Measuring public support for European integration from population-level data using a Bayesian IRT model

Michele Scotto di Vettimo *
King's College London

Working paper
presented at the 116th American Political Science Association's Annual Meeting
10 September 2020

Abstract

This study proposes the use of Bayesian item response theory (IRT) models to measure public preferences towards Europe. IRT models address the limitations of single-question indicators and produce valid estimates of public latent attitudes over long time periods, even when available indicators change over time or present interruptions. The approach is compared with an alternative technique recently introduced in the study of EU public opinion, the Dyad Ratios algorithm. It shows that IRT models can both incorporate a more theoretically grounded individual-level model of response and produce more accurate estimates of latent public support for Europe. The measure is validated by showing its association both to alternative measures of EU support and to the vote share of Eurosceptic parties across Europe.

*Department of Political Economy, King's College London, michele.scotto_di_vettimo@kcl.ac.uk.

1 Introduction

Theories of European integration, like liberal intergovernmentalism and neo-functionalism, have generally neglected the role of public preferences in the integration process. However, since the 1990s and the end of the "permissive consensus" (Hooghe and Marks, 2009), there has been a growing interest in understanding the impact of public opinion on the European Union's (EU) developments. Against a background of growing contestation and politicisation of the EU, public attitudes towards European integration are now central to understanding European-level policy-making (Hagemann et al., 2017; Wratil, 2019) as well as national-level party strategies (Hutter and Grande, 2014; Hoeglinger, 2016; Rauh et al., 2020). Nevertheless, the precise measurement of public attitudes towards the EU has received relatively limited attention and a wide range of measures has been employed to analyse public support for Europe without clarifying "what we actually mean when we refer to and measure support for European integration" (Hobolt and De Vries, 2016, 415).

In the last decade, multidimensional approaches to the study of public support for the EU have gained prominence (Lubbers, 2008; Hobolt and Brouard, 2011; Hobolt and De Vries, 2016; Boomgaarden et al., 2011; De Vries, 2018). These approaches acknowledge that support for the EU can be directed towards different objects - the polity as a whole or its policies - and can be of various nature - with affective, utilitarian or other connotations. Although there is little agreement on which dimensions are crucial for understanding public attitudes towards the EU, all these contributions "highlight that a one-dimensional approach is insufficient" (De Vries, 2018, 42). This, in turn, calls for new strategies to measure public attitudes towards the EU, as scholars have mainly relied on existing single-question indicators, which can only account for one dimension of EU support at a time. These indicators, therefore, offer only a partial approximation of the concept of interest (Hobolt and De Vries, 2016; Guinaudeau and Schnatterer, 2019) and most of the times have been retrofitted to suit a particular analysis (Anderson and Hecht, 2018, 621). Additionally, they rarely allow for the measurement of policy-specific preferences for EU integration over relatively long periods (Zhelyazkova et al., 2019; Guinaudeau and Schnatterer, 2019).

The shortcomings of single-question indicators have recently been addