Better Measures and Fewer Questions? Improving Measures of Political Knowledge.

Maxime Walder [*], Nathalie Giger [†] and Denise Traber [‡]

October 31, 2024

early draft, comments welcome; please do not cite or circulate

Abstract

Political knowledge is a concept frequently used in political science research. While numerous studies include an indicator of political knowledge as explanatory or moderating variables, measures of the concepts are often very limited. One of the key issues is that knowledge questions need to be updated over time, and that we do not have consistent ways to compare knowledge based on a set of different items. In this paper, we first propose a way to develop surveys to have a comparable measure of political knowledge over-time using Bayesian Item-Response-Theory models. In addition, we also use posterior estimates to select the best items in future surveys. In doing so, we contribute to the literature on the measurement of political knowledge and survey research. Our results show that using the proposed framework can increase the measurement validity and enable us to position respondents of different surveys in a similar latent space.

[*]University of Geneva, maxime.walder@unige.ch
[†]University of Geneva, nathalie.giger@unige.ch
[‡]University of Basel, denise.traber@unibas.ch

Introduction

The literature in political science has long recognized the immense relevance of well-informed, knowledgeable citizens for democracy: "An informed citizenry, able to appreciate its interests and make intelligent judgments, is essential to democracy" (Bennett 1989: 422). Political knowledge is indispensable for the role that democracy gives to its citizens: to identify their interests and to act upon them when choosing political personnel; to evaluate the governments' and parties' actions, and to sanction them if necessary.

While the general concept of political knowledge is uncontested, the exact measurement remains a challenge. The most common approach dates back to the seminal work of Carpini and Keeter (1996) who define political knowledge as "the range of factual information about politics stored in long-term memory." Their five-item index, which asks respondents about electoral and institutional processes, has become the standard in the literature whenever scholars attempt to measure political knowledge. In practice, this means correctly identifying the Chancellor in Germany or knowing that Switzerland has two chambers of parliament. As recognized by Kraft and Dolan (2023), despite the large variety of political science work that tackles the concept of political knowledge, there exists a remarkable convergence when it comes to measurement: most often scholars apply measures that focus on the rules of government and institutions and the identification of political leaders. As such, being knowledgeable about politics can be defined as answering correctly a couple of questions about politics.

This is not to say that no criticism exists on how to measure political knowledge. Scholars raised general concerns about the practice to base broad generalizations of citizens' competence on a relatively small set of idiosyncratic, fact-based questions (e.g. Lupia, 2016). Others criticized the focus on institutional facts (Barabas, Jerit, Pollock, & Rainey, 2014; Rapeli, 2022). Further, a methodological debate has emerged around how the (online) survey environment impacts the provision of correct answers (Boudreau & Lupia, 2011; Kleinberg, 2022; Munzert & Selb, 2017) and whether closed-ended questions should be replaced by open-ended questions as a more valid alternative (Kraft & Dolan, 2023). What is missing in our view, though, is a debate on how to select the most valid questions to measure political knowledge, i.e. those who are best

able the capture the underlying concept. Also, the question how to aggregate the single responses to one single measure of the underlying concept – given the difference in difficulty between questions (see also Bullock & Rader, 2022) – has not been discussed prominently, and simple summary measures (e.g. the mean) prevail.

In this paper, we propose an innovative way to measure the concept of political knowledge in a cost-efficient and robust way with few but well-defined questions. We provide insights on how many and which questions are appropriate to evaluate the political knowledge of citizens on a latent scale. We use Monte Carlo simulations and Bayesian Item Response Theory (IRT) models to show how questions with different parameter values impact the quality of posterior estimates, i.e our measure of knowledge. We then use survey data to validate and test the simulations, and show how our technique can be used to assess political knowledge over time in a flexible way. In the end, we give recommendations on how to measure political knowledge in surveys, as well as ways to improve the measure of the concept over time using item randomization.

We offer two major innovations: First, by randomizing questions across respondents and applying a Bayesian IRT model, we get an estimate of the latent concept of the underlying scale of political knowledge with only a handful of questions asked to each respondent in the survey. Second, our approach allows to assess the validity and quality of the individual questions asked, i.e. the ability to measure the underlying concept. This in turn allows us to improve the questions asked by selecting those which are better suited only.

By this approach we circumvent another measurement problem, the aggregation of the single items to a scale that is reliable and ideally also comparable across time and contexts (see e.g. Elff, 2009; van Heerde-Hudson, 2020). So far, only few attempts exist to study political knowledge over time (Bathelt, Jedinger, & Maier, 2016; Lizotte & Sidman, 2009) as the availability of constant items over time is very limited. Our approach is more flexible and thus renders such an endeavour also possible with a small sample of items.

Measuring political knowledge

In their seminal work, Carpini and Keeter (1996) define political knowledge as factual information, i.e. knowledge about political facts with questions that have one objectively right, and (potentially) several wrong answers. The authors group knowledge about politics into three basic categories: the *rules of the game*, *people and parties* and the *substance of politics*, i.e. knowledge of specific policy areas. In general, there is an astonishing similarity in the measurement approaches in the literature. In fact, existing research measures the concept often only with a handful of questions, mostly focusing on institutional facts (Kraft & Dolan, 2023; Lupia, 2016).

What these questions should be about exactly and whether to asked them in closed-ended or open question formats has been subject to discussions, however. While the later debate has not been resolved (see e.g. Bullock & Rader, 2022; Kleinberg, 2022; Pietryka & MacIntosh, 2022), the consensus on the former is that political knowledge should be assessed with policy-specific and general questions (Barabas et al., 2014), and questions of varying difficulty (e.g. Bullock & Rader, 2022; Kleinberg, 2022). Asking who is the president of the USA, for example, can be considered an easier question than asking about the minority party leader in Germany. In fact, if most people know the correct answer, the question is not helpful to differentiate levels of political knowledge among respondents. An addition, given the online environment of most surveys nowadays, the questions should not be easily searchable online to avoid cheating (e.g. Kleinberg, 2022; Munzert, Ramirez-Ruiz, Barberá, Guess, & Yang, 2024; Munzert & Selb, 2017). It has been shown for example that the visual display of politicians helps circumvent cheating.

What is rather absent from this literature is a discussion of how valid these questions are, i.e how well they are able to capture the underlying concept of political knowledge. For instance, if a question relates to a policy that recently heavily impacted specific types of farming, chances are that knowing about this policy is more related to another concept – such as the closeness to the agricultural sectors – more than it is to the concept of general political knowledge. So, we want questions that directly tap into somebody's political knowledge, i.e. where the ability to give a correct answer is a valid demonstration of the level of political sophistication of the respondent.

