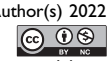
Article

EUP

European Union Politics
2022, Vol. 23(2) 171–191


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14651165221080400
journals.sagepub.com/home/eup
SAGE

Measuring public support for European integration using a Bayesian item response theory model

Michele Scotto di Vettimo
Department of Politics, University of Exeter, Exeter, UK

Abstract
This study proposes the use of Bayesian item response theory models to measure aggregate public support for European integration. This approach addresses the limitations of other indicators and produces valid estimates of public attitudes over long time periods, even when available indicators change over time or present interruptions. I compare Bayesian item response theory models with alternative approaches used in the study of support for European integration, and demonstrate that they produce more accurate estimates of latent public opinion. The estimates are validated by showing their association both to alternative public opinion measures and to the vote share of Eurosceptic parties across Europe. I show that Bayesian models solve unaddressed issues like ensuring cross-country comparability of the estimates and modelling responses with multiple answer options.

Keywords
European integration, Euroscepticism, item response theory, public opinion, public support

Introduction

Against a background of growing politicisation of the European Union (EU), public attitudes towards EU integration have become central to understanding European-level policy-making (e.g. Bølstad, 2015; Toshkov, 2011; Wratil, 2019) as well as national-level party strategies (e.g. Hutter and Grande, 2014). Yet, the precise measurement of

Corresponding author:
Michele Scotto di Vettimo, Department of Politics, University of Exeter, Amory Building, Rennes Drive, EX4 4RJ, UK.
Email: m.scotto-di-vettimo@exeter.ac.uk

aggregate public opinion towards the EU has received relatively limited attention. A wide range of measures have been employed, but few of these efforts have been justified with reference to established conceptualisations of EU support (Hobolt and De Vries, 2016).

In particular, EU support possesses two properties calling for a careful empirical treatment. Firstly, people might like certain aspects of the EU, but not others and present ambivalent attitudes towards the EU integration rather than just showing support or opposition (De Vries, 2013). Secondly, EU support has a multilevel in nature. According to the 'benchmark theory', EU attitudes are the result of a comparison between citizens' EU and national evaluations. National contextual factors are 'part-and-parcel of the way in which people evaluate the EU level, and vice versa' (De Vries, 2018: 28). Hence, survey indicators tapping into EU support are rarely understood in the same way across countries, raising serious concerns for the cross-national comparability of measures of EU support based on survey indicators (Ariely and Davidov, 2011: 272).

Yet, existing measurement strategies are rarely capable of dealing with these two issues satisfactorily. Scholars exploring aggregate-level EU support have mainly relied on existing single-question indicators, selected mostly because they constitute the only data source that allows for cross-national and longitudinal comparisons (Hobolt and De Vries, 2016: 416). However, the cross-national 'equivalence' of these indicators is rarely assessed. Furthermore, they often have to be retrofitted to suit a particular analysis (Anderson and Hecht, 2018: 621).

Dimension reduction techniques, like the Dyad Ratios (DR) algorithm (Stimson, 1991, 2018), have also been employed to estimate latent public EU support from a set of different indicators. In the EU context, the algorithm represents the most advanced approach so far employed to estimate EU support starting from aggregate-level data (Anderson and Hecht, 2018; Guinaudeau and Schnatterer, 2019). Nevertheless, DR estimates inherit all the problems related to the lack of cross-national comparability of the raw indicators, and the algorithm is poorly designed to deal with neutral answer options.

Thus, I propose the use of Bayesian item response theory (IRT) models for the estimation of public EU support starting from aggregate-level data. The EU context represents a promising area of extension for this technique. Existing surveys provide for identically worded questions over relatively long, though irregular, time periods. Additionally, current approaches to the measurement of aggregate-level EU support fail in appropriately dealing with key issues like ambivalence and the comparability over time and across countries. Bayesian IRT, instead, can produce estimates of aggregate EU support with a sounder theoretical grounding (Caughey and Warshaw, 2015; McGann, 2014), and that are more tailored to the established conceptualisations of EU support.

This article contributes to the literature about the measurement of aggregate-level EU support by applying, for the first time, a Bayesian IRT model. I explain why this approach is superior to other available alternative techniques as it starts from an explicit model of individual behaviour, offers a theory-based approach to deal with neutral responses and to appropriately capture more ambivalent attitudes towards the EU, and produces measures of public preferences that are comparable both over time and between countries (McGann et al., 2019). I show that these advantages enable Bayesian IRT to produce measures of EU support that are both more precise and more in line with the conceptual properties of

EU support. Finally, this technique is applied to the examination of aggregate support for European policy integration in specific areas, a domain currently characterised by a lack of appropriate data (Zhelyazkova et al., 2019).

Measuring support for Europe: Challenges and approaches

The increasing politicisation of the EU has led scholars to incorporate the role of the public into theories of supranational integration and to study the nature of EU attitudes (Hobolt and De Vries, 2016; De Vries, 2018). Yet, 'the precise measurement of public attitudes towards European integration has […] received somewhat limited consideration' (Hobolt and De Vries, 2016: 416). On the conceptual level, the literature highlighted two features of EU support that should be taken into account when measuring this concept: the ambivalence of public attitudes towards the EU and their multilevel nature.

Firstly, the literature on attitudinal ambivalence questions the assumption that citizen's attitudes towards the EU can be categorised as either positive or negative (Stoeckel, 2013; De Vries, 2018). In practice, individuals may simultaneously like and dislike a certain aspect of the EU, for different reasons (Stoeckel, 2013; De Vries and Steenbergen, 2013). These ambivalent attitudes 'are held with less certainty, […] tend to be less stable over time, [and] are more likely to be driven by whatever considerations are momentarily salient' (Stoeckel, 2013: 24). Yet, far from indicating political detachment, ambivalence increases in more sophisticated individuals and when political elites are more divided on EU issues (cf. De Vries, 2013). Ultimately, attitudinal ambivalence denotes a fundamental uncertainty about public stances on European issue (De Vries and Steenbergen, 2013: 122), and calls for a careful empirical treatment (cf. Guinaudeau and Schnatterer, 2019: 1190). Focusing just on the relative balance between pro- and anti-EU positions can give the impression of clear swings from mostly favourable to mostly unfavourable (or vice versa) attitudes, whereas in reality, public opinion towards the EU is not so clear-cut (De Vries and Steenbergen, 2013: 137). Thus, an appropriate empirical treatment of attitudinal ambivalence is of key importance for a correct measurement of aggregate EU support.

Secondly, recent developments in the conceptualisation of EU support explicitly conceive it as multilevel in nature, thus emphasising the role of national contextual factors in shaping it. The 'benchmark theory' of EU support states that EU attitudes are the result of a comparison between citizens' EU and national evaluations. Hence, 'people's national evaluations are themselves part-and-parcel of the way in which people view and evaluate the EU level, and vice versa' (De Vries, 2018: 28). Therefore, we should not expect measurement instruments to work in the same way across all countries. For instance, people might express more support for EU integration if they think that it adds value to the policy-making or service provision in a specific field and, therefore, it is to be preferred to the 'alternative state' of no integration. However, the extent to which the EU can *add* value crucially depends also on what kind of performance national-level institutions can attain. Hence, the rosier the 'alternative state', the more 'difficult' is to express support for EU integration.

