6.00 ± 0.04	10.23 ± 0.10	24.3 ± 0.4	Includes estimated recs.
Posterior	1 sample	9.20 ± 0.05	10.46 ± 0.09	26.2 ± 0.6	Includes sampled recs.
Posterior	10 samples	5.11 ± 0.02	9.17 ± 0.09	25.1 ± 0.6	Includes sampled recs.
Posterior	100 samples	4.26 ± 0.02	9.00 ± 0.09	28.5 ± 0.6	Includes sampled recs.
Posterior	1000 samples	4.13 ± 0.02	8.98 ± 0.09	61.0 ± 0.6	Includes sampled recs.

Notes: The fluctuation is computed as the sum of absolute successive differences. A higher value means the estimations are less smooth. The error gives the mean absolute difference between the estimated CRA and the true CRA. The duration includes computing the estimated CRA and stable recommendations. All values are means and their standard errors.

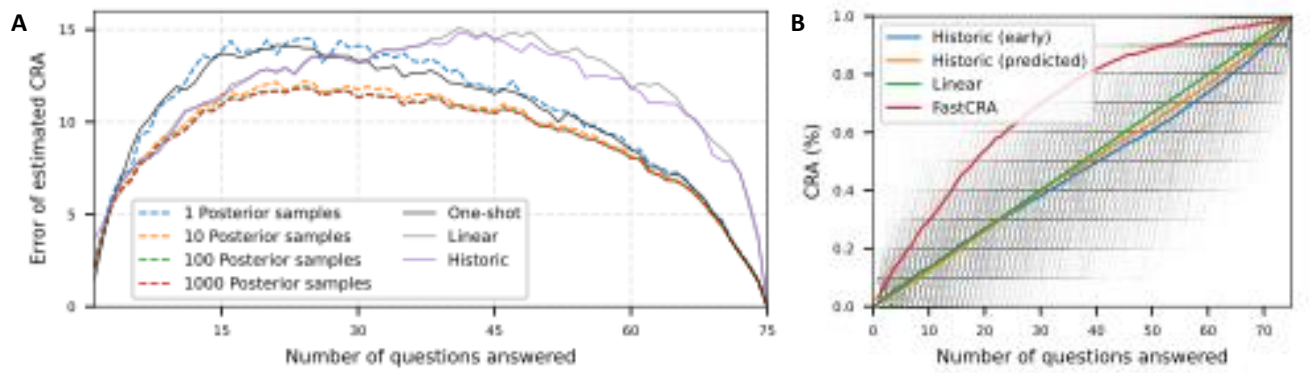

Figure 10. Mean absolute error of the estimation algorithms compared to the true CRA for k=10. Notes: (A) For all estimation algorithms, the error increases from 2% to around 12-14% after fifteen questions. The user-agnostic predictions, Linear and Historic, perform worse than Posterior estimates. (B) Each voter has an individual progression of the true CRA (gray lines). The Historic estimates correspond to the average CRA across all voters.