The selection of knowledge questions thus determines the measurement quality of

the overall concept of political knowledge. While there have been several recommendations on how to develop survey questions for the measurement of political knowledge (e.g. Kraft & Dolan, 2023), a discussion on how to select from a pool of questions to improve measurement validity has been largely absent so far

Measurement approach

In this paper we use a Bayesian Item-Response theory (IRT) to estimate the latent concept of political knowledge. While such an approach is not unknown to this literature (see e.g. Carpini & Keeter, 1993; Pietryka & MacIntosh, 2013; Tsai & Lin, 2017) and in fact, already Carpini and Keeter (1993) use it for model checks, the dominant approach in this literature is still using a naive summing up of indicators. In our view, the IRT model is closer to the underlying measurement model of political knowledge which has been described as an effect model (Pietryka & MacIntosh, 2013). Most importantly though, an IRT approach allows to test the quality of the single questions of political knowledge and provides approved procedures evaluate not only the difficulty but also the degree to which these questions tap into the ability of respondents to answer political knowledge questions (the discrimination parameter).

We use the Bayesian variant of an IRT model because its flexible form has several advantages in our view. First, it allows to sample questions from a wider universe of political knowledge items. In doing so we can keep the number of items used to measure the latent scale low, while still being able to test the quality of a larger number of items. Second, the flexible form allows us to estimate political knowledge scores over time, even if not all questions are asked in all surveys.

Bayesian IRT models have long been used in educational research (König & van de Schoot, 2018). In political science, these models have largely been used to measure the ideal position of actors in latent space (see for instance Caughey, O'Grady, & Warshaw, 2019; Caughey & Warshaw, 2015; O'Grady, 2019; O'Grady & Abou-Chadi, n.d.). However, in our view, when measuring political knowledge, the Bayesian IRT model is closer to the original concept in educational research than recent developments in political science.

Bayesian IRT models define a latent space where each respondent has an ideal

position. In our example, the latent space represents the space of political knowledge. Each respondent j has a position on the latent space summarized by the parameter θ, – the ability parameter. Additionally, each question i has two parameters: the difficulty a of the question and its discrimination b. In sum, two parameters are evaluated for each question: how difficult it is and how indicative of the position of actors in the latent space it is. Formally, the model can be written as:

$$Y_{ij} \sim Bernoulli(logit^{-1}(a_i + b_i\theta_j)) \tag{1}$$

Where Y_{ij} is the outcome variable and takes the value 1 if respondent j gave the correct answer to question i and 0 otherwise. a is the difficulty parameter of question i. Indeed, if all respondents answer the question correctly, then the a parameter will have a high value and the correct response is not a function of the knowledge of the respondent but of the question's difficulty. Finally, the parameter b_i is the discrimination of the question and θ_j is the position of the actor j in the latent space – in our case the level of political knowledge.

There is a key difference in the Bayesian IRT models for knowledge and other social constructs, such as ideology. Given that our outcome variable Y_{ij} indicates whether the respondents gave the correct answer or not, it can only impact the level of political knowledge in one direction: the correct answer can only increase the level of political knowledge or at least keep it constant.[1]

To take this into account, we need to adapt the priors of the parameters in the model, and more specifically, the discrimination parameter b_i with a half-normal distribution enabling only positive values for this parameter.[2]

[1]This is not the case for ideology. One conventional approach to place actors in the latent space is to consider how they voted in parliament for instance. However, depending on the content of the policy, voting "yes" may indicate a more conservative or a more progressive position. Thus, voting on ballot proposals may impact the ideal position of actors in the latent space in two directions.

[2]Let's consider we aim to measure the ideal position of actors in a latent space where positive values represent the progressive position and negative values represent the conservative position. In this case, when the policy at stake aims for conservative changes, the discrimination parameter b should take negative values to account for the fact that more conservative positions are more likely to vote yes and more progressive actors are less likely. Multiplying the positive parameter of progressive actors with a negative discrimination parameter gives a negative product which relates to the lower probability these actors have to support the policy. For conservative actors, it would mean multiplying

From the formalization of the IRT model it becomes clear that two parameters are important here: the discrimination parameter b and the difficulty of the question (a). We now discuss in turn which characteristics of these parameters are optimal for a reliable and cost-efficient way to measure political knowledge.

While questions with low discrimination do not add much information to the ability or ideal position of respondents, questions with higher discrimination parameters are a substantial benefit for the model's performance. This parameter is equivalent to the learning curve as described by Delli Carpini and Keeter (1996). In short, the learning curve is steeper with higher discrimination. This means that differences in the outcome between low and high ability are more probable with higher discrimination parameters. Figure 1 illustrates how the discrimination parameters affect the probability of a correct answer.

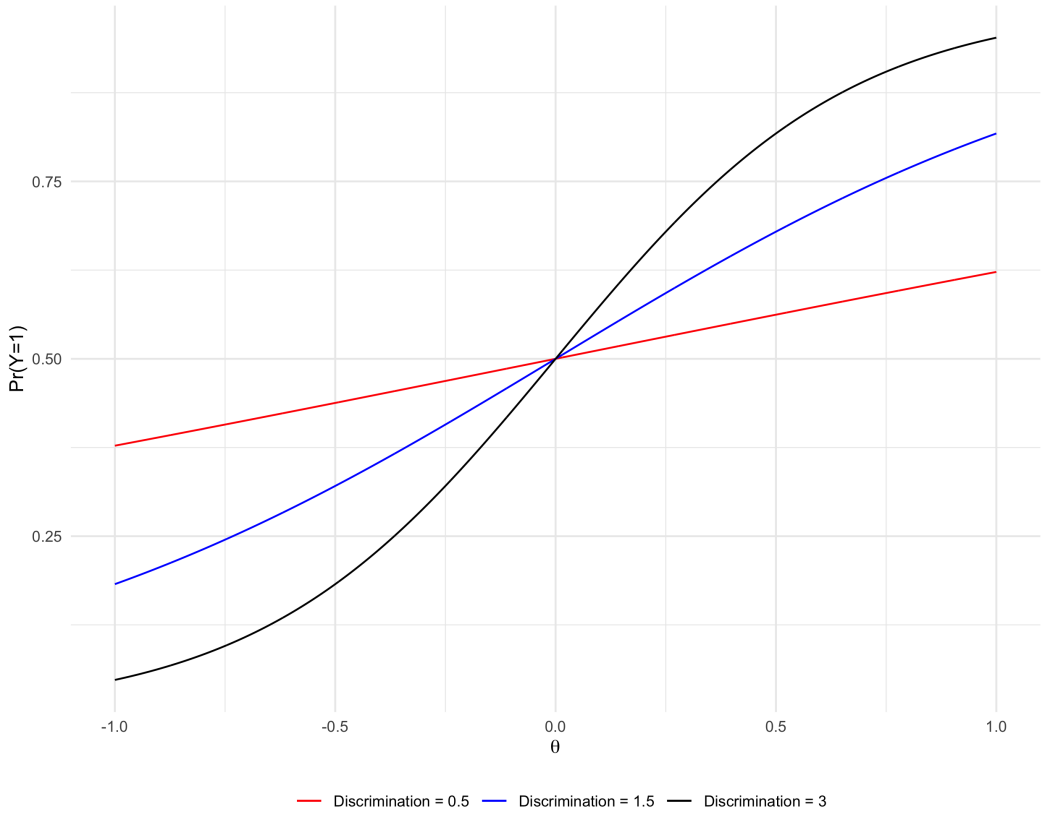

Figure 1: Effect of discrimination magnitude on the probability of correct answer.

the negative position by negative discrimination which indicates a higher probability to support the proposal. This logic is the exact inverse if we consider votes on policies that aim at progressive changes. However, the model works differently if we consider a space where negative values indicate less political knowledge and positive values indicate higher levels of political knowledge.