These counter-factual evaluations focus on the national context, and affect the extent to which an instrument (e.g. a survey question) is associated with the latent concept being measured (e.g. EU support) across multiple countries. Should this association vary across countries, the items used to measure the concept of interest do not work in the exactly similar way across the various contexts (a situation known as lack of 'measurement invariance'; cf. Heath et al., 2005). In this case, observed cross-national differences can be due to both real differences in the presence of the underlying concept and to differences in the functioning of survey questions (cf. Ariely and Davidov, 2011).

Therefore, to produce estimates that satisfy the conceptual properties of EU support, the measurement technique should deal with the issues of attitudinal ambivalence as well as cross-national (and cross-time) comparability of the estimates in an explicit way. Yet, existing approaches to the measurement of aggregate-level support for the EU fare poorly with regard to both tasks. Below I discuss two common approaches to the measurement of aggregate-level EU support – the use of single-question indicators and of dimension reduction techniques – and assess their ability to deal with attitudinal ambivalence and the lack of measurement invariance.

Single-question indicators of EU support

To ensure cross-national comparability, two necessary conditions are that available surveys: (a) employ the same question wording across countries; and (b) are administered with relatively similar timings. Cross-national surveys represent, therefore, the most obvious candidates as data sources and, among these, Eurobarometer (EB) surveys have become by far the most used.

The EB surveys have been conducted at least once every semester since September 1973 in all EU member states and most candidate countries, and EB questions tap into various aspects of public EU attitudes. Nonetheless, changes in question wordings and interruptions in the administration of a specific question make the use of particular items impractical, thus forcing the analyst to choose the 'best' single-question indicator among the few available ones. Ultimately, scholars had to rely on the few indicators available over longer periods (e.g. Eichenberg and Dalton, 2007; Franklin and Wlezien, 1997; Hobolt and De Vries, 2016), which are often retrofitted to suit the needs of a particular analysis (Anderson and Hecht, 2018).

However, the use of faithfully translated questions is not enough to guarantee that measurement scales are invariant across countries (Ariely and Davidov, 2011: 272). Although there are various approaches to deal with the problem of the 'contextualisation' of survey items (Heath et al., 2005), measurement invariance cannot just be assumed a priori (Ariely and Davidov, 2011: 282). In the context of a multilevel nature of EU support (De Vries, 2018), the use of identically worded questions does not ensure that the estimated support is comparable across countries. Furthermore, each of these items tap into a specific aspect of EU attitudes (Anderson and Hecht, 2018: 618), and it is problematic to consider them as viable ways of capturing EU support more broadly.

Dimension reduction techniques and latent EU support

To address the shortcomings related to the specificity and gaps of single-question indicators, scholars have started to employ measures combining different survey items related to the latent concept of interest.

The combination of multiple items allows for the measurement of the public's latent attitudes that cannot be properly captured by single-item indicators (Anderson and Hecht, 2018). Yet, the fact that these items are usually asked over non-overlapping time frames makes it impossible to estimate the covariance matrix, which is at the base of scaling techniques like factor analysis (Stimson, 2018: 203). Hence, other methodologies (cf. Beck, 1989; Claassen, 2019; Voeten and Brewer, 2006) have been developed to deal with fragmented aggregate-level data and, among these, only the DR algorithm (Stimson, 1991, 2018) has been extended to the estimation of cross-national series of aggregate EU support (Anderson and Hecht, 2018; Guinaudeau and Schnatterer, 2019).

The algorithm was developed primarily to deal with the problem of data series interruption, and builds on the assumption that the variations in the responses to available questions 'reflect something more general' than attitudes towards the particular item asked (Stimson, 2018: 203). The DR algorithm estimates this latent construct by imitating principal component analysis, and offers an ingenious procedure to get around the problem posed by fragmented survey data preventing the estimation of the covariance matrix required for the principal component technology to operate.

The DR algorithm has been extended to the study of EU support only recently (Anderson and Hecht, 2018; Guinaudeau and Schnatterer, 2019). In brief, it starts from the ratio of the share of positive responses over the total of positive and negative given to each question, and then calculates the rate of change of this quantity between two consecutive appearances of the question. These ratios represent a first estimate of the relative change of the underlying latent attitudes over time, and are used in an iterative process to calculate the 'mood' in each time point (Stimson, 2018).

The algorithm currently represents the most advanced technique used to estimate long trends of EU support from aggregate-level data. Nonetheless, it has been criticised on different grounds. For instance, the algorithm assumes that any change in the raw indicators is linearly associated with changes in the underlying 'mood' (McGann, 2014: 117). However, individual-level models of response suggest that the relationship between the latent construct of interest and the observed response patterns indicator measure is likely to be a non-linear one (Caughey and Warshaw, 2015; McGann, 2014; Nunnally and Bernstein, 1994).

With regard to the EU, the DR algorithm fails in filling one important gap and in addressing two key issues related to the measurement of aggregate EU support. Firstly, a key limitation of public opinion and political representation studies in the EU is the lack of measures of public support for EU integration in specific policy areas and over a long period of time (Zhelyazkova et al., 2019: 1718). However, the extension of the DR algorithm to the EU case has not yet addressed this important gap.

Secondly, the DR algorithm offers only a suboptimal way of dealing with more ambivalent attitudes, as there is no way to explicitly model middling responses.

Yet, when the share of neutral responses is relatively high, the decision to discard them leads to the loss of potentially relevant information on citizens' attitudes and makes the estimated levels of support appear more extreme than they actually are. For instance, the Anderson and Hecht's DR algorithm estimates for Germany illustrate the impact of this problem. The omission of intermediate responses makes the raw items more prone to marked swings (cf. De Vries, 2013), and all these features of the raw items are eventually transferred to the estimated mood.

Figure 1(a) shows that the DR estimate closely tracks the movements of the 'membership' question (Anderson and Hecht, 2018: 627). This indicator represents the paradigmatic case of an item that can be better handled by the proposed Bayesian IRT approach. It has a neutral answer option which, on average, represents the 26.5% of the responses. Therefore, although only 58.7% of Germans have, on average, answered that membership is 'a good thing', by focusing only on the ratio of positive responses over non-neutral ones, the average share of positive responses is a more extreme 85.3%.

Given the high 'weight' that the DR algorithm attributes to the membership question, the estimated mood inherits many features of this indicator. Consequently, as Figure 1(b) shows, the decline in the German 'mood' occurs between 1991 and 1997, when the share of respondents saying that EU membership was 'a good thing' decreased from 71% to 36%. Still, respondents largely migrated to neutral, rather than negative, attitudes towards membership, as the proportion of 'don't know' and neutral answers grew from 23% to 48% in the same period, whereas the share of negative responses only increased from 6% to 15%.

