Supplementary Material

Measuring Legislators' Ideological Position in Large Chambers using Pairwise-Comparisons

Part

Appendix

Table of Contents

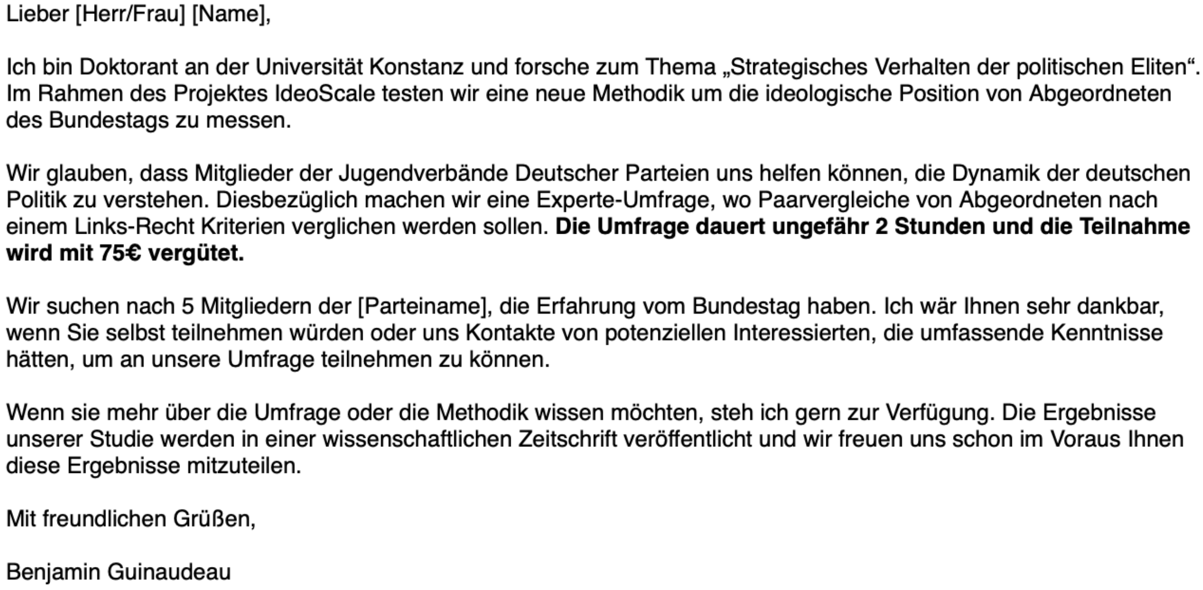
A Recruitment and Participation

The recruitment for this study started by email on April, 29th 2020. It ended on June, 11th 2020. Table 1 lists the organization we contacted. For each organization, we identified and emailed all members of the national committee (or the highest committee managing the organization). Recruitment was based on a "first come, first served" principle, with a limit of 5 members per party. Each participant who completed the 500 comparisons was rewarded with €75. Except for the Junge Alternative (AFD), we managed to have at least 4 participants completing the survey for each party. The German template of recruitment email we used is presented in Figure 1.

Table 1: Contacted organizations and their responses

Youth Party	Party	Contacted	Accepted	Completed
Solid	Die Linke	10	6	5
Grüne Jugend	Bündnis 90/ Die Grünen	10	6	4
Jusos	SPD	11	6	5
Julis	FDP	11	6	4
Junge Union	CDU/CSU	14	6	5
Junge Alternative	AfD	13	2	1

Figure 1: Recruitment Email



Lieber [Herr/Frau] [Name],

Ich bin Doktorant an der Universität Konstanz und forsche zum Thema „Strategisches Verhalten der politischen Eliten“. Im Rahmen des Projektes IdeoScale testen wir eine neue Methodik um die ideologische Position von Abgeordneten des Bundestags zu messen.

Wir glauben, dass Mitglieder der Jugendverbände Deutscher Parteien uns helfen können, die Dynamik der deutschen Politik zu verstehen. Diesbezüglich machen wir eine Experte-Umfrage, wo Paarvergleiche von Abgeordneten nach einem Links-Recht Kriterien verglichen werden sollen. **Die Umfrage dauert ungefähr 2 Stunden und die Teilnahme wird mit 75€ vergütet.**

Wir suchen nach 5 Mitgliedern der [Parteiname], die Erfahrung vom Bundestag haben. Ich wär Ihnen sehr dankbar, wenn Sie selbst teilnehmen würden oder uns Kontakte von potenziellen Interessierten, die umfassende Kenntnisse hätten, um an unsere Umfrage teilnehmen zu können.