Figure 1 shows the fundamental importance of the discrimination parameter in IRT models. The probability of answering the question correctly only slightly changes if we consider questions with low discrimination parameters. Between low ability (θ = -1) and high ability (θ = 1) the probability of answering the question correctly changes from 37% to 63%. Although this already differentiates between different levels of knowledge, there are still 4 out of 10 chances that respondents with low ability answer correctly and that respondents with high ability give the wrong answer. However, we see that the difference in the probability of correct answers between respondents with high and low ability increases with the magnitude of the discrimination parameter. Indeed, for the question with a discrimination of 1.5, the probability grows from 18% to 82%, and for the questions with a discrimination parameter of 3, it grows from 5% to 95% of probability.

Thus, the discrimination parameter has a large impact on the learning curve. Higher discrimination lowers the probability of correct responses from respondents with low ability, and increases the probability of correct responses from respondents with high ability. At the extreme, a discrimination parameter of 0 would not change the probability of a correct answer with an increased ability. Thus, we know that higher discrimination in questions is better to evaluate the political knowledge of respondents.

In practice, imagine a project that drafts m questions on political knowledge. However, the resources only enable the researcher to ask n questions to each respondent, and $n < m$. So, the number of questions each respondent received is lower than the total number of questions drafted. How should researchers select n questions to build the best measure of political knowledge? One crucial aspect of our methodology is the randomization of questions shown to the respondents (subset n). This approach allows us to utilize all m questions in the survey, enhancing the precision of our political knowledge measures over time. By employing this randomization, we can create more accurate and detailed profiles of citizens' political knowledge.

In the IRT model, there is a second parameter of interest: the difficulty parameter. In line with Bullock and Rader (2022); Kleinberg (2022) we define difficult questions as questions that are less known. The observable implication is that difficult questions have fewer respondents who answer them correctly.

Figure 1 shows the effect of the discrimination parameter on the probability that respondents with different abilities answer the question correctly. In Figure 1, the difficulty parameter is constant and set to 0. However, this parameter can also vary and in interaction with the item's discrimination parameter affect the probability to answer a question correctly. Indeed, questions that are too easy or too hard – where all or none of the respondents have the correct answer – do not help estimating respondents' ability parameter (the position on the knowledge scale). As soon as respondents' answers vary, there may be some variance in the discrimination parameter. Imagine a hypothetical question where 50% of respondents have a correct answer. This question could have high discrimination because the respondents' ability is what differentiates the correct from the wrong answer. However, if the question is very difficult, in the extreme the 50% correct answers are due to pure luck and would not be caused by the ability of respondents to answer the question correctly. Thus, the most interesting questions to select for measuring knowledge are those with high discrimination and varying difficulty.

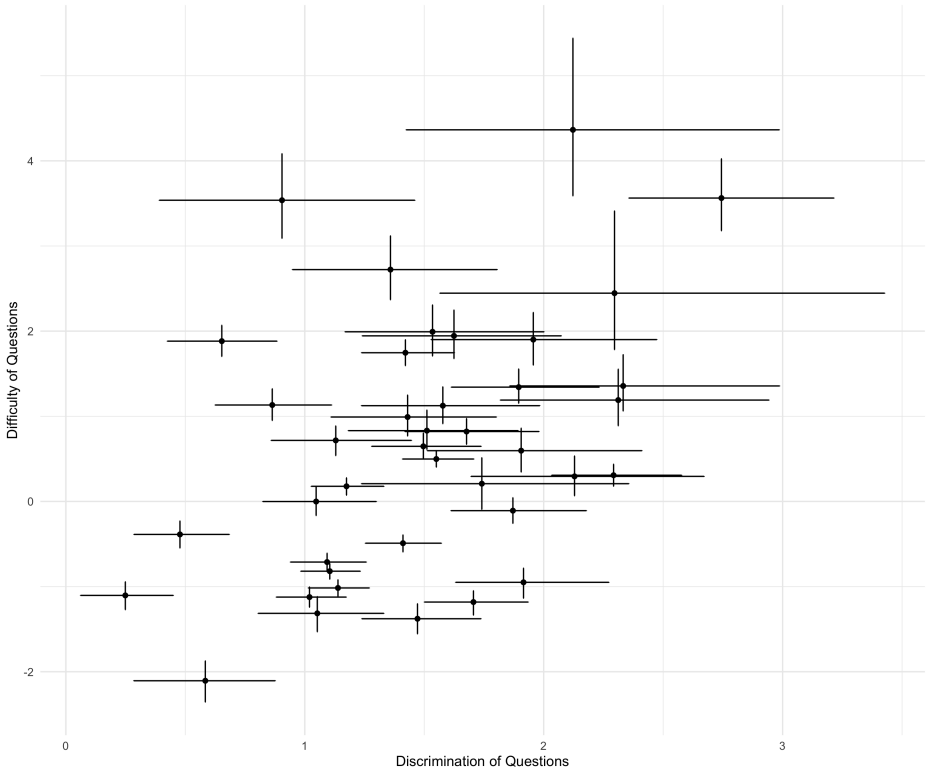

Figure 2: Posterior estimates (median and 95% credible intervals) for difficulty and discrimination parameters (40 questions in three different survey waves.)

To exemplify this variation, Figure 2 presents the distribution of the discrimination and difficulty parameters from 40 political knowledge questions we asked respondents in three survey waves.[3] There is a clear pattern visible in Figure 2 between the difficulty parameters (y-axis) and discrimination parameters (x-axis) : easy questions (negative difficulty parameters) also tend to have relatively low discrimination. Despite this correlation (r=.427), however, it is important to note that there is important variation in difficulty for the same discrimination estimates, and variation in discrimination for the same difficulty estimates, vice-versa. Indeed, the estimated difficulty parameters for items with discrimination above 2 range from 0 to more than 4. Second, we see that the discrimination parameter can be below 1 or above 2 for questions with high difficulty.