Thirdly, the algorithm cannot deal with the lack of cross-national scale equivalence between each raw indicator (Heath et al., 2005), hence compromising the comparability of the estimates (Ariely and Davidov, 2011; Avvisati et al., 2019). All constructs based on single-question indicators become non-comparable across countries if the measurement

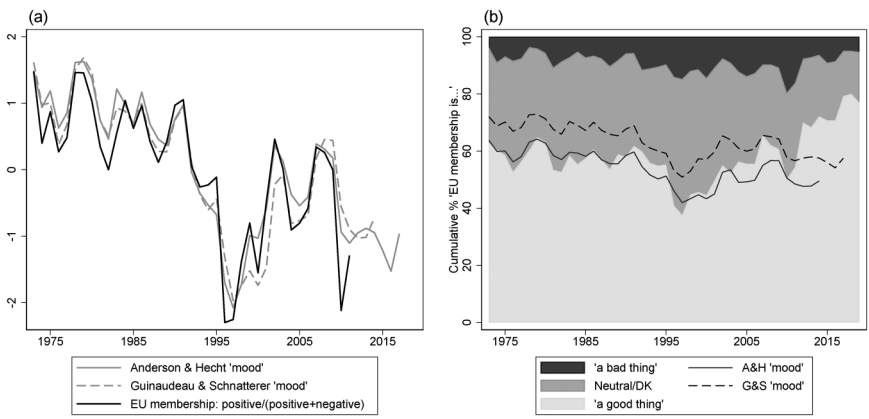


Figure 1. Dyad Ratios (DR) algorithm estimates compared to the 'membership' question in Germany.

invariance of the raw indicators is not explicitly addressed. The DR algorithm ensures that items whose trends are closer to those of the latent concept carry more weight during the estimation process. Yet, this procedure is conducted within each country analysed, though the same item might carry a lot of weight in one context and very little in another. Thus, differences in the estimated mood will indicate differences in the latent trait of interest only as far as the underlying items work similarly across groups, but there is no way of adjusting DR estimates accordingly if this is not given (Solt, 2020).

To better deal with these problems, I propose the use of Bayesian IRT models to create comparable country-specific measures of EU support starting from aggregate-level data. In the following, I adapt the approach of McGann et al. (2019) to extend it to the EU case. IRT models offer a valid approach to address two empirical issues faced by studies of EU public opinion – the fact that single-question indicators offer only irregular and partial measures of EU support and the lack of policy-specific measures of support. Moreover, an IRT-based approach is superior to other available techniques in fulfilling these tasks because it incorporates an individual-level model of response, allows for the explicit treatment of neutral responses, and uses question-specific features to recover the level of the estimated latent dimension allowing comparability across EU countries. Therefore, the approach offers a substantive contribution to the study of public EU attitudes by allowing for a measurement technique that is more in accordance with the established conceptualisations of EU support.

An item response theory model of support for Europe

IRT models were developed in psychometric theory to measure latent traits and attributes of individuals. These models have recently been adapted to deal with aggregate-level data, and used to estimate latent public preferences from the percentages of responses to a set of different survey items (McGann, 2014; McGann et al., 2019).

Hence, the intuition behind Bayesian IRT for aggregate survey data starts from an explicit individual-level model of response. Imagine we want to measure respondent i's latent support for Europe (henceforth the attribute) by asking her a few questions related to different aspects of EU support. The probability that she answers in a pro-European way ('correct' response) is determined by three factors. First, all things being equal, the higher the individual's attribute (θ), the more likely she is to give a pro-EU answer to the question asked. Secondly, each question q has its intrinsic difficulty b_q. For instance, we would expect it to be more 'difficult' to answer positively to a question asking whether the EU should become a federation rather than to a question asking whether more EU action is needed on matters related to development cooperation. Whenever a respondent's attribute exceeds the difficulty of the question, she is more likely to answer correctly than incorrectly. Finally, changes in the respondent's attribute affect the likelihood of answering correctly, but the effect of these changes is not constant across items. Item discrimination, a_q, shapes the relationship between the probability of a correct response and the latent attribute (see the Online appendix for a more detailed discussion).

Bayesian IRT with aggregated data incorporates this individual-level model into an aggregate-level one. By knowing each question's difficulty and discrimination, as well as the distribution of latent EU support in the population, it would be possible to predict the percentage of pro-European responses to question q at time t. Yet, we find ourselves precisely in the opposite situation, where only the number of pro-EU responses is a known quantity. A Bayesian approach, therefore, is used to determine the most likely value of the question difficulty and discrimination of latent EU support, *given the observed response patterns*.[1] Bayesian IRT starts from some prior information about the probability distribution of these unknown quantities ('parameters'), then looks at the observed data (actual responses) and estimates the probability ('likelihood') that the data could be generated by a model with a given set of parameters (Kruschke, 2014).

As long as at least one question is asked at a given time point, by knowing its difficulty and discrimination parameters, it is possible to recover the distribution of the population's latent EU support for that time point. Hence, Bayesian IRT models can reduce the reliance on a particular item and overcome the problem posed by the interruptions and missingness in the series of the various questions used as input, whilst also producing uncertainty measures around the estimated model parameters.

The advantages of the Bayesian IRT methodology are manyfold. Firstly, IRT does not assume that the responses to a given indicator and the latent attribute of interest are linearly related. In the IRT model, if latent support is already quite high, we would expect a large movement towards even more positive attitudes to have little effect on the proportion of observed pro-European responses (virtually everyone is manifesting support already). If, however, latent support is more ambivalent, a change in latent mood could have a greater effect (McGann, 2014). Moreover, questions with different discrimination parameters react differently to changes in the latent mood.

Secondly, the IRT approach allows for a sounder treatment of those neutral attitudes representing a non-trivial share of the responses to some EB questions. In the IRT framework, it is possible to estimate more difficulty parameters for each item. In this case, a $b_{2,q}$ measuring the amount of θ_i required to answer the question in a positive way and a $b_{1,q} < b_{2,q}$ representing the difficulty of answering q at least in a neutral way (McGann et al., 2019: 51). This property represents an important advantage of Bayesian IRT models, and allows an explicit modelling of the middle response category and of more ambivalent of EU attitudes (De Vries, 2013; Stoeckel, 2013).

Finally, Bayesian IRT models improve the interpretation and comparability of the estimated constructs. Firstly, they use the information coming from items' parameters to estimate a measure of latent support with an interpretable metric. Questions with very high positive response rates and a low difficulty or discrimination do not pull the estimated support upwards, as for these questions a higher share of positive responses can either be expected or it should not be considered as representative of the underlying mood. Secondly, the accurate cross-national comparison of EU support requires that the scales used are equivalent (Ariely and Davidov, 2011; Heath et al., 2005). The benchmark theory of EU support (De Vries, 2018), however, suggests that national-level contextual factors are very likely to influence how the indicators work in each context. Hence, the process of preference formation suggests that the assumption of measurement

invariance might be violated and, as a consequence, the cross-national comparability of EU support compromised. However, Bayesian IRT models can deal with the problem of measurement invariance by applying appropriate constraints to the estimated item parameters, so that items work similarly across countries.