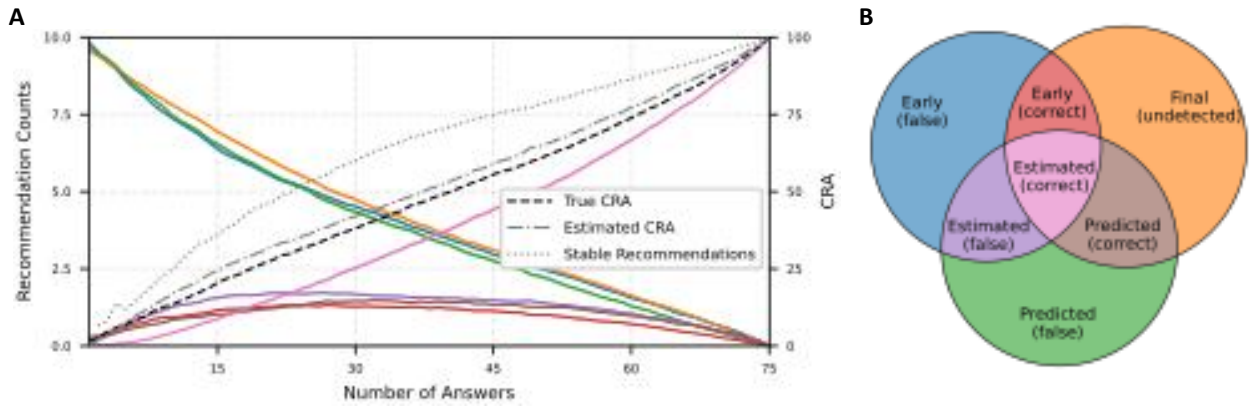

Figure 11. Venn diagram of the early, predicted, and final recommendations for k=10. Notes: (A) The blue, green, and orange lines show the number of candidates that are only in one of the three sets of recommendations. They decrease from 10 (when all sets are distinct) to 0 (when all sets are identical). The pink line shows the intersection of all three sets. The dotted line shows the fraction of stable recommendations in the estimated recommendations. (B) In the corresponding Venn diagram, the true positives (correct) and false positives (false) are colored according to the lines in the graph.

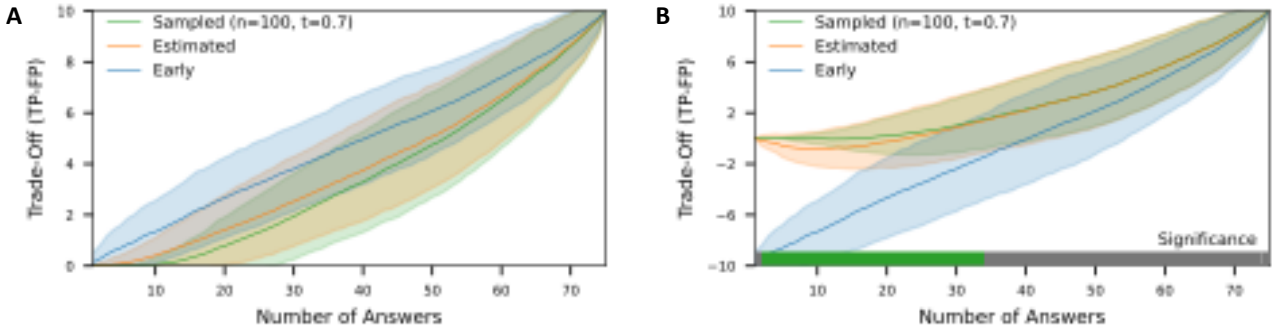

Figure 12. The precision of the stable recommendations for three different selection algorithms for k=10. Notes: (A) The y-axis shows the number of true positives per question across all voters (mean and standard deviation). The Early recommendations constantly have a higher score than the Sampled and Estimated ones. (B) The y-axis shows the difference between true and false positives per question across voters (mean and standard deviation). The Early recommendations start with many false positives. The horizontal bar shows where the difference between Estimated (orange) and Sampled (green) recommendations is statistically significant.