In sum, we argue that (1) choosing a subset of questions with the highest discrimination parameters and varying difficulty increases the validity and performance of posterior estimates, and that (2) question randomization enables the testing of more questions, which increases the pool of questions researchers can choose from without decreasing the quality of posterior predictions. To estimate how well the postulated argument is empirically valid, we present the results of different simulated models. First, we show how question randomization affects model performance. Second, we show how increasing the discrimination parameter enhances the posterior prediction.

Simulations

Monte Carlo simulations have generally been used to assess model performance. The main advantage of simulations is that we can compare the "true" value of our parameters of interest (in this case θ), and its distribution. The true value comes from a synthetic distribution that we generate first. The simulation allows us to compare the simulated parameter $\hat{\theta}$) with the true parameter θ. In doing so, it is possible to assess which models perform the best for the precise purpose of the study. Hopkins et al (2024) suggest several indicators to evaluate the performance of models based on Monte Carlo simulation. Among others, they suggest computing the *bias*, *standard deviation*, and *root mean squared error (RMSE)* of estimates.

[3]Table 4 presents the data in detail.

Based on the explanation above, we need two types of simulations: First, we claim that item randomization should allow us to increase the pool of knowledge questions even though space is constrained in the survey. Thus, the first simulations compare the performance of models with and without randomization, keeping discrimination and difficulty constant. Specifically, we compare simulations where $\hat{\theta}$ is estimated with 1 out 1, 1 out 3, or 1 out of 5 questions.

Second, we postulate that questions with higher discrimination improve the performance of the model. For this second simulation, we compare models with similar item difficulty but with discrimination parameters following a half-normal distribution centered on 1, 2, or 3, with a standard deviation of 0.1. In this step, we expect better performance of models with higher discrimination parameters.

Comparing models with and without item randomization

We create the simulated data as follows:

1. The respondents' true ability θ follows a standard normal distribution.

2. The question difficulty a is distributed based on a normal distribution centered on 0 with a standard deviation of 2.[4]

3. The question discrimination parameters b are defined as a half-normal distribution centered on 1.5 with a standard deviation of one.[5]

Following the recommendation of Hopkins et al. (2024), we compare simulations based on 10 responses to questions in three different settings: No question randomization, which means each simulated actor replies to the 10 same questions (1); a setting where respondents see 10 out of 30 questions (2); and a setting where respondents see 10 out of 50 questions (3). Comparing these models informs us about the drawbacks of question randomization for the model performance. Table 1 presents the different values of the indicator of model performance for the three models.

[4]These are realistic values based on our observational data.

[5]These are realistic values based on our observational data.

	No randomization	One out of three	One out of five
Bias	-0.030910799	0.003683538	0.032074252
Standard deviation	0.7435248	0.8076239	0.8433687
Root Mean Squared Error	28.90007	29.55817	29.99976

Table 1: Bias, Standard deviation, and RMSE for models with and without item randomization.

Table 1 shows that the performance indicators are essentially similar between the models. Indeed, we see that the RMSE is very close for the models with and without item randomization. Although the model with one out of five questions shown to respondents has the highest value, the difference compared to the model with no randomization is negligible. Further, item randomization does not systematically bias the ideal point estimates more than models with no randomization. Finally, the standard deviation is close to the true standard deviation (equal to 1 by definition) in the model with the largest item randomization. In sum, the RMSE, bias, and standard deviation of ideal point estimates are only minimally affected by item randomization. Thus, the simulation shows that item randomization enables researchers to test as many questions as possible which will increase model performance to estimate ideal positions of actors.

Comparing models with different discrimination parameters and number of questions

The second set of simulations compares the performance of models including items with different distributions of the discrimination parameter. We compare simulations of 1000 respondents on 15 questions where the discrimination is a half-normal distribution centered on 1, 2, or 3 with a standard deviation of 0.1. Table 2 presents the indicators' values for the different models.

Table 2 shows that in all of the metrics, the model including discrimination parameters centered on 3 has a better performance to compute the ideal position of actors than models with lower discrimination scores. Indeed, we see that the bias and RMSE are higher, and that the standard deviation is further away from the true value (1) with lower discrimination values. In sum, the model with the highest discrimination

	Center on one	Centered on two	Centered on three
Bias	-0.013849422	-0.008030910	-0.008882072
Standard deviation	0.7841315	0.8993759	0.9311952
Root Mean Squared Error	19.53118	13.13782	10.06769

Table 2: Bias, Standard deviation, and RMSE for models with different discrimination parameter

value is better suited to catch the true knowledge score in the latent space.

While changes in discrimination parameters affect the model performance to estimate the ideal position of actors, it is important to keep in mind how the number of questions also affects model performance. Ideally, researchers would ask as many questions as possible. In reality researchers are constrained by resources, data, and survey time. Choosing the questions with the highest discrimination parameter increases the models' performance; however, it is not clear what the trade-off between more questions with lower discrimination or fewer questions with higher discrimination is. To investigate this trade-off, we compare models with different discrimination values and different numbers of questions included in the model.

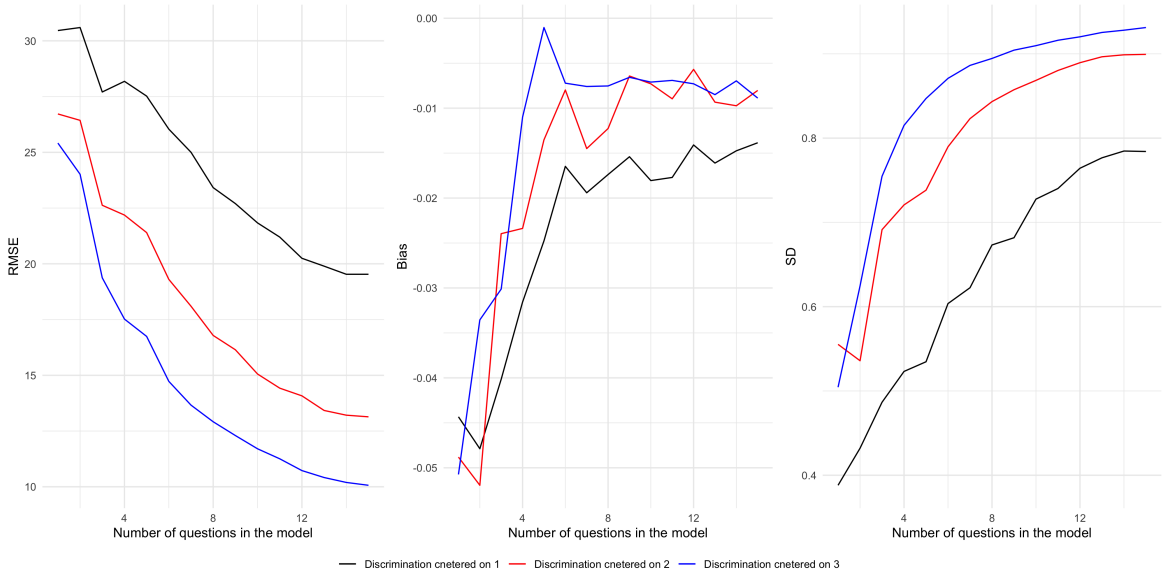

Figure 3: RSME, bias, and SD for models with different discrimination values.