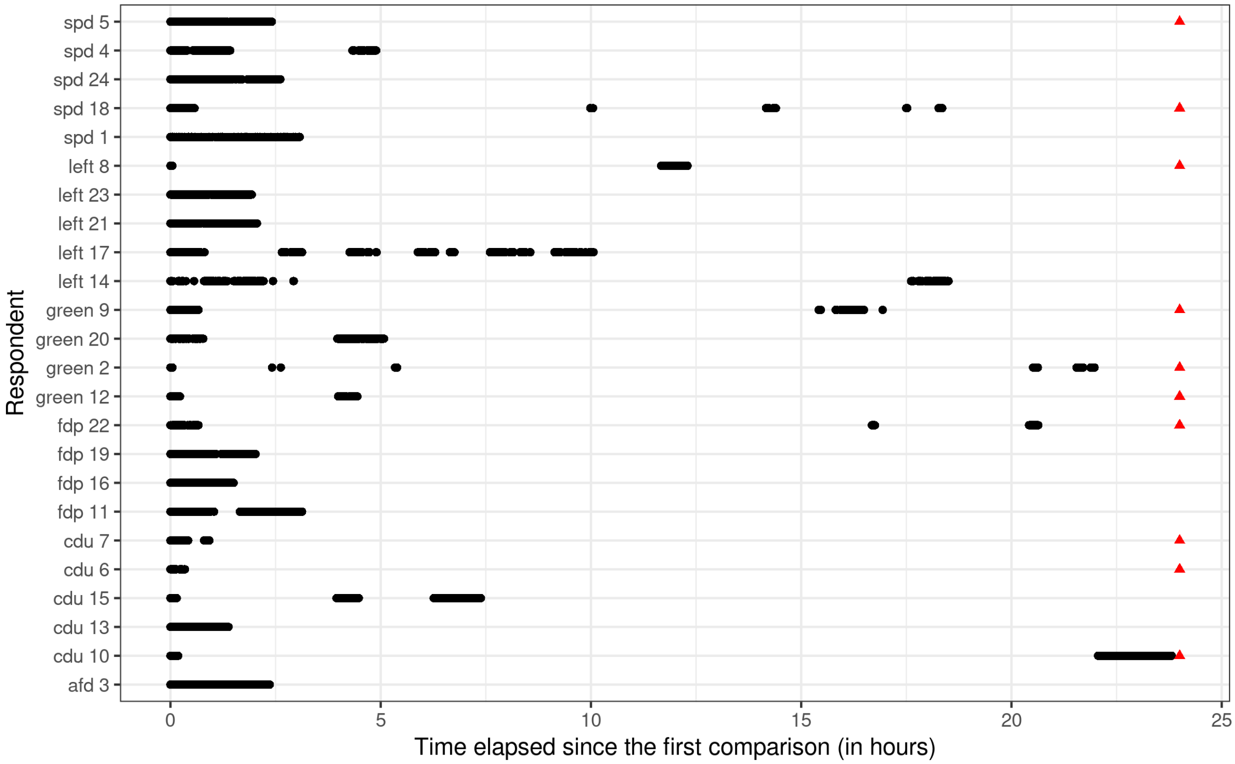
Wenn sie mehr über die Umfrage oder die Methodik wissen möchten, steh ich gern zur Verfügung. Die Ergebnisse unserer Studie werden in einer wissenschaftlichen Zeitschrift veröffentlicht und wir freuen uns schon im Voraus Ihnen diese Ergebnisse mitzuteilen.

Mit freundlichen Grüßen,

Benjamin Guinaudeau

Each participant rated an average of 477 pairs, comparing on average 281 unique MPs. Respondents had the opportunity to declare an MP as unknown (and were encouraged to do so). On average, participants declared 53 MPs as unknown. The median time required to complete a comparison was 13 seconds. Completing the survey required around 2 hours and 33 minutes. Because participants did not necessarily complete the survey in one shot - but also took breaks - it is difficult to assert precisely how long each participant needed to complete the survey. Fig 2 presents an overview of the comparison pace for each respondent.

Figure 2: Comparison over time



Note: Each dot represents one comparison rated by one of the 24 coders (y-axis). The x-axis represents for each coder the time elapsed since the first comparison. Many respondents completed the whole survey in less than 3 hours. Some coders took a break and completed the survey in several waves. Red triangles indicate that the respondents completed comparisons more than 24 hours after their initial comparison.