Therefore, Bayesian IRT can be used to produce measures of the population's latent EU support with a grounding in a model of individual-level behaviour and without the unrealistic assumption of a linear relationship between changes in expressed preferences and changes in the latent attribute. These models can be adapted to the treatment of neutral responses and produce estimates of support that can be compared across countries, two features of key importance in the EU context. Though the IRT approach requires the same type of input data as other techniques, such as the DR, it opens the way to new applications in the estimation of overall or policy-specific support for EU integration to reflect more closely the conceptual properties of EU support.

The formalisation of the model is reported in the Online appendix. The model represents an improvement over recent applications of Bayesian IRT on aggregate-level data (McGann, 2014; McGann et al., 2019). In particular, it combines McGann et al. (2019)'s solution to deal with non-binary responses with Claassen (2019)'s dynamic specification, and proposes an efficient approach to address the lack of cross-national equivalence.

Estimates of public support for Europe

I define EU support in line with the two-dimensional conceptualisation proposed by De Vries (2018). In particular, I take into account both attitudes towards the EU as a polity (regime support) and preferences towards EU policy integration (policy support). To implement the Bayesian IRT model, I identify 201 EB questions tapping into one of these two dimensions, which were asked at least twice between September 1973 and July 2020 (see the Online appendix). In line with De Vries (2018)'s definition of regime support, the first group of questions comprises items on trust in the EU or in its institutions, the evaluation of one's country membership in the EU, and dispositions towards European unification. The second and bigger (about 81%) category of questions, instead, measures preferences for the level of government which should be responsible for a particular policy, whether the EU should do more or less in a certain field, and whether the respondent favours or opposes the creation of an EU common policy in a given domain.

The size of the dataset ranges from 594 question-year dyads for Croatia to 1448 for the nine countries already member in 1973.[2] I employ data from 168 EB surveys and, if a question is asked more than once in the same year, responses are averaged over the available surveys. For each observation, the percentage of responses expressing positive attitudes towards Europe is used as input for the Bayesian IRT model. When a neutral response option is present, the percentage of respondents giving at least a neutral response is also used as input (McGann et al., 2019). Refusal and 'don't know' answers are excluded.

I implement the Bayesian IRT model using Gibbs sampling, a randomised algorithm used to generate sequences of observations from approximations of the probability

distributions of the desired parameters. The algorithm starts from an arbitrary value of one of the parameters of interest, then generates an instance of another parameter, conditional on the current values of all the other parameters and variables, and progressively approximates the joint probability distribution of all the unknown quantities (Kruschke, 2014: 137).

To ensure cross-national comparability of the estimated latent support, the questions asked should work similarly across the various countries. That is, their parameters should not differ much between countries. McGann et al. (2019) estimated a single model with country dummies for the latent dimension, so as to constrain question parameters to same values for all the countries considered, whereas the dummy shifts the position of the latent attribute in each country.

However, the same solution would not be optimal for the task at hand, as the assumption of perfect measurement invariance across a wide range of countries is a too unrealistic imposition. Therefore, I build on the concept of 'approximate measurement invariance', and allow question parameters to vary within a narrow 'wiggle room' across countries (Avvisati et al., 2019: 4-5). First, I estimate a Bayesian IRT model using opinion data representative of the whole EU public. Then, for each country, I use the mean and the standard deviation of the EU-average item parameters to provide informative priors for the country-specific item parameters, so that items show a similar behaviour across countries (see the Online appendix). The models are estimated with the open source software JAGS (Plummer, 2003), and convergence is checked using the Geweke and Gelman-Rubin diagnostics.

The estimation produces annual measures of EU support for all EU member states and the EU as a whole from 1973 (or from when EB data are available) to 2020. The series are plotted in Figure 2. Previous analyses of EU support have identified 'core' and 'periphery' trends of public opinion (Bølstad, 2015), with founding members consistently more supportive of European integration than countries joining in 1973. This pattern emerges also in Figure 2, with Denmark and especially the United Kingdom showing lower levels of support if compared to the six founding member states. Concerning the over time trends, the 1980s have been characterised by a steady rise in EU support, which peaked at the moment of the adoption of the Maastricht treaty in 1992, and eventually fell due to the so called 'post-Maastricht blues' (Eichenberg and Dalton, 2007). Again, these trends can be seen in most of the countries already member in late 1980s, as well as for the EU as a whole. Similarly, the negative impact of the eurozone crisis on the level of support is also visible from 2009 onwards, particularly for the more affected Southern member states.

Therefore, estimated trends are in line with the general understanding of EU support, and are reassuring in terms of face validity of the measure. More generally, Figure 2 highlights few other patterns. In most countries EU support has increased in recent years, though it is just returning to its pre-2008 levels. Only in countries like Austria, Finland, Germany, Ireland, Portugal, the United Kingdom and few others the estimated EU support is higher now than it used to be before the sovereign debt crisis. Also, there are member states – like Czech Republic, Greece, France and Italy – in which EU support did not bounce back to the pre-crisis levels at all.