Figure 3 shows that models with the lowest discrimination of questions perform worst in all performance indicators (blue line). However, models that include questions with a higher discrimination perform better, even with a low number of items. For example, in terms of RMSE, the performance of a model that includes five questions with discrimination around 3 is similar to the performance of a model that includes 15 questions with discrimination around 1. Further, when looking at bias and SD, we see that the models with the lowest discrimination score do not reach the same performance as the models with higher discrimination values. In sum, this shows that a careful selection of questions based on preliminary analysis can substantially increase the model performance in estimating ideal positions of actors.

Taking these results into account, the following step is to assess *how many* questions the researcher should ask to provide good estimates of the latent knowledge scale and peoples' abilities. As shown in Figure 3, the performance indicators improve with more questions – no matter the discrimination score. Asking a minimum of 5 or 6 questions significantly increases the performance of the model. However, we also see some stagnation after a certain number of questions and the performance indicators follow an exponential curve.[6] In sum, the number of questions for the best model performance depends on the quality of the questions.

In the next section, we turn to a comparison between a real-world example and the simulations. Afterwards, we list the different recommendations to take into account when measuring political knowledge before presenting concluding remarks.

[6] Additional questions increase the performance indicator at a lower rate with a higher number of questions. For instance, there is great improvement in Bias and SD between the first 6 to 8 questions, but questions 9 to 15 do not have a large impact on the performance. Similarly, we see that the RMSE is decreasing fast with the first 6 to 8 questions, but less so afterwards.

Real-world data 1: Swiss Survey 2022

We first investigate data from a survey among Swiss citizens that was conducted between the 26^{th} of September and the 9^{th} of November 2022. Overall, the survey contains the responses of more than 2000 respondents to 11 questions on political knowledge. These 11 questions are drawn from a sample of 24 political knowledge questions (see Table 4 for more information). The topics of the questions vary strongly. They include institutional facts (e.g. allocation of seats in the lower house) but also feature party competition and knowledge about the positioning of parties on the left-right scale. Furthermore, the introductory text included a request not to use direct google search to improve the measure of the concept of political knowledge and not the efficiency in internet search (Kleinberg, 2022; Munzert et al., 2024) and all questions had 4 answer categories to ensure that guessing is not easy (Bullock & Rader, 2022).

In the following, we compare the model performance of the real-world data with the simulated data based on the same items. First, we compute the observed model. We then take the value of the discrimination and the difficulty parameters from the observed model and generate the outcomes based on these values and the true position of actors in the simulation. In doing so, we simulate data that is as close as possible to the observational data. One challenge is to compute performance indicators for the observed model. Indeed, in this case, we do not have the true value θ. To estimate this value we consider that the model closest to the true value is the one including 11 questions.

Figure 4 shows the evolution of the performance indicators with an increasing number of questions from the simulated and the observed data.

Figure 3 shows that the RSME and Bias are difficult to compare due to the importance of the true value θ in the calculation of the performance indicator. However, we see that generally speaking, the Bias has different intercepts but similar slopes. Similarly, we see that the RSME has a similar shape in the first couple of models. However, the difference between the estimated value with 11 models and the true parameter value gradually increases the bias between the true RMSE and the observed RMSE. However, we see that for the SD, the values for the observed model and the simulated model follow the same path. Overall, this suggests that the simulations give

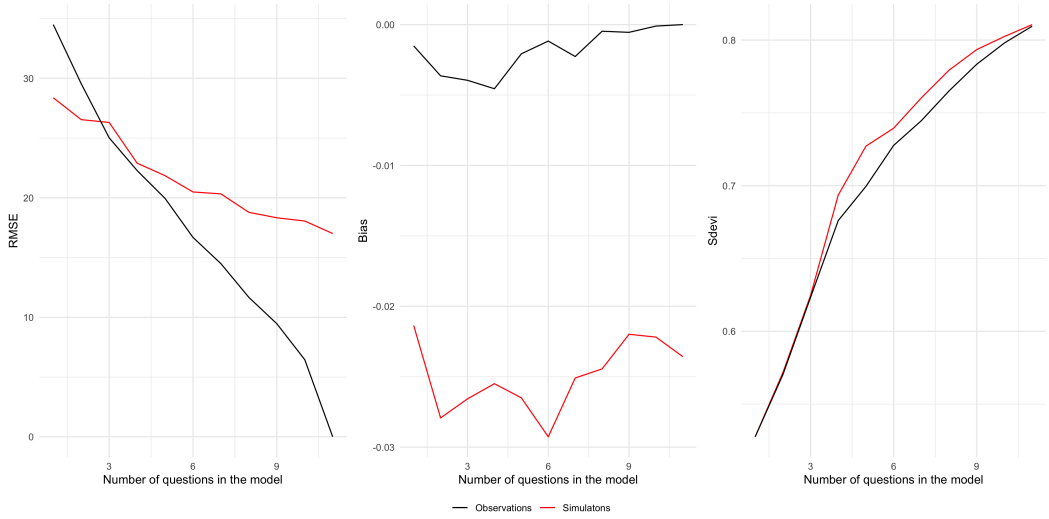

Figure 4: Bias, SD, and RSME for observed and simulated data, for model with one up to eleven questions

an image that is in line with the observation made in the same condition - meaning similar items.

The last remaining question is how the model performance would be affected in the best and the worst-case scenario. In our observed case, we drafted 24 knowledge questions and asked respondents a subset of 11 questions. Our paper suggests that questions with higher discrimination parameters enhance the model performance to measure the ideal positions of actors. Thus in the best-case scenario, we would select the 11 questions with the highest discrimination, and in the worst-case scenario, we would have selected the 11 questions with the lowest discrimination. In our case, the 11 questions with the lowest discrimination vary between .6 and 1.2. On the other side, the questions with the highest discrimination parameter are from 1.4 to 2.5. How much would performance increase between the worst case and the best case scenarios of question selection, and the observed case where we used item randomization and did not select questions? Table 3 presents the value of the different performance indicators for these three scenarios.