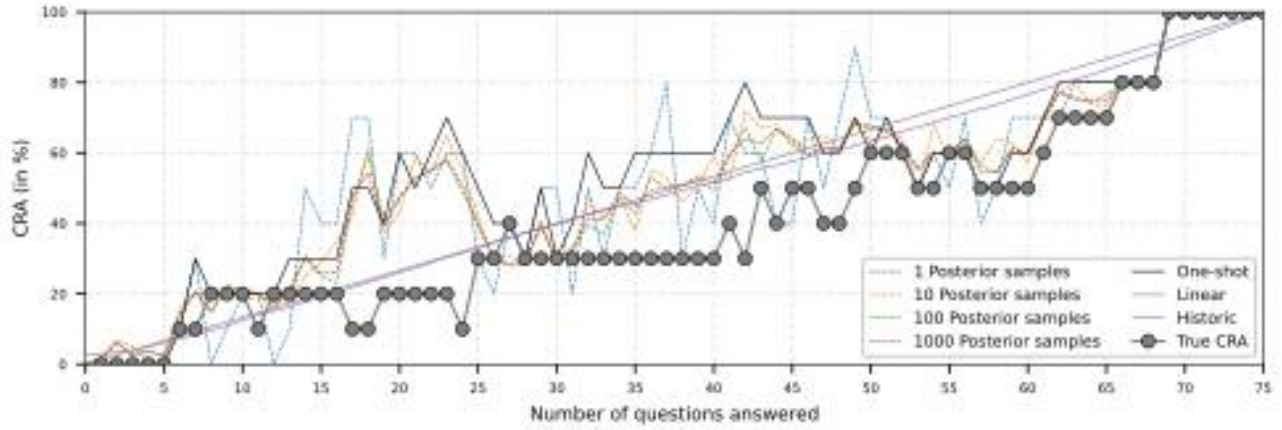

Figure 13. Progression of the true CRA compared to different estimation algorithms for user 3613 and k=10. Notes: The gray dots show the true CRA. The colored lines indicate the estimated CRA for the different algorithms.

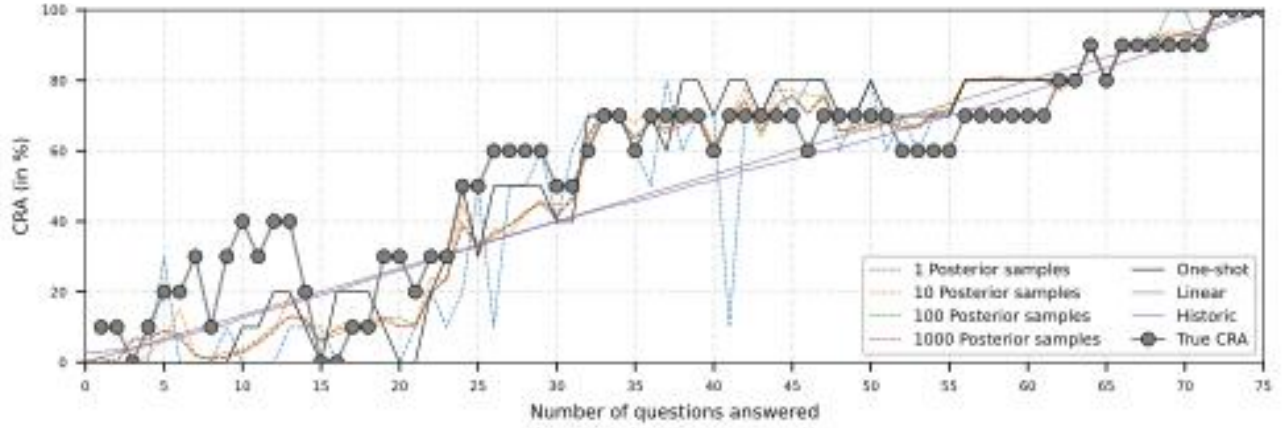

Figure 14. Progression of the true CRA compared to different estimation algorithms for user 19274 and k=10. Notes: The gray dots show the true CRA. The colored lines indicate the estimated CRA for the different algorithms.