B Sampling algorithm

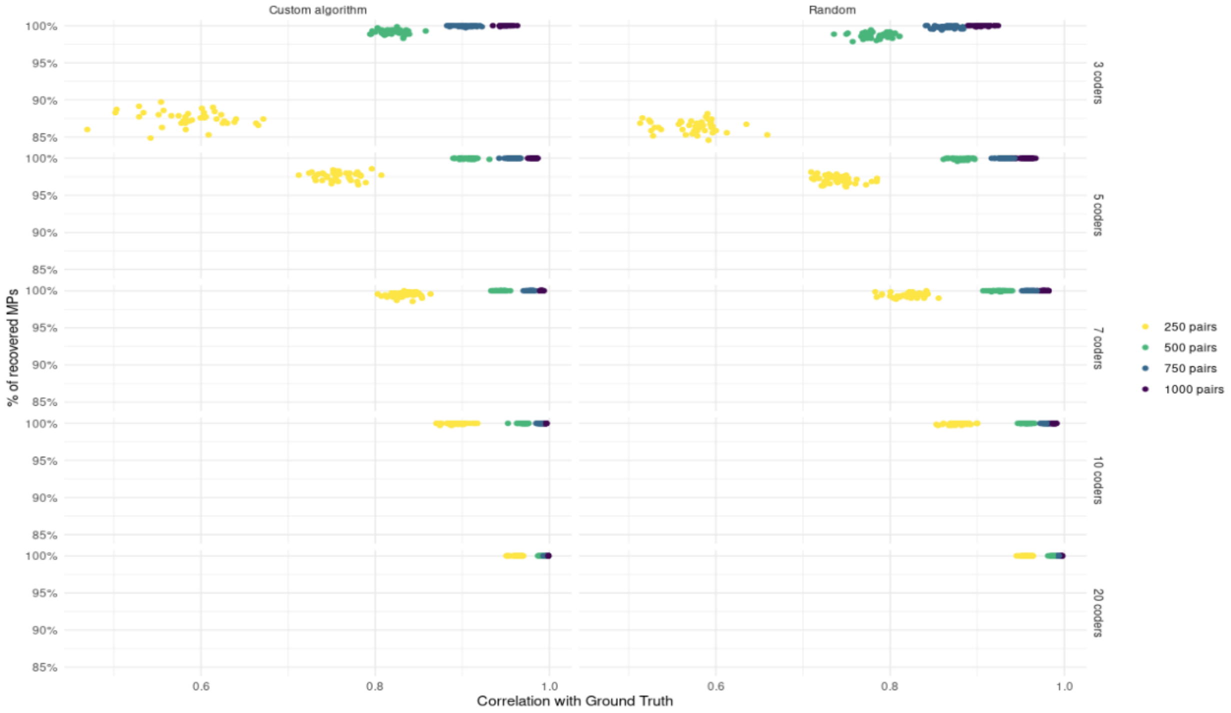
To estimate the impact of the sampling algorithm on the quality of the estimates, we run simulations and compare our custom algorithm against random sampling. As a reminder, the custom algorithm assumes transitivity across one coder's comparisons and identifies informative pairs. In doing so, it avoids comparing far-right MPs with far-left MPs, maximizing the amount of information extracted from a fixed number of comparisons.

The simulation protocol follows three steps. First, we simulate the ideological distribution of 700 MPs by drawing 700 values from a uniform distribution. This distribution determines the ground truth, i.e. the ideological position of the 700 simulated MPs. Second, we simulate 2000 surveys involving a varying number of respondents (3, 5, 7, 10, and 20) and a varying number of pairs for each respondent (250, 500, 750, and 1000). Most importantly, half of these simulated surveys rely on the custom algorithm to draw pairs, while the other half randomly explores the comparison space. Third, we fit one Bradley-Terry model for each simulated survey and evaluate the results against the ground-truth values.

Each time a new pair is drawn, there is a trade-off between exploration – i.e. comparing MPs that have not been yet compared – and exploitation – i.e. comparing MPs that were already compared to increase the precision of their estimates. We hence evaluate the simulated results with two different metrics capturing these two dimensions. For the exploration, we compute the proportion of MPs that were recovered - i.e. which were at least compared once - in the survey. Because the custom algorithm focuses on informative pairs, we expect it to explore faster especially when the number of comparisons is low relative to the number of MPs. For the exploitation, we compute the correlation between the estimates and the ground truth (Pearson's correlation).

The raw results are presented in Fig 3. Each point represents a simulated survey, for which we computed the proportion of recovered units and Pearson's correlation coefficient between the estimates and the ground truth. Generally, we can observe that increasing the number of coders (vertical faceting) or the number of comparisons (colors) generally improves the performance of the survey on both the exploration and exploitation dimensions.