The results presented in Table 3 show that performance indicators are better in the best-case scenario than in the worst-case scenario, which shows that selecting questions with higher discrimination parameters increases the model performance. Indeed, we see that the RSME is lower in the best-case scenario. Although the bias is larger in

	Worst selection	Randomization	Best selection
Bias	-0.01606938	-0.0235758	-0.02807936
Standard deviation	0.7854578	0.8106069	0.8406342
Root Mean Squared Error	19.21578	17.01115	15.2809

Table 3: Bias, Standard deviation, and RMSE for models with different question selection

the best-case scenario, this performance simply indicates whether the source of the error is biased. Thus, even though the source of the error is more biased in the models with higher discrimination of questions since the RSME is lower, this indicates that the model still performs better. Furthermore, Table 3 shows that for each performance indicator, the observed model is in between the best and worst-case scenarios. Thus, in addition to enabling the test of a larger pool of questions, question randomization also ensures an average model performance.

In sum, our comparison between the simulation and the observations has shown that question randomization enables to testing of a large set of questions. Selecting questions based on item discrimination improves the model performance, and question randomization ensures an on-average model performance. In the following section, we will summarize the results of the paper and suggest how to measure political knowledge and how to use simulations to evaluate model performance in estimating ideal positions.

Real world data 2: Political knowledge over-time: Selecting questions and comparing knowledge

In this paper, we investigated the model performance of Bayesian IRT to estimate the ideal position of actors and increase the measurement accuracy of political knowledge. We used the Mote-Carlo simulation to first show that using item randomization does not impact the model's performance. In the second step, we show how increasing the discrimination parameter value of questions increases the model's performance. Finally, we have shown how simulations are in line with observed models and how we can use simulation to estimate the gain in performance in different scenarios. In this

last section, we present results from the observational data we collected from three different samples of respondents. We show that our approach can propose a single space in which we can position each respondent's ability to answer political knowledge questions.

We collected data over three survey waves following Swiss popular ballots and in a National panel survey in September 2022, June 2023, and March 2024. The main issue with comparing such measures is that knowledge of specific items may vary between waves. Furthermore, our approach proposes to update the questions based on the posterior prediction of the discrimination parameter of the questions - higher discrimination being associated with better accuracy for the same number of questions. In the different surveys, respondents answered between 8 and 12 knowledge questions, randomized from a sample of questions. However, the sample of questions also varied between the different surveys. In total, we developed 40 questions for the three different surveys. Some were asked in all waves, and others were only asked once.

These questions cover different underlying dimensions of political knowledge. According to Carpini and Keeter (1996), the institutional dimensions, knowledge of the party competition, and policy-specific questions are three important dimensions of political knowledge. Based on their theoretical discussion, we drafted the 40 political knowledge questions relating to the institutional setting and the knowledge of the party competition. In addition, Barabas et al. (2014) theorize knowledge's static and dynamic dimensions in the institutional and policy space. We thus drafted questions related to the static (i.e., number of seats in the parliament, the left-right position of parties) and dynamic (i.e., party of federal councilor, party vote recommendation on specific direct democratic ballots) questions related to the institutional setting and the knowledge of the party competition. Table 4 presents the labels as well as the distribution of the different questions in the three waves we conducted between September 2022 and March 2024.

Table 4 shows that we discontinued some questions and we drafted new questions. There may be several reasons to do so. First, we use the direct democratic setting to ask specific questions about the party's position on the direct democratic ballot. However, these questions need to be updated occasionally as the ballots' timing contextualizes the voters' knowledge about the parties' vote recommendation for these

Question label	September 2022	June 2023	March 2024
Party in favor of affordable housing initiative	Yes	Yes	No
Party against the modification of the stample Law	Yes	Yes	No
Party against the measures to support the media Law	Yes	Yes	No
Party for the expansion of social welfare programs	Yes	No	No
Party opposed to European integration	Yes	No	No
Party advocating for the limitation of immigration	Yes	No	No
Issue does the FDP pay the most attention to	Yes	No	No
Issue do the Greens pay the most attention to	Yes	No	No
Issue does the SP pay the most attention to	Yes	No	No
vote recommendations on the self-determination initiative	Yes	Yes	Yes
vote recommendations on the CO2 law	Yes	Yes	Yes
vote recommendations on the federal law on measures against terrorism	Yes	Yes	Yes
Estimated left-right placement of SVP, SP, GLP and FDP	Yes	Yes	Yes
Most conservative party	Yes	No	No
Which parties merged to form the Center	Yes	No	No
Party with the most seats in the National Council	Yes	No	No
Allocation of seats in the National Council	Yes	Yes	Yes
How can party lists be modified	Yes	Yes	No
Direct-democratic institution requiring 50,000 signatures	Yes	Yes	Yes
What changes when popular initiative adopted	Yes	Yes	Yes
Who elects the members of the Federal Council	Yes	Yes	Yes
How often changes the President of the Confederation	Yes	Yes	Yes
Who is the current President of the Confederation	Yes	No	No
Party composition of Federal Council	Yes	No	Yes
Party of Alain Berst	No	Yes	No
Party of Albert Rosti	No	Yes	Yes
Party of Elisabeth Baume-Schnider	No	Yes	Yes
Party of Guy Parmelin	No	Yes	Yes
Party of Ignacio Casis	No	Yes	Yes
Party of Karine Keller Sutter	No	Yes	Yes
Party of Viola Amherd	No	Yes	Yes
Party of Beat Jans	No	No	Yes
Most used direct democratic institution between 2010 and 2020	No	Yes	No
Rules on finales votes bettween parliamentary chambers	No	Yes	No
How are state councillors selected	No	Yes	Yes
How many state councillors has the largest canton	No	Yes	Yes
How many member does the State council's have	No	Yes	No
Party against initiative on intensive breeding	No	No	Yes
Party against the law on additional financing of Pensions	No	No	Yes
Party against the initiative on Wedding for all	No	No	Yes

Table 4: List of questions asked in the different survey waves.

direct democratic ballots. Thus, although asking about a 2021 ballot in the first (in 2022) or even the second (in 2023) survey wave can be done, it is unlikely that the probability of answering these questions correctly is unrelated to the timing of the survey. We drafted new questions on more recent ballots in the last wave for these items. Second, there may be some institutional changes that require changes in questions. For instance, Alain Berset stepped down as Federal councilor at the end of 2023, leaving his place to Beat Jans. Thus, we needed to remove the question on the political party of Alain Berset for the one on the party of Beat Jans. Finally, the last reason for changing the pool of questions is related to their ability to improve the accuracy of the respondents' ability to parameter posterior predictions. Indeed, as shown in the Monte-Carlo simulation section, the selection of questions with low discrimination parameters does not help to improve the accuracy of posterior prediction for the ability parameters. This is the case, for instance, for the question on the party that merged to form the center or the question on the most used direct democratic institution.

In sum, Table 4 shows a great deal of variation in the pool of questions between surveys. In addition, within each survey, we randomized the pool of questions each respondent saw. The remaining questions are: Are we able to position respondents in different points in time on the same ability space despite this great variation of questions, and did we improve the accuracy of the measures by selecting questions with higher discrimination parameters?