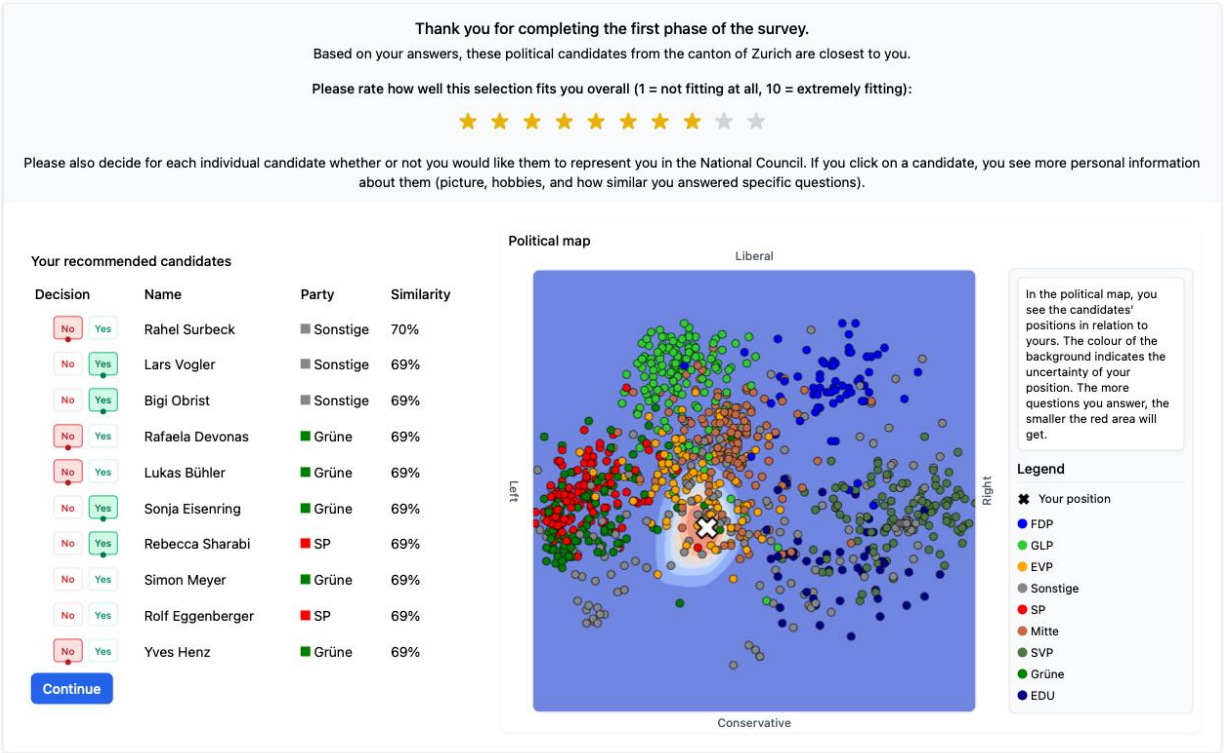

Figure 15. Candidate recommendation page of our VAA platform. Notes: On the recommendation page, users select which candidates seem relevant to them. The similarity scores indicate the distance of their answers to the user's answers. The political map on the right locates candidates' and the user's ideology in a two-dimensional political space.

Table 5. Pre-survey of the user experiment (screening questions).

Key	Question	Translation	Response type
Zurich	Sind Sie im Kanton Zürich stimmberechtigt?	Are you eligible to vote in the canton of Zurich?	Single Choice
Age	Wie alt sind Sie?	How old are you?	Numerical
Gender	Was ist Ihr Geschlecht?	What is your gender?	Single choice
Location	In was für einem Gemeindetyp wohnen Sie?	What type of municipality do you live in?	Single choice
Education	Was ist Ihr höchster Bildungsabschluss?	What is your highest level of education?	Single choice
Party	Welcher Partei stehen Sie am nächsten?	Which political party are you closest to?	Single choice
Politics	Wie stark interessieren Sie sich für Politik?	How interested are you in politics?	Numerical
Smartvote	Wie oft haben Sie bereits das Online-Tool "Smartvote" benutzt, um sich vor einer Wahl über Kandidierende zu informieren?	How often have you used the online tool "Smartvote" to inform yourself about candidates before an election?	Numerical

Notes: For single choice, users could select from a predefined list. A numerical slider was used for the other questions.

Table 6. Post-survey of the user experiment (Likert-type questions).