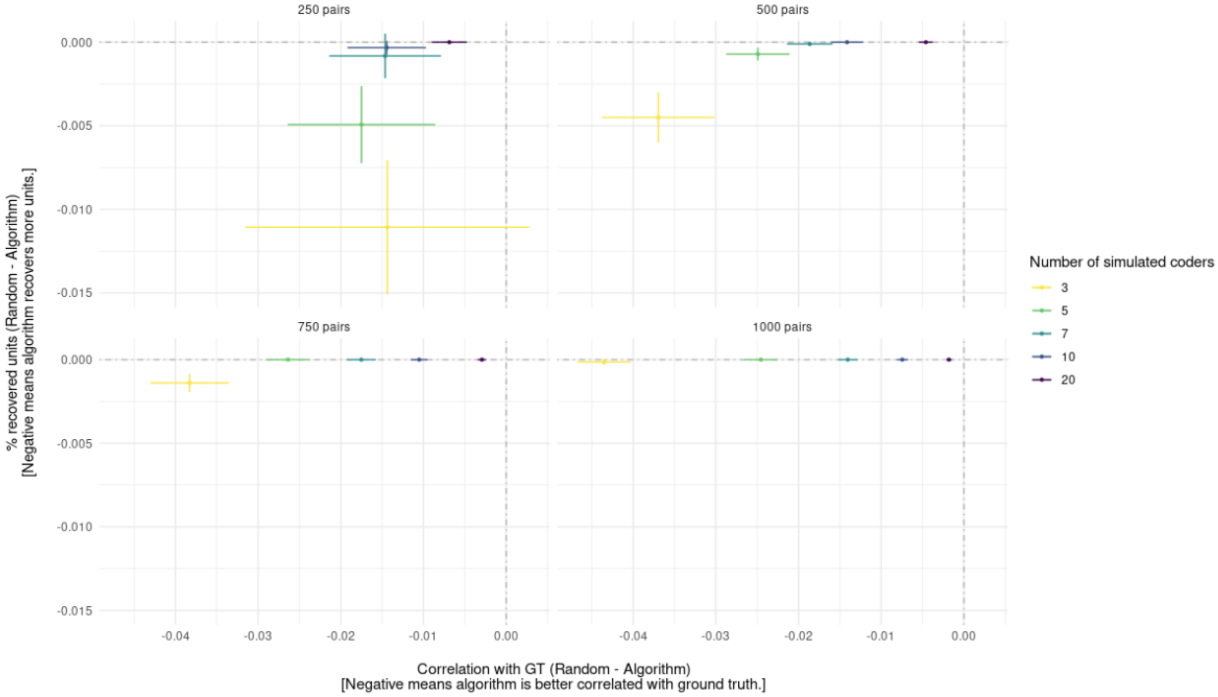
Figure 3: Raw results of the simulations



Note: This figure presents the results of 2000 simulated surveys under varying conditions. Y-axis represents the proportion of MPs for which a score was obtained. Small samples are not able to obtain enough data on all MPs. X-Axis represents the correlation between the obtained estimates and the ground truth.

To ease the comparison of the two approaches, we compute, for each combination of parameters - number of coders, number of pairs, and sampling algorithm -, the average performance and eventually compare the performance of the custom algorithm against the performance under random exploration. The results are presented in Fig 4. Each dot represents the average differences in performance on both dimensions for a given combination of parameters (corresponding to the facets and colors). The horizontal bars represent the 95% confidence interval for the average differences of Pearson's correlation, while the vertical bars represent the confidence interval for the differences in proportions of recovered MPs.

Figure 4: Performance of the custom algorithm against the random exploration



Note: This figure shows the differentiated performance of our custom sampling algorithm when compared to random sampling. We compare the two sampling algorithms across a range of varying survey conditions. Vertical and horizontal bars represent the 95% confidence interval for the difference in performance on the two considered metrics (exploration and exploitation). These computations rely on approximately 2000 surveys with 40 observations for each combination of parameters.