To try to answer these questions, we conducted a Bayesian Item Response Theory with the responses of all respondents to the political knowledge questions we surveyed them on. Figure 5 presents the distribution of the ability parameter in the three surveys.

Figure 5 shows first a similar distribution of ability parameters for respondents across the different surveys. The results in the Figure 5 suggest that, in general, the model gives a relatively similar distribution of knowledge for three sample respondents, at different points in time, surveyed with different questions from a different pool of questions. However, Figure 5 also shows that the "SELECTS" panel distribution of knowledge is slightly higher than for the respondents in another setting. As the SELECTS respondents are drawn from a panel survey meaning that respondents are already several waves into the project, it is not unreasonable to consider that, gener-

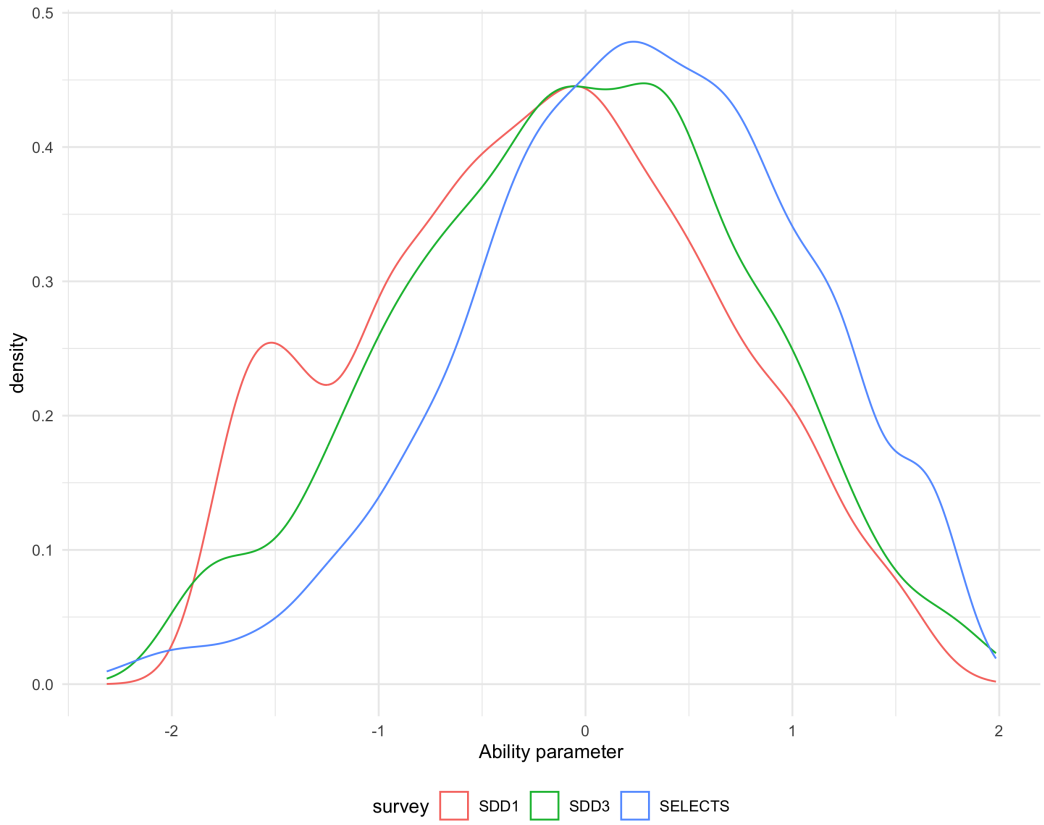

Figure 5: Distribution of ability parameter - political knowledge of respondents - for the different survey waves.

Survey wave	Bias	Standard deviation	Root Mean Squared Error
Randomization of 5 questions			
Selects	0.03516288	0.7451756	19.66273
SDD1	0.04217951	0.7581694	19.43559
SDD3	0.04191394	0.7680746	19.03557
Randomization of 10 questions			
Selects	0.03991732	0.8264079	15.80005
SDD1	0.04142501	0.8440506	15.47164
SDD3	0.0432289	0.8479119	14.74186
Randomization of 15 questions			
Selects	0.04254701	0.8743675	13.16286
SDD1	0.04845151	0.8827296	13.04014
SDD3	0.04278401	0.8853675	13.23503

Table 5: Bias, Standard Deviation and RMSE for the models with 5, 10 and 15 random question in the different survey waves.

ally, the respondents from the panel survey are slightly more knowledgeable than the respondents from the other samples. In sum, the results confirm that the computation method used enables the position of respondents from different samples who answered different questions on the same ability space.

The second question we addressed was whether we could enhance the accuracy of the measure by selecting questions based on the magnitude of the discrimination parameter. To test this assumption, we conducted three additional Monte-Carlo simulations, each considering the posterior estimates of the discrimination and difficulty of items given by the model with the observed data. We ran three separate models for each design, with a random allocation of 5, 10, and 15 questions respectively. Table 5 presents the bias, standard deviation, and root mean squared error of the ability parameter median's posterior estimate - the parameter indicating the knowledge of respondents. The results of these simulations offer promising insights into the potential for improving measurement accuracy, providing a hopeful outlook for the implications of our research.

Table 5 shows quite stable Bias, Standard deviation and RMSE among the different

question samples. However, we notice that with lower number of questions - the models with 5 and 10 questions respectively - the model's accuracy is slightly better for the last survey conducted (with SDD3 questions). Overall, this shows first that dropping and selecting new questions does not impact the overall quality of posterior predictions of our models. Second, and most importantly, it suggest that the selection of questions slightly improve the accuracy of our measures mostly when only a limited number of questions are asked to respondents. In sum, the selection of questions based on the magnitude of the posterior estimate of the discrimination parameter improves the accuracy of posterior estimates for the ability parameter indicating the knowledge of respondents. This can be used in surveys wich aim to measure political knowledge with a limited number of questions to ensure a maximal accuracy of the measure using Item-Response theory models.

Conclusions and Recommendations: Measuring political knowledge with Bayesian IRT models.

The measure of political knowledge has had a lengthly conversation in the literature. Many scholars give recommendations on how to measure political knowledge (Barabas et al., 2014; Boudreau & Lupia, 2011; Carpini & Keeter, 1996; Kraft & Dolan, 2023; Rapeli, 2022). In this paper, we aim to participate to this debate and propose a novel approach based on recent development on the measurement of latent concepts using Bayesian Item-Response Theory models. These models are computationally intensive, and the development of computational power, as well as the updated and more accessible software to use such methods, makes the current approach accessible to researchers.