Key	Question	Translation	Response
To what extent do you agree with these statements about the website (questionnaire and recommendations) overall?			
Platform_1	Die empfohlenen Kandidaten spiegelten meine politische Meinung wider.	The recommended candidates reflected my political opinion.	6-Likert
Platform_2	Mit der Webseite kann man sich gut über die Kandidierenden informieren.	With this platform you can inform yourself well about the political candidates.	6-Likert
Platform_3	Ein solcher Fragebogen ist hilfreich, um passende Empfehlungen zu erhalten.	The questionnaire was helpful to get suitable recommendations.	6-Likert
Platform_4 (Usage)	Ich würde diese Webseite einem Familienmitglied oder Freund empfehlen.	I would recommend this platform to a friend or family.	6-Likert
Only for the treatment groups: *During the questionnaire you saw a prediction of the certainty of your recommendations [include screenshot of their condition]. To what extent do you agree with the following statements about the displayed certainty?*			
Certainty_1 (Understanding)	Ich habe die Bedeutung dieser Prognose verstanden.	I understood the meaning of this prediction.	6-Likert
Certainty_2	Die Prognose wirkte für mich konsistent über die Fragen hinweg.	The prediction seemed consistent to me across questions.	6-Likert
Certainty_3 (Relevance)	Ich habe diese Prognose als interessant empfunden.	I found this prediction interesting.	6-Likert
Certainty_4 (Accuracy)	Ich habe diese Prognose als zutreffend empfunden.	I found this prediction accurate.	6-Likert
Certainty_5 (Influence)	Die Prognose hat beeinflusst, nach wie vielen Fragen ich mir die Empfehlungen habe anzeigen lassen.	The prediction influenced how many questions I answered before viewing the recommendations.	6-Likert
Only for the control group: *Imagine you saw a prediction of the certainty of your recommendations during the questionnaire [include a screenshot of each interface: certainty bar (A) and stable recommendations (B)]. To what extent do you agree with the following statements about the displayed certainty?*			
Control_A1	Ich verstehe diese Prognose der Zuverlässigkeit.	I understand this certainty estimate.	6-Likert

Control_A2	Ich fände so eine Prognose der Zuverlässigkeit interessant.	I would find such a certainty estimate interesting.	6-Likert
Control_A3	Ich kann mir vorstellen, dass so eine Prognose der Zuverlässigkeit zutreffend ist.	I can imagine that such a certainty estimate is accurate.	6-Likert
Control_B1	Ich verstehe diese Anzeige der wahrscheinlichen Empfehlungen.	I understand this display of stable recommendations.	6-Likert
Control_B2	Ich fände so eine Anzeige von wahrscheinlichen Empfehlungen interessant.	I would find such a display of stable recommendations interesting.	6-Likert
Control_B3	Ich kann mir vorstellen, dass so eine Anzeige von wahrscheinlichen Empfehlungen zutreffend ist.	I can imagine that such a display of stable recommendations is accurate.	6-Likert

Notes: 6-Likert scales were labeled from "Strongly disagree" (-3), "Disagree" (-2), "Somewhat disagree" (-1), "Somewhat agree" (1), "Agree" (2), and "Strongly agree" (3). Users in the control group evaluated all different interfaces in the post survey while the other users only evaluated the interfaces they had seen during the experiment.

Table 7. Post-survey of the user experiment (open-ended questions).

Key	Question	Translation	Response type
Only for the users with less than 75 questions answered: *You answered only [insert their number of answers] out of 75 questions in the questionnaire before requesting your recommendations. Why didn't you complete the entire questionnaire?*			
Option_L1	Ich wusste ohnehin schon, wen ich wählen würde.	I already knew whom I would vote for.	Multiple choice
Option_L2	Ich fand den Fragebogen langweilig/unpassend.	I found the questionnaire boring/inappropriate.	Multiple choice
Option_L3	Es gab keinen Anreiz für mich, ihn weiter auszufüllen.	There was no incentive for me to continue filling it out.	Multiple choice
Option_L4	Ich war schon fast fertig.	I was almost finished anyway.	Multiple choice
Option_L5	Ich habe auf die Prognose geachtet.	I paid attention to the prediction.	Multiple choice
Option_L6	Die Zuverlässigkeit der Empfehlungen war ausreichend.	The certainty of the recommendations was sufficient.	Multiple choice
Option_L7	Ein anderer Grund	Another reason	Open-ended
Only for the users with all 75 questions answered: *You answered 75 out of 75 questions in the questionnaire before the recommendations were displayed. Why didn't you request them earlier and end the questionnaire incompletely?*			
Option_F1	Ich habe nicht gewusst, dass das möglich ist.	I didn't know that was possible.	Multiple choice
Option_F2	Ich war neugierig auf die verbleibenden Fragen.	I was curious about the remaining questions.	Multiple choice
Option_F3	Ich wollte die Empfehlungen möglichst genau machen.	I wanted the recommendations to be as accurate as possible.	Multiple choice
Option_F4	Ich wollte sicherstellen, dass ich alles ausfülle, um die volle Entlohnung von Bilendi zu erhalten.	I wanted to make sure I completed everything to receive the full compensation from Bilendi.	Multiple choice
Option_F5	Ich war ohnehin schon fast fertig.	I was almost finished anyway.	Multiple choice
Option_F6	Ich habe auf die Prognose geachtet.	I paid attention to the certainty.	Multiple choice
Option_F7	Ein anderer Grund	Another reason	Open-ended
Please describe any final thoughts you had about the recommendations, the certainty, or the questionnaire.			
Thoughts	Bitte beschreiben Sie abschließend weitere Gedanken, die Sie zu den Empfehlungen, den Prognosen oder dem Fragebogen hatten.	Please describe any final thoughts you had about the recommendations, the certainty, or the questionnaire.	Open-ended