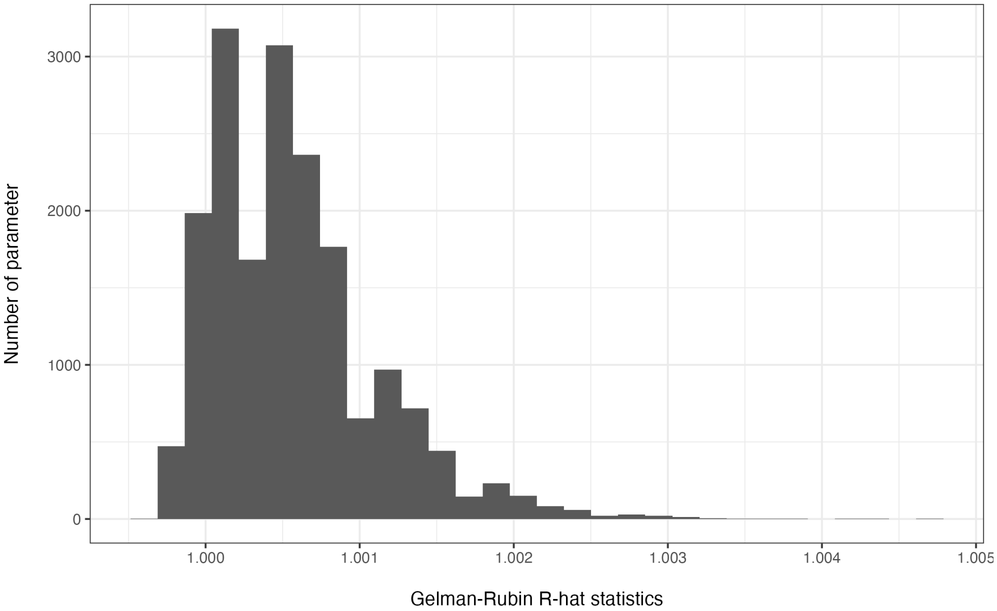
On the exploration dimension, we observe that the performance of the custom algorithm is at least as good as the random exploration of the comparison space. The custom algorithm can recover on average approximately 1% more observations for the smallest simulated survey (250 comparisons with 3 respondents). Overall, the two approaches perform similarly on the exploration dimension and produce most of the time an estimate for all MPs.

On the exploitation dimension, the simulated surveys relying on the custom algorithm are systematically and significantly more precise than the random exploration. The difference between the two approaches attenuates as the sample size grows. However, even for the largest simulated survey (20 coders with 1000 comparisons each 20000 comparisons), the custom algorithm produces estimates whose correlation with the ground truth is on average 0.01 higher than with random sampling.

C Convergence

To evaluate the convergence of the model, we relied on various diagnostics. Our model features hundreds of parameters, so that qualitative methods to assert convergences, such as trace plots, are not practicable to display here. Still, we manually evaluated the trace plots for dozens of parameters without detecting any issues. In addition, we computed Gelman-Rubin R-hat statistics for each parameter. In most cases, they were close to one (see distribution presented in Figure 5) suggesting convergence.

Figure 5: Gelman-Rubin R-hat



D Alternative modelling strategies

The main estimates presented in this paper rely on a Davidson model and feature "similar positions" as an option for a rating. Because "similar positions" are informative, they should be used for estimation. To assess the robustness of our estimates, we benchmark the main estimates with estimates from a plain Bradley-Terry model. This model excludes "similar positions". After excluding "similar positions", we obtained 9522 ratings and used these to estimate a Bradley-Terry model. Figure 6 shows the individual estimates for each MP according to the Davidson model (x-axis) and the Bradley-Terry model (y-axis). The very high correlation (r = .98) suggests that including the "similar positions" does not bias the estimates. However, the credible intervals from the Davidson models are smaller: Figure 7 compares the width of the individual credible intervals and shows that the Bradley-Terry model systematically produced larger estimates than the Davidson model. This benchmark suggests the benefit of a Davidson model and adding a "similar positions" rating, which produces similar but more precise estimates than a Bradley-Terry model.

Figure 6: Comparison between estimates from a Davidson and a Bradley-Terry model.

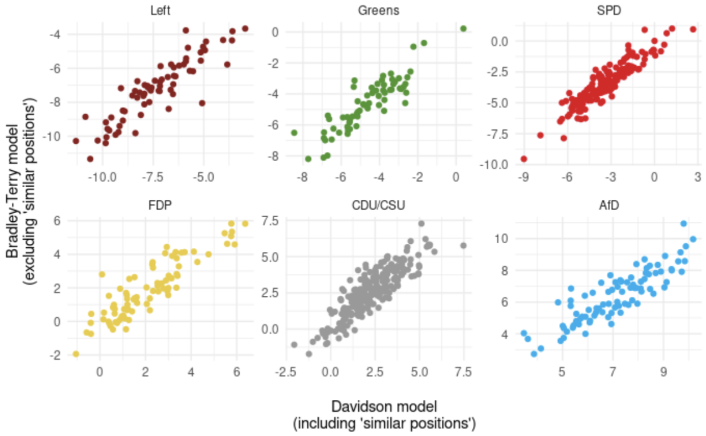
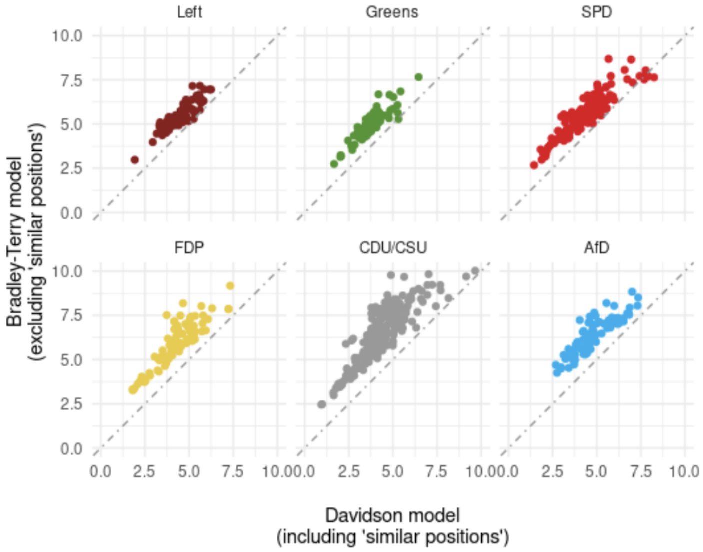