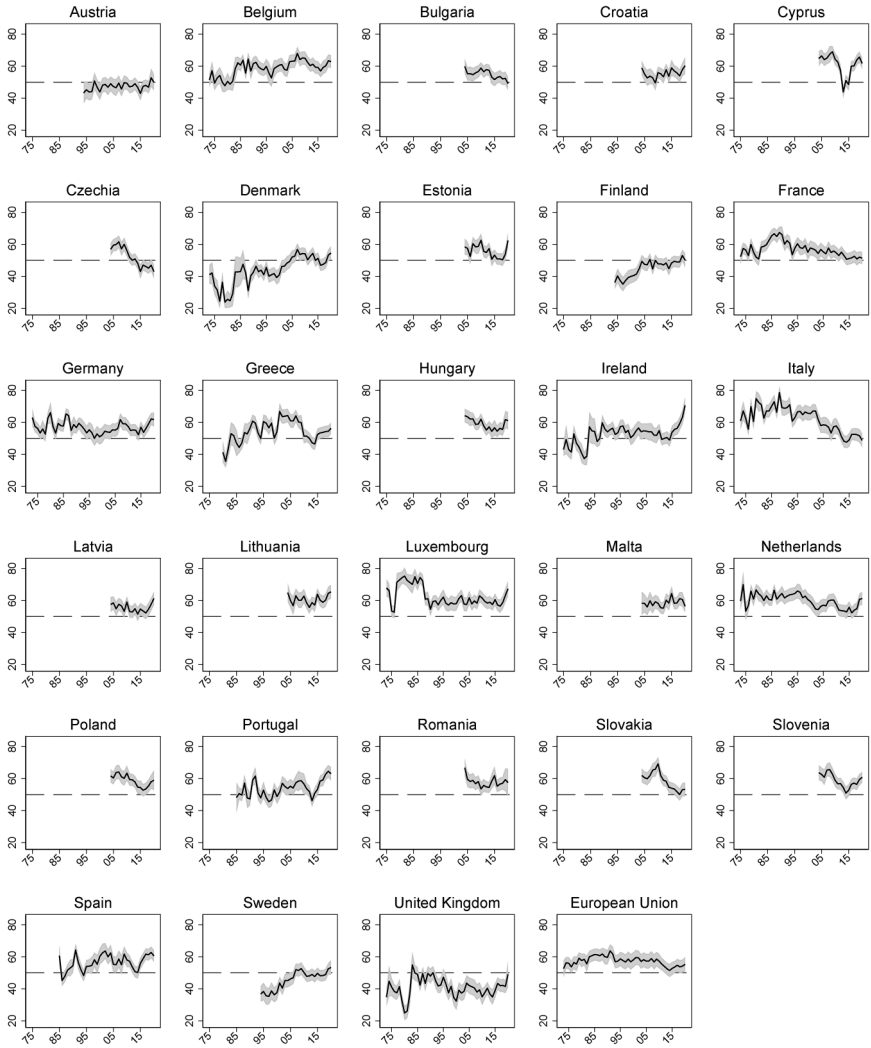


Figure 2. EU support estimated with a Bayesian IRT model. Shaded area indicates ± 2 standard deviations. IRT: item response theory; EU: European Union.

Finally, the trends of Figure 2 can be compared with those estimated using the DR algorithm and the same set of raw indicators (see the Online appendix). As expected, there are stark differences in the levels of the estimated support. By ignoring middling responses, the DR algorithm produces much higher levels of support in all countries. Additionally, only in eight out of 29 cases the correlation between the two measures is above or close to 0.85. In a similar number of instances it is equal to 0.4 or lower, with

two countries even exhibiting 0 or negative correlations. Overall, there seem to be only a slight convergence between the two estimates. They always capture very different levels of EU support (with an average difference of 14.8 points) and in a majority of cases the over time trends are also far from identical.

Support for EU integration in specific policy domains

Another substantive contribution brought by the use of Bayesian IRT models is the measurement of policy-specific support for EU integration. The EU landscape is characterised by the lack of long trends of public preferences towards integration in specific policy fields (Zhelyazkova et al., 2019). A single-question indicator capturing, for example, preferences for economic policy integration comparable in terms of continuity to the 'membership' question does not exist. Rather, there are different items, asked with varying regularity, measuring support for a direct EU taxation system, the EU being responsible for domestic redistributive policies and so on. These items can be used in a Bayesian IRT model to produce estimates of public support for EU integration in the economic domain, by combining the information coming from different items which, on their own, offer only an issue-specific information about public preferences for economic integration.

Of course, the same problems characterise other policy domains as well. Yet, to the extent that it is possible to identify an appropriate set of indicators, this solution can be replicated for various policy areas (see McGann et al., 2019). This exercise can be used to learn more about cross-national dynamics of EU support in specific areas. For instance, Sánchez-Cuenca (2000) shows that EU support is inversely related to public spending on social protection, arguing that citizens see EU integration as a threat to a

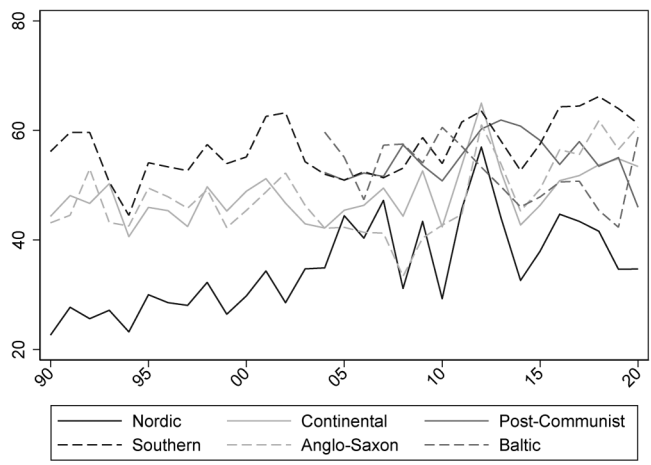


Figure 3. Support for European Union (EU) integration in social affairs by type of welfare model.

powerful welfare state. With the proposed strategy, we can assess whether this dynamic is also visible with regard to aggregate preferences for EU integration in social affairs more specifically.

The country figures are reported in the Online appendix. Figure 3 plots the average support for six groups of countries reflecting an adaptation of Esping-Andersen (1990) welfare models to EU-28 countries (Zaidi et al., 2018). Despite some signs of convergence in recent years, the difference between countries with comparatively higher levels of transfers and equity like the Nordic ones and Southern, Baltic and Post-Communist countries is particularly suggestive, as the pattern is in line both with the literature on the relation between EU support and welfare policies and with the implications of the benchmark theory of EU support.

Similarly, it is also possible to explore the trends in EU support with regard to the increasingly salient domain of migration policies. Figure 4 shows the average trends across six groups of member states, identified on the basis of their geographical location. In practice, the groupings are similar to the ones presented above. The only difference is that Central and Eastern European (CEE) member states are now divided into Visegrad and Non-Visegrad countries.[3] The main patterns show that Nordic and Anglo-Saxon countries have traditionally held more sceptic views towards policy integration in migratory affairs, whereas Southern member states display a steady high-level support for such initiatives. The effect of the so-called 'migrant crisis' is also visible at around 2015. Nordic, Anglo-Saxon and Continental countries converge towards higher levels of support for EU integration in the field of migration policies, whereas CEE countries adopt more hostile views. In particular, despite timid signs of convergence in the last decade, Visegrad countries are now the ones showing the lowest support for integration in this domain.

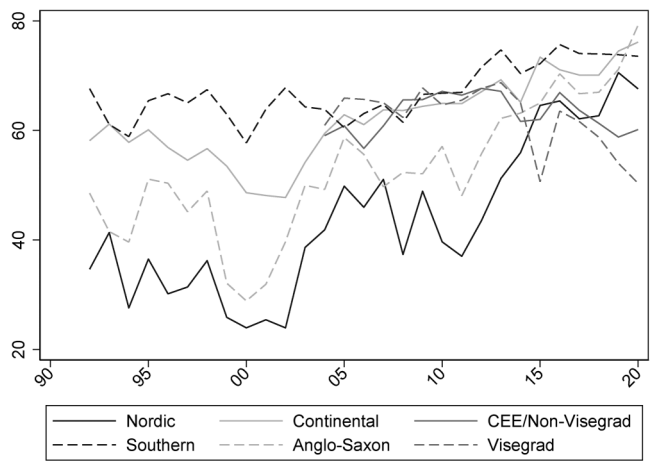


Figure 4. Support for European Union (EU) integration of migration policies in six groups of countries.