Notes: Depending on the user's point of dropout, the framing of the questions varied slightly. Users could add their own response to the multiple choice list, or provide their thoughts in the open-ended feedback.

Table 8. Statistics of the open-ended feedback in the post-survey.

Label	Positive	Negative	Examples
Clarity of certainty explanation	3	4	*Ich […] fand die Prognose am Anfang etwas verwirrend, dacht sie zeigt an wie sehr ihr mir mit meinen Aussagen vertrauen könnte/wie fest ihr beurteilen könnt, das ich wählen kann.*
General VAA experience	25	2	*Gute Neuerung für Halb-Interessierte. Für mich persönlich nicht nötig.*
Length of questionnaire	0	2	*Etwas lange aber gut gemacht.*
Ambiguity of questions	0	8	*Manchmal wäre es hilfreich gewesen, meine Auswahl in einem kurzen Bemerkungsfeld erläutern zu können.*
Learning through the tool	12	4	*Es hat mich als nicht stark politisch interessierte Person überrascht, welche Personen mir empfohlen wurden.*
Display of the evaluation page	5	3	*Einfache, kurze Fragen, übersichtliche Darstellung der Ergebnisse/Empfehlungen.*

Notes: Six themes arose from the survey feedback. Each response can be applicable to multiple themes. Responses were labeled positive, negative, or neutral towards each theme.

Table 9. Simulation results for the optimal threshold.

Method	Samples	Threshold	Q=15	Q=30	Q=45	Q=60	Q=75
Early			-21.62 ± 0.18	-14.17 ± 0.19	-7.84 ± 0.20	-2.02 ± 0.20	4.05 ± 0.18
Predicted			-20.66 ± 0.26	-12.28 ± 0.29	-5.54 ± 0.30	0.46 ± 0.29	6.34 ± 0.25
Estimated			-2.20 ± 0.10	0.91 ± 0.13	4.42 ± 0.15	8.14 ± 0.16	12.44 ± 0.15
Sampled	1	20%	-1.74 ± 0.09	0.83 ± 0.12	4.17 ± 0.15	7.87 ± 0.16	12.21 ± 0.15
Sampled	1	50%	-1.74 ± 0.09	0.83 ± 0.12	4.17 ± 0.15	7.87 ± 0.16	12.21 ± 0.15
Sampled	1	70%	-1.74 ± 0.09	0.83 ± 0.12	4.17 ± 0.15	7.87 ± 0.16	12.21 ± 0.15
Sampled	1	80%	-1.74 ± 0.09	0.83 ± 0.12	4.17 ± 0.15	7.87 ± 0.16	12.21 ± 0.15
Sampled	1	90%	-1.74 ± 0.09	0.83 ± 0.12	4.17 ± 0.15	7.87 ± 0.16	12.21 ± 0.15
Sampled	10	20%	-2.78 ± 0.12	0.26 ± 0.15	3.87 ± 0.16	7.68 ± 0.17	12.06 ± 0.15
Sampled	10	50%	-0.22 ± 0.07	2.03 ± 0.12	5.16 ± 0.14	8.67 ± 0.16	12.87 ± 0.14
Sampled	10	70%	0.15 ± 0.04	2.12 ± 0.09	5.12 ± 0.13	8.61 ± 0.14	12.81 ± 0.13
Sampled	10	80%	0.16 ± 0.02	1.85 ± 0.07	4.77 ± 0.11	8.28 ± 0.13	12.54 ± 0.13
Sampled	10	90%	0.08 ± 0.01	1.24 ± 0.05	3.93 ± 0.09	7.47 ± 0.12	11.86 ± 0.12
Sampled	100	20%	-3.60 ± 0.13	-0.34 ± 0.16	3.44 ± 0.17	7.35 ± 0.17	11.80 ± 0.16
Sampled	100	50%	-0.07 ± 0.08	2.21 ± 0.13	5.33 ± 0.15	8.83 ± 0.16	13.00 ± 0.15
Sampled	100	70%	0.28 ± 0.04	2.38 ± 0.10	5.39 ± 0.13	8.85 ± 0.15	13.01 ± 0.14
Sampled	100	80%	0.21 ± 0.03	2.14 ± 0.08	5.13 ± 0.12	8.61 ± 0.14	12.82 ± 0.13
Sampled	100	90%	0.10 ± 0.01	1.57 ± 0.06	4.48 ± 0.10	8.03 ± 0.13	12.33 ± 0.12
Sampled	1000	20%	-3.62 ± 0.13	-0.38 ± 0.16	3.40 ± 0.17	7.31 ± 0.17	11.77 ± 0.16
Sampled	1000	50%	-0.00 ± 0.08	2.25 ± 0.12	5.36 ± 0.15	**8.85 ± 0.16**	**13.02 ± 0.15**
Sampled	1000	70%	**0.29 ± 0.04**	**2.39 ± 0.10**	**5.40 ± 0.13**	**8.85 ± 0.15**	**13.02 ± 0.14**
Sampled	1000	80%	0.21 ± 0.03	2.13 ± 0.08	5.11 ± 0.12	8.59 ± 0.14	12.80 ± 0.13
Sampled	1000	90%	0.09 ± 0.01	1.55 ± 0.06	4.46 ± 0.10	8.00 ± 0.13	12.31 ± 0.12