Figure 7: Width of credible intervals across models.



Moreover, to evaluate the influence of the decision to estimate the model in a Bayesian framework, we estimate a frequentist Bradley-Terry model and compare its results with the main estimates, i.e. the Davidson model. This model is fitted using the R package BradleyTerry2 after excluding the rating "similar positions".

Figure 8: Distribution of estimates across Bayesian and Frequentist frameworks

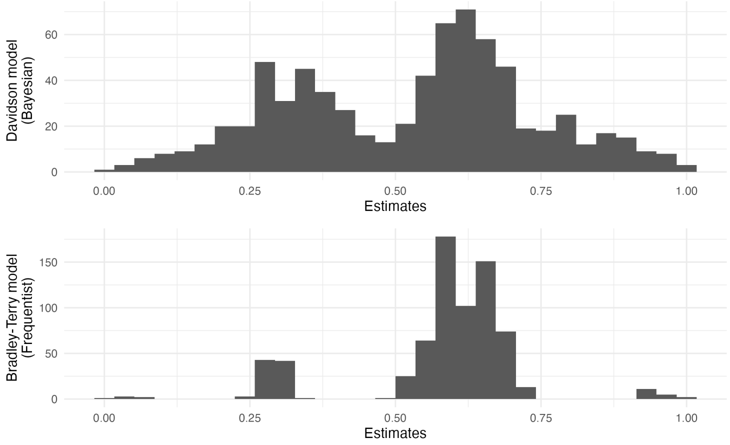


Figure 8 shows the distribution of the ideological scores for both the main model and the frequentist Bradley-Terry. If both distributions are bi-modal, the frequentist Bradley-

Terry produces a disjunct and more concentrated distribution, which suggests it is less able to detect contrast between the MPs' ideological positions.

Figure 9: Comparison between estimates from the Davidson model and the frequentist Bradley-Terry model

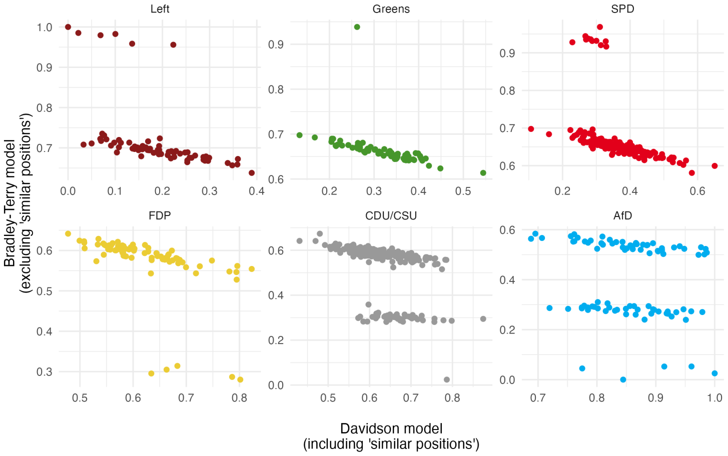
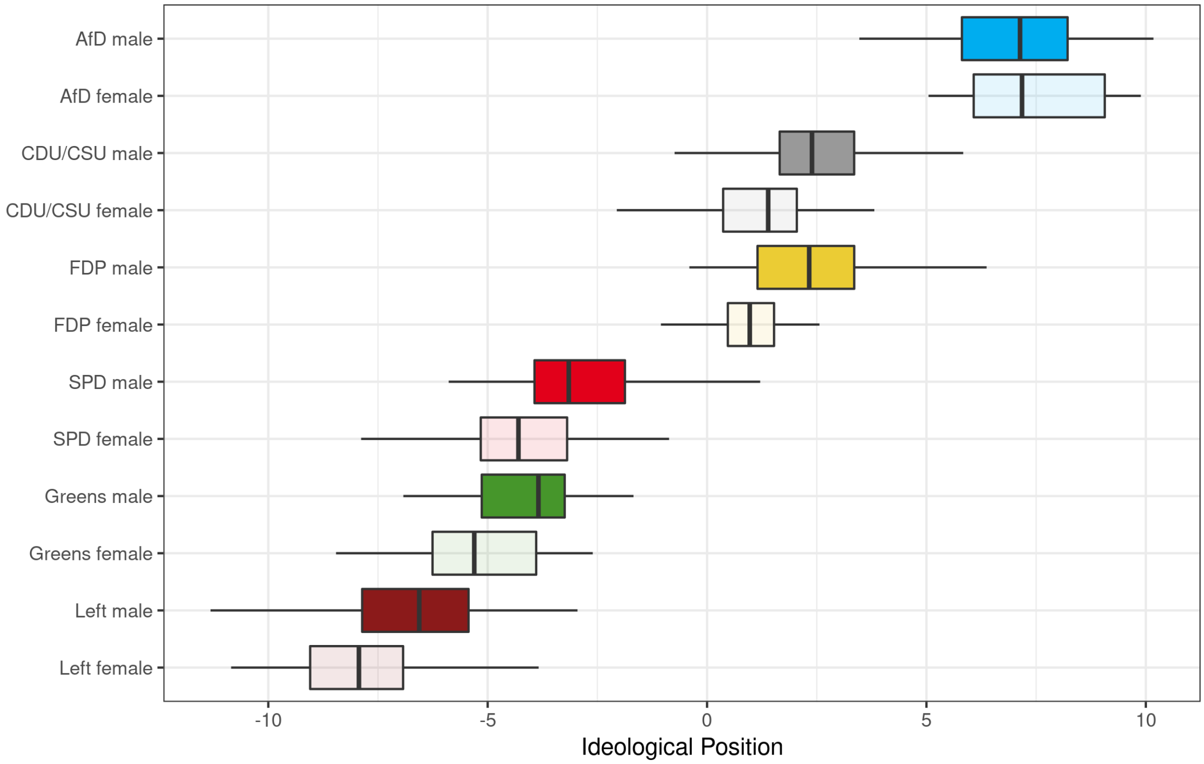