Our results yield two main points. First, using knowledge questions in a Bayesian IRT framework makes it possible to select empirically the "best" subset of questions to measure the latent concept. Second, using the Bayesian IRT model enables one to compare the knowledge of individuals in the same space using different questions at different points in time. Our findings have significant implications for survey design as researchers may not need to inherit the heavy legacy of the 90's political knowledge questions for their comparative capabilities. Furthermore, far from forgetting what has

been done, this paper shows that using this model and analyzing the posterior estimates of item parameters enables one to objectively evaluate the quality of a question to measure the latent concept of political knowledge.

So, what is the takeaway of our paper? How can the results from this contribution improve the measure of knowledge in a single survey or comparative survey? Based on our findings, we give a set of recommendations to measure political knowledge with survey items. These recommendations enable the comparison but can also be used to measure.

1. *Discrimination scores of existing questions:* To build reliable indicators, we recommend to run IRT model with existing knowledge questions. Survey research has long asked questions on political knowledge. Although several criticism have been made about the measure of political knowledge that these questions generate, some usually asked question can be helpful to measure the political knowledge of survey respondents. By analysing the discrimination parameter of existing questions, researcher can select the ones with the highest discrimination and de-select the others. This creates a strong basis for a valid and reliable measure of political knowledge.

2. *Drafting new questions and use item randomization:* As we have shown, measure of political knowledge can reach high model performance with only a handful of questions. However, to do so, it is necessary to have at least 5 or 6 questions and to consider questions with very high discrimination scores. To reach this efficiency, we argue that the best way is to test as many survey questions as possible. Thus, the second recommendation is to draft relevant survey questions and add them to the pool of existing questions collect prior or in phase 1. Finally, in order to test the largest pool of questions possible, we recommend to use item randomization to test as many questions as possible. We have shown that this randomization does not lower the performance of the models and insures an on average model performance - compared to the best and worst selection of questions.

3. *Estimate the performance of the measure:* The final step of the process is to use the results of the survey conducted by following step 1 and 2, and run an IRT model with it. Then, researchers can use Monte Carlo simulations to estimate the performance of the measure based on the parameter values in the observed model. This way, researcher can give more information on the measure's performance which

enables more reliable way to compare measure between studies.

4. *Repeat the process in a following survey:* With the results of the model and the simulations performed in the third stage, it is possible to repeat the whole process by selecting the questions based on their discrimination scores, add new questions, and create a new measure. Over time selection of better questions enables to efficiently improve the performance of the model and the validity of the ideal position estimates. One draw back is that respondents do not answer the same question in every survey, thus making the comparison more complicated. While this is true for the construction of scales, measure of ideal position with Bayesian IRT only requieres a hanful of similar items to estimates ideal position on the same latent space. Thus, as long as researchers include a handful of already used questions, overtime comparisons are still possible.

In sum, this paper has shown that implementing questions randomization when measuring political knowledge does not affect the Bayesian IRT models' performances and enables to pick and choose questions with the highest discrimination parameter to increase the accuracy of ideal position estimates without using too much survey space.

References

Barabas, J., Jerit, J., Pollock, W., & Rainey, C. (2014, November). The Question(s) of Political Knowledge. *American Political Science Review*, *108*(4), 840–855.

Bathelt, S., Jedinger, A., & Maier, J. (2016). Politische kenntnisse in deutschland: Entwicklung und determinanten, 1949–2009. *Bürgerinnen und Bürger im Wandel der Zeit: 25 Jahre Wahl-und Einstellungsforschung in Deutschland*, 181–207.

Boudreau, C., & Lupia. (2011). Political Knowledge. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge Handbook of Experimental Political Science* (pp. 171–183). Cambridge: Cambridge University Press.

Bullock, J. G., & Rader, K. (2022). Response options and the measurement of political knowledge. *British Journal of Political Science*, *52*(3), 1418–1427.

Carpini, M. X. D., & Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science*, 1179–1206.

Carpini, M. X. D., & Keeter, S. (1996). *What americans know about politics and why it matters*. Yale University Press.

Caughey, D., O'Grady, T., & Warshaw, C. (2019, August). Policy Ideology in European Mass Publics, 1981–2016. *American Political Science Review*, *113*(3), 674–693.

Caughey, D., & Warshaw, C. (2015). Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model. *Political Analysis*(2), 197–211.

Elff, M. (2009). Political knowledge in comparative perspective: the problem of cross-national equivalence of measurement. In *Mpsa 2009 annual national conference* (pp. 2–5).

Hopkins, V., Kagalwala, A., Philips, A. Q., Pickup, M., & Whitten, G. D. (2024). How do we know what we know? learning from monte carlo simulations. *The Journal of Politics*, *86*(1), 000–000.

Kleinberg, M. S. (2022). Measuring political knowledge in online surveys: How question design can improve measures. *Public Opinion Quarterly*, *86*(3), 736–747.

König, C., & van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review*, *70*(4), 486–509.

Kraft, P. W., & Dolan, K. (2023). Asking the right questions: A framework for developing gender-balanced political knowledge batteries. *Political Research Quarterly*,

76(1), 393–406.

Lizotte, M.-K., & Sidman, A. H. (2009). Explaining the gender gap in political knowledge. *Politics & Gender*, *5*(2), 127–151.

Lupia, A. (2016). *Uninformed: why people know so little about politics and what we can do about it*. New York: Oxford University Press.

Munzert, S., Ramirez-Ruiz, S., Barberá, P., Guess, A. M., & Yang, J. (2024). Who's cheating on your survey? a detection approach with digital trace data. *Political Science Research and Methods*, *12*(2), 390–398.

Munzert, S., & Selb, P. (2017). Measuring political knowledge in web-based surveys: An experimental validation of visual versus verbal instruments. *Social Science Computer Review*, *35*(2), 167–183.

O'Grady, T. (2019, October). How do Economic Circumstances Determine Preferences? Evidence from Long-run Panel Data. *British Journal of Political Science*, *49*(4), 1381–1406.

O'Grady, T., & Abou-Chadi, T. (n.d.). Not so responsive after all: European parties do not respond to public opinion shifts across multiple issue dimensions. , 7.

Pietryka, M. T., & MacIntosh, R. C. (2013). An analysis of anes items and their use in the construction of political knowledge scales. *Political Analysis*, *21*(4), 407–429.

Pietryka, M. T., & MacIntosh, R. C. (2022). Anes scales often do not measure what you think they measure. *The Journal of Politics*, *84*(2), 1074–1090.

Rapeli, L. (2022, August). What is the best proxy for political knowledge in surveys? *PLOS ONE*, *17*(8), e0272530.

Tsai, T.-h., & Lin, C.-c. (2017). Modeling guessing components in the measurement of political knowledge. *Political Analysis*, *25*(4), 483–504.

van Heerde-Hudson, J. (2020). Political knowledge: Measurement, misinformation and turnout. In *The routledge handbook of elections, voting behavior and public opinion* (pp. 369–382). Routledge.