Notes: The difference of true and false positives is computed at five different stages in the simulation (after 15, 30, 45, 60, and 75 questions). Four algorithms to identify stable recommendations are compared (Early, Predicted, Estimated, and Sampled). For 1000 sampled recommendations and a threshold of 70%, the difference of false and true positives is maximal.

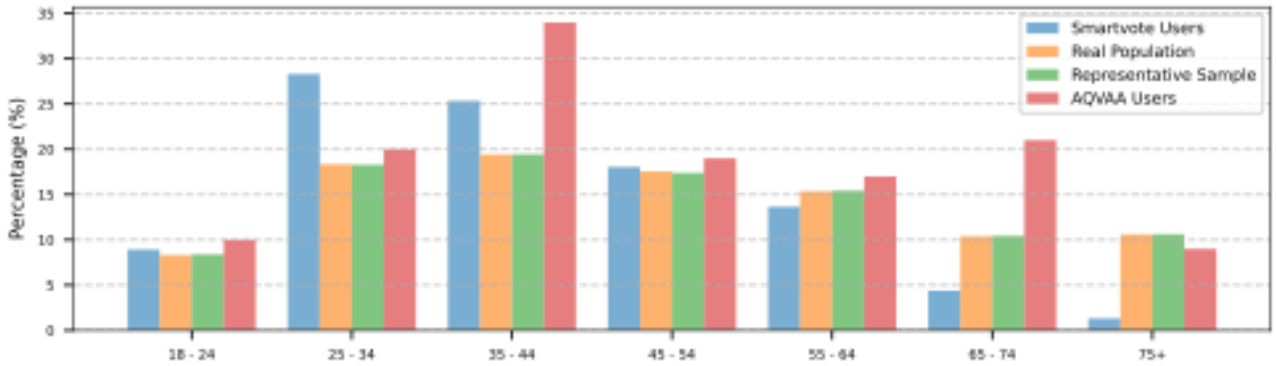

Figure 16. Distribution of voters across age groups. Notes: The overall users of the *Smartvote* VAA (blue) are younger than the real population from the Canton of Zurich (orange). The representative sample of voters in the *Smartvote* data (green) reduces the number of users such that the demographic information corresponds to the real distribution. The participants in the user experiment (red) are older than the real population.