Figure 9 presents the benchmarked estimates for the two models. To project the two distributions on the same scale, we used min-max normalization. Overall, the estimates from the two models vary in the same direction, but again the disjunct distribution from the frequentist models induces unusual clusters in the data. For instance, most of the MPs from the CDU/CSU are estimated at around 0, but a large group of MPs obtains scores of around -20 with almost no MPs rated between -5 and -15.

E Additional survey result

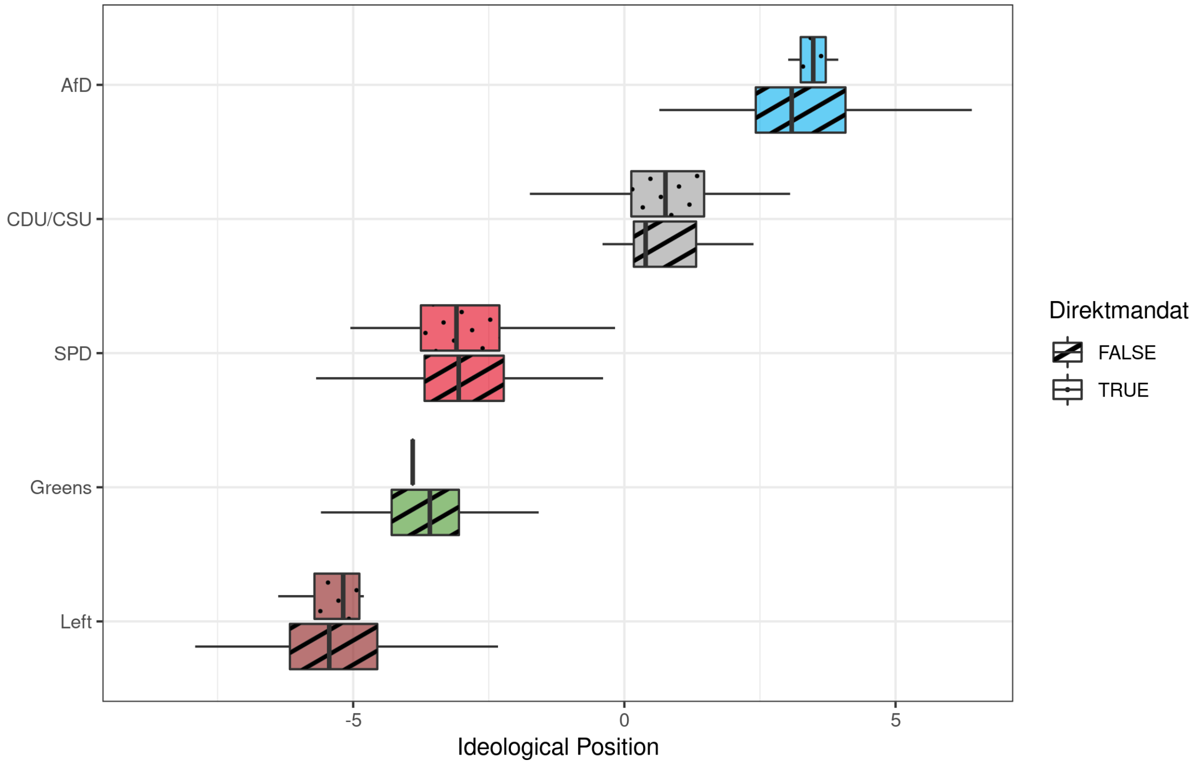
In this section, we present descriptive results of the estimates and how they relate to important individual traits of German MPs. Figure 10 compares within each party the ideological position of women and men. Women hold systematically more leftist positions than men. Figure 11 compares for each party the ideological score of MPs elected on the party list and MPs directly elected by a constituency (*Direktmandat*). The positions of directly elected MPs do not diverge systematically from those coming from lists. Finally, Figure 12 compares MPs from East and West Germany. Eastern MPs hold on average slightly more leftist positions.