The Online appendix reports other exercises of this kind for economic and fiscal policy and competition, consumer protection and single market rules. This exploration of policy specific series of EU support enables a more nuanced understanding of the phenomenon – as it allows scholars to single out specific objects of EU support – and highlights the degree of sophistication of the various national publics, which are capable of orienting their support or opposition towards specific aspects of EU integration. Moreover, as the member states that are more (or less) supportive of EU integration vary from one policy domain to another, these measures can be useful in studying whether and how citizens' preferences influence the log-rolling and vote-trading activities between national governments during EU-level decision-making processes.

Measurement validity

To fully assess the validity of the estimates, I conduct more stringent validation procedures (Adcock and Collier, 2001). First, I compare the IRT measure with the trends of few established single-question indicators widely employed in the literature exploring aggregate trends of EU support ('convergent validity'). Then, I assess the fit that the IRT measure has with the observed data by exploring to what extent it can be used to reconstruct the response patterns of the raw indicators. Finally, I assess the 'construct validity' of the estimates – that is the association with measures of theoretically related phenomena – by examining the association with vote share of Eurosceptic parties in national and European elections.

Convergent validity

If the product of the Bayesian estimation provides a valid measure of EU support, then it should also be correlated to other valid indicators of this concept. Figure 5 shows the IRT

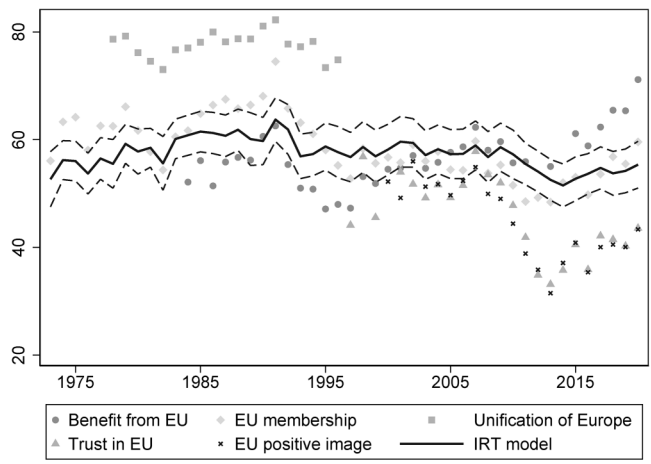


Figure 5. Comparison between the IRT model and other indicators of EU support. IRT: item response theory and EU: European Union.

series along with the share of pro-European responses given to five EB indicators frequently employed in the EU studies literature (e.g. Franklin and Wlezien, 1997; Hobolt and De Vries, 2016; Toshkov, 2011).

As expected, the response patterns of the various items move together with trends of the Bayesian measure. The 'membership' and 'EU benefit' indicators track the IRT trends closely. However, the former shows higher values of EU support up to the 1990s, whereas the latter lower ones between 1985 and 2005. Additionally, both indicators have an higher idiosyncratic variation than the estimated latent construct, and are also more prone to marked swings (Stimson, 2018). Even more problematic would be the use of the 'unification', 'EU trust' or 'EU image' questions as comprehensive indicators of EU support. The first problem with these indicators is their availability. The 'unification' question has been asked only up to 1996, and the 'EU trust' and 'EU image' ones only since 1997 and 2000, respectively. Furthermore, the 'unification' question also shows very high levels of expressed support, whereas the two items tapping into more affective aspects of EU support illustrate diverging trends in the post-crisis period.

Therefore, Figure 5 indicates that there is an overall 'convergence' between the estimated latent construct and some established single-question indicators of EU support, in the sense that the former seems to strike a balance between these partial measures, whilst also sharing the overall trend that is visible in all indicators. However, as far as one is interested in a comprehensive measure of EU support, the IRT construct clearly needs to be preferred over the single-question indicators, as the latter focus just on specific aspects of support and can display levels (e.g. the 'unification' indicator) or time trends (the 'trust' indicator) of support which cannot be taken as indicative of the overall of citizens' attitudes towards the EU.

Although these indicators also feature in the estimation of the Bayesian model, they are not linearly related to the estimated latent attribute, as the extent to which a shift in any raw indicator is reflective of a shift in the underlying mood is captured by each indicator's discrimination. For instance, the 'membership' question has an average discrimination (see the Online appendix), whereas the 'benefit', 'unification' and 'EU image' questions have relatively low ones, meaning that their response patterns are not very informative of the underlying latent construct of interest. Additionally, question parameters are also employed to rescale the level of EU support on an interpretable metric and, therefore, not just the trends but also the levels of the Bayesian measure should reflect more the true contribution of each item.

To further explore whether the IRT estimates represent a valid summary measure of EU support, I compare their fit with real data. This means to plug into the model the estimated support for each semester and the question parameters to predict the proportion of pro-European responses to each question. To avoid overfitting, the measures of fit are calculated using 5-fold cross-validation.

A first measure of fit is the root mean square error (RMSE), calculated as the square root of the average squared difference between predicted values and observed ones (McGann, 2014: 124). For the Bayesian model, the RMSE equals 4.46. Using item means to guess the responses of all the questions produces an RMSE of 4.88. From the RMSE, it is possible to calculate a by-item R^2, which represents the proportion of

the variance that the model explains, over and above what is explained by the item means (*ibidem*). The IRT model has a by-item R^2 of 0.161. This fit can be compared with that of other measures of EU support like one estimated using the DR algorithm and the EB 'membership' question. As Table 1 shows, the latter explains just 7.8% of the variance in response rates, whereas the DR algorithm does, surprisingly, even worse with an explained variance ranging from 0.7% to 5.3%. The IRT model, therefore, explains more than the double of the best alternative measure.[4]

Construct validity

Construct validation aims at showing that a measure is in line with established hypotheses about the relationship between the concept being measured and other concepts (Adcock and Collier, 2001). Moreover, with this exercise it is also possible to assess whether this approach would yield interestingly different results when used instead of other indicators in subsequent analyses.

I test whether the IRT estimates of EU support are related to the vote share of Eurosceptic parties in 259 elections since 1999. If the Bayesian measure is a valid operationalisation of EU support, it should be negatively related to Eurosceptic vote share. To identify Eurosceptic parties, I use Chapel Hill Expert Survey data (Bakker et al., 2020). The survey collects party positions on EU integration and was first conducted in 1999, with additional waves in 2002, 2006, 2010, 2014, 2017 and 2019. A party is classified as Eurosceptic if the expert score is equal or lower than 2 on a scale ranging from 1 ('strongly opposed' to EU integration) to 7 ('strongly in favour'). Party positions between waves are estimated as the weighted average of all positions for that party from all available waves, where the weights are the inverse distances between the missing year and the available waves (Broniecki, 2018). Data on Eurosceptic vote share comes from the ParlGov database (Döring and Manow, 2020). Tobit models are used to account for the fact that the dependent variable is bounded between 0 and 100.

The models include political and economic controls commonly used in the analysis of support for Eurosceptic parties. First, country-level confounders are accounted for with country fixed effects. Secondly, as economic factors might be both an argument to cue voters against the EU (De Vries and Edwards, 2009) and a dimension underlying, on a more durable basis, EU support (Boomgaarden et al., 2011), annual unemployment

Table 1. Model fit using the item response theory (IRT) measure and alternative measures (5-fold cross-validation).