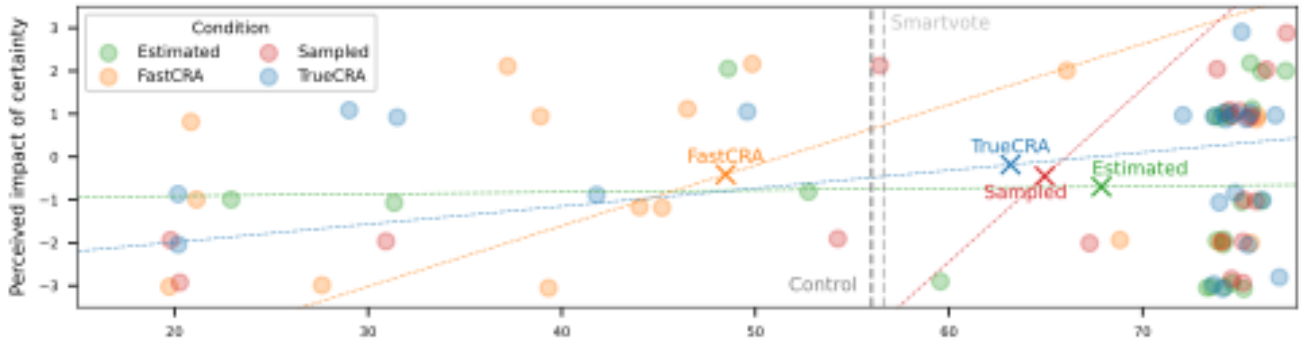

Figure 17. Number of answered questions for users with different interfaces. Notes: Users in the Control group finish the questionnaire on average after 56 questions. Users in FastCRA finish after 49 questions. Users in all other conditions (TrueCRA, Sampled, and Estimated) finish after 63 to 68 questions. The y-axis shows users' perceived impact of the displayed recommendation certainty on their number of answered questions. For Sampled there is a positive correlation between these variables (r=0.403, p=0.078).

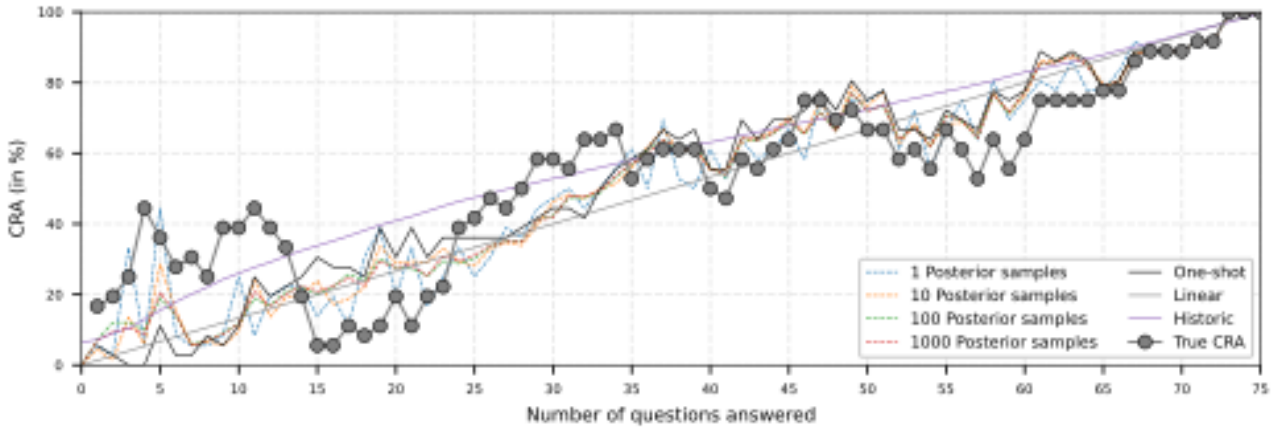

Figure 18. Progression of the true CRA for user 19274 and k=36. Notes: The gray dots show the true CRA. While initially high, the CRA suddenly drops below 5% after 14 questions. Then, the increase in CRA is moderate and equally well approximated by posterior-based samples and user-agnostic estimates. The colored lines indicate the estimated CRA for the different algorithms.