Figure 10: Gender



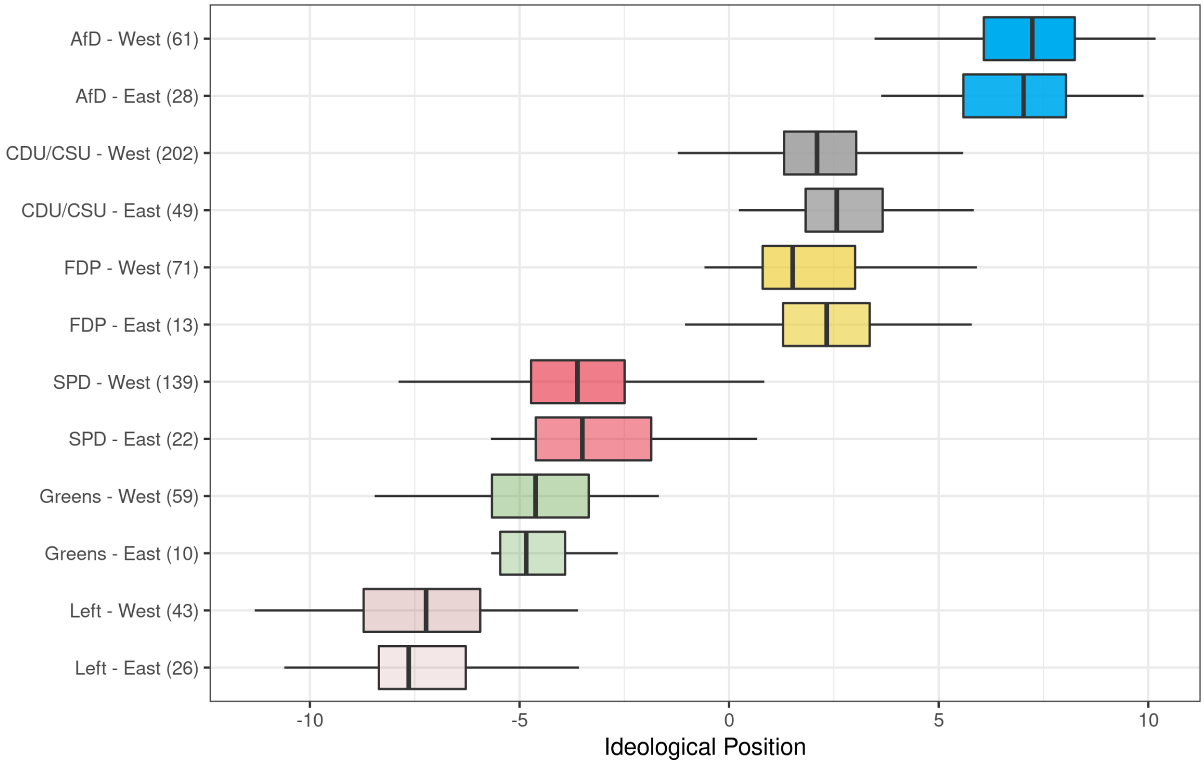
Notes: Each boxplot represent the ideological scores of groups of MPs depending on their party and their gender.

Figure 11: Direct Mandate



Notes: Each boxplot represent the ideological scores of groups of MPs depending on their party and whether they have been directly elected by a constituency (Direktmandat).

Figure 12: East-West comparison



Notes: Each boxplot represent the ideological scores of groups of MPs depending on their party and whether they are from East-Germany.