	RMSE	By-item R^2
Bayesian IRT model	4.46	0.161
'Membership' question	4.62	0.078
Dyad Ratios algorithm (%pos/%pos+%neg)	4.75	0.053
Dyad Ratios algorithm (%pos)	4.86	0.007
Item means	4.88	0 (by construction)

and GDP growth rates are included to control for the possibility of a spurious relationship. Similarly, both support for Eurosceptic parties and EU attitudes have been associated with growing concerns for national identity (Hooghe and Marks, 2009) and discontent with increasing immigration from other countries (Treib, 2014). The share of individuals with exclusive national identity, and the percentage of foreigners in the labour force and the population as a whole are used to measure the strength of identitarian concerns as well as the perceived threat to national identity.[5] Also, I consider the net budgetary position of the country, as the financial redistribution between member states might affect both EU support and the popularity of nationalistic platforms (Borin et al., 2021). Finally, I control for the Eurosceptic vote in the previous election, for the fact that Euroscepticism has increased over time because of the growing politicisation of the EU, and for the effect of less salient EP elections.

Figure 6 plots the marginal effects of the various predictors, whereas the models are presented in the Online appendix. Recalling that the primary purpose of this analysis is to assess the construct validity of the Bayesian measure, the results of model using the IRT-estimate construct are particularly encouraging. The previous Eurosceptic vote ($p < 0.05$) and the time trend variable ($p < 0.01$) are significantly associated, in the expected direction, with Eurosceptic vote share. Although these controls have a strong association with Eurosceptic vote share, the coefficient of the public preference variable remains correctly signed and well below the conventional 5% threshold. Therefore, the analysis provides further evidence of the validity of the Bayesian estimation. Furthermore, the use of alternative measurement approaches (e.g. the DR algorithm or the 'membership' indicator) leads to marginally worse model fit. Additionally, the DR

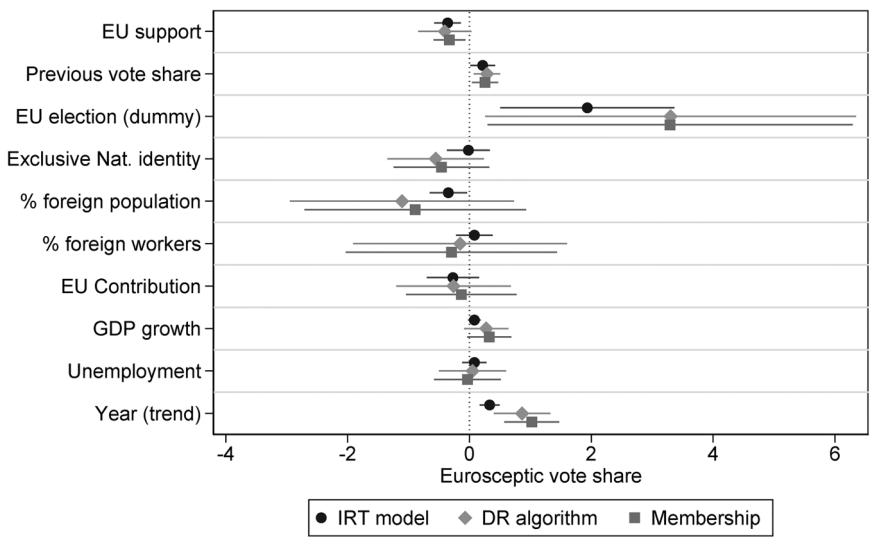


Figure 6. Marginal effects of the predictors on Eurosceptic vote share in three models.

algorithm estimates fail to reach statistical significance at 95%, whereas in the model using the membership question the relationship between EU support and Eurosceptic parties' electoral performances is substantially weaker and other factors play a bigger role in explaining Eurosceptic vote share.

Conclusion and discussion

This study has shown that Bayesian IRT models allow for the production of valid estimates of public EU support, comparable across member states and over time. I argued that Bayesian IRT models represent an improvement over existing approaches to the measurement of EU support, like single-question indicators and the DR algorithm, and that, therefore, such models can also be used to estimate measures of public attitudes where no appropriate single-question indicators exist.

Using the responses to 1448 EB questions, I generate estimates of public EU support in each member state for every year between 1973 and 2020. These measures have been validated by inspecting their correlation with existing alternative approaches and by looking at their association with the vote share of Eurosceptic parties in 259 elections conducted over the last 20 years. Additionally, an IRT-based approach produces a more precise estimate of the underlying level of public EU support. Bayesian IRT models proved to be superior to other alternative measurement techniques both in terms of fit with expressed preferences and with regard to the grounding in an individual-level model of behaviour.

Although the use of single-question indicators is easier and more efficient to implement, it has different limitations. The 'membership' question, for instance, accounts for just about half of the variance in expressed preferences explained by the Bayesian estimates, confirming that single-questions indicators represent only partial measures of a latent phenomenon like EU support (Anderson and Hecht, 2018). Hence, a preference for more efficient measurement strategy comes at the price of less precise estimates of public support. This consideration adds up to other concerns issues to the use of single-question indicators, like the high level of idiosyncratic variations, the risk of interruptions in the data series and the need to retrofit existing measures to the needs of the current analysis.

On the contrary, Bayesian IRT models can easily accommodate the interruptions in a specific data series, as far as there are other indicators tapping into the presence of the underlying attribute of interest. This can also be partially achieved with the DR algorithm. However, given the grounding in an individual-level model of response, the capacity to deal with neutral answer options and the possibility of ensuring the cross-national comparability of the items, Bayesian IRT models provide, in fact, more robust estimates of the latent public EU support. Indeed, by using the latter, we can reconstruct the observed response patterns with a smaller average error and accounting for a larger proportion of the variance of each question. Additionally, the DR algorithm and Bayesian IRT models require exactly the same kind of input information and both can be implemented with accessible free software. Hence, there are no efficiency-precision trade-off as in the case of single-question indicators.

To conclude, this methodological novelty comes with both theoretical and empirical gains. On the one hand, Bayesian IRT models offer a more precise and theoretically grounded alternative to the use of the DR algorithm for the estimation of latent opinion and allow to operationalise EU support in a way that is closer to its established conceptualisations. On the other hand, Bayesian IRT frees the researcher from an over reliance on a narrow set of existing single-question indicators which may or may not fit well with the analysis at hand, and may sometimes not even be available for a specific task. In this respect, they have a clear potential in helping researchers to address the lack of policy-specific measures of EU support starting from aggregate-level data.

Acknowledgements

For valuable comments on earlier drafts of this article I am very grateful to four anonymous referees, Gerald Schneider, Edoardo Bressanelli, Christel Koop, Stuart Turnbull-Dugarte, and the participants of the panel 'Exit, Voice, and Dissatisfaction with Democracy in Europe' at the Annual Meeting of the American Political Science Association (9–13 September 2020). All remaining errors remain my own.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Michele Scotto di Vettimo https://orcid.org/0000-0002-7627-3975

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Bayesian IRT differs from frequentist IRT in that the latter uses on maximum likelihood estimation, whereas the former relies on the sampling of the posterior distributions of the parameters of interest. Though the two frameworks reach relatively similar conclusions when applied to the same situations, the Bayesian framework provides more precise point and uncertainty estimates of the parameters in more complex models (Burkner, 2020: 15).
2. EB included also Bulgaria, Romania and Croatia in most surveys from 2004. Therefore, these three countries have estimates pre-dating their accession in 2007 and 2013.
3. The Visegrad group is a political alliance to advance cooperation in military, cultural, economic, and EU matters, and consists of Czech Republic, Hungary, Poland and Slovakia. These

countries have recently voiced more hostile attitudes towards more burden-sharing in the framework of the EU immigration policy.
4. As robustness check, for the DR the input is arranged in two different ways. See the Online appendix.
5. For the first, see the 'EUid1' question in the Online appendix. The other two come from the ILO and OECD data, respectively.

References

Adcock R and Collier D. (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3): 529–546.

Anderson CJ and Hecht JD (2018) The preference for Europe: Public opinion about European integration since 1952. *European Union Politics* 19(4): 617–638.

Ariely G and Davidov E (2011) Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Value Survey. *Social Indicators Research* 104(2): 271–286.

Avvisati F, Le Donné N and Paccagnella M (2019) A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Meas Instrum Soc Sci* 1(8): 1–10.

Bakker R, Hooghe L, Jolly S et al. (2020) *Chapel Hill Expert Survey. Version 2019.1*. Available on chesdata.eu. Chapel Hill: University of North Carolina.

Beck N (1989) Estimating dynamic models using kalman filtering. *Political Analysis* 1: 121–156.

Bølstad J (2015) Dynamics of European integration: Public opinion in the core and periphery. *European Union Politics* 16(1): 23–44.

Boomgaarden HG, Schuck AR, Elenbaas M et al. (2011) Mapping EU attitudes: Conceptual and empirical dimensions of Euroscepticism and EU support. *European Union Politics* 12(2): 241–266.

Borin A, Macchi E and Mancini M (2021) EU transfers and euroscepticism: can't buy me love?. *Economic Policy* 36(106): 237–286.

Broniecki P (2018) Informal bargaining in bicameral systems: Explaining delegation by the Council of the European Union and the European Parliament. *Doctoral dissertation, University College London, London*.

Bürkner P-C (2020) Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *Journal of Intelligence* 8(1): 1–18.

Caughey D and Warshaw C (2015) Dynamic estimation of latent opinion using a hierarchical group-level IRT model. *Political Analysis* 23(2): 197–211.

Claassen C (2019) Estimating smooth country–year panels of public opinion. *Political Analysis* 27(1): 1–20.

De Vries CE (2013) Ambivalent Europeans? Public support for European integration in east and west. *Government and Opposition* 48(3): 434–461.

De Vries CE (2018) *Euroscepticism and the Future of European Integration*. Oxford: Oxford University Press.

De Vries CE and Edwards EE (2009) Taking Europe to its extremes: Extremist parties and public Euroscepticism. *Party Politics* 15(1): 5–28.

De Vries C and Steenbergen M (2013) Variable opinions: The predictability of support for unification in European mass publics. *Journal of Political Marketing* 12(1): 121–141.

Döring H and Manow P (2020) ParlGov 2020 Release, *Harvard Dataverse*. DOI: 10.7910/DVN/Q6CVHX

Eichenberg RC and Dalton RJ (2007) Post-Maastricht blues: The transformation of citizen support for European integration, 1973–2004. *Acta Politica* 42(2-3): 128–152.

Esping-Andersen G (1990) *The Three Worlds of Welfare Capitalism*. Princeton: Princeton University Press.

Franklin MN and Wlezien C (1997) The responsive public: Issue salience, policy change, and preferences for European unification. *Journal of Theoretical Politics* 9(3): 347–363.

Guinaudeau I and Schnatterer T (2019) Measuring public support for European integration across time and countries: The 'European mood' indicator. *British Journal of Political Science* 49(3): 1187–1197.

Heath A, Fisher S and Smith S (2005) The globalization of public opinion research. *Annual Review of Political Science* 8: 297–333.

Hobolt SB and De Vries CE (2016) Public support for European integration. *Annual Review of Political Science* 19: 413–432.

Hooghe L and Marks G (2009) A postfunctionalist theory of European integration: From permissive consensus to constraining dissensus. *British Journal of Political Science* 39(1): 1–23.

Hutter S and Grande E (2014) Politicizing Europe in the national electoral arena: A comparative analysis of five west European countries, 1970–2010. *Journal of Common Market Studies* 52(5): 1002–1018.

Kruschke J (2014) *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Cambridge: Academic Press.

McGann A, Dellepiane-Avellaneda S and Bartle J (2019) Parallel lines? Policy mood in a plurinational democracy. *Electoral Studies* 58: 48–57.

McGann AJ (2014) Estimating the political center from aggregate data: An item response theory alternative to the Stimson dyad ratios algorithm. *Political Analysis* 22(1): 115–129.

Nunnally JC and Bernstein IH (1994) *Psychometric Theory*. New York: McGraw-Hill.

Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *in* 'Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)'.

Sánchez-Cuenca I (2000) The political basis of support for European integration. *European Union Politics* 1(2): 147–171.

Solt F (2020) *Modeling dynamic comparative public opinion*, SocArXiv, 10.31235/osf.io/d5n9p.

Stimson JA (1991) *Public Opinion in America: Moods, Cycles, and Swings*. Boulder: Westview Press.

Stimson JA (2018) The dyad ratios algorithm for estimating latent public opinion: Estimation, testing, and comparison to other approaches. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 137(1): 201–218.

Stoeckel F (2013) Ambivalent or indifferent? Reconsidering the structure of EU public opinion. *European Union Politics* 14(1): 23–45.

Toshkov D (2011) Public opinion and policy output in the European Union: A lost relationship. *European Union Politics* 12(2): 169–191.

Treib O (2014) The voter says no, but nobody listens: Causes and consequences of the Eurosceptic vote in the 2014 European elections. *Journal of European Public Policy* 21(10): 1541–1554.

Voeten E and Brewer PR (2006) Public opinion, the war in Iraq, and presidential accountability. *Journal of Conflict Resolution* 50(6): 809–830.

Wratil C (2019) Territorial representation and the opinion–policy linkage: Evidence from the European Union. *American Journal of Political Science* 63(1): 197–211.

Zaidi A, Harper S, Howse K et al. (2018) *Building Evidence for Active Ageing Policies: Active ageing index and Its Potential*. Singapore: Palgrave Macmillan.

Zhelyazkova A, Bølstad J and Meijers MJ (2019) Understanding responsiveness in European Union politics: Introducing the debate. *Journal of European Public Policy* 26(11): 1715–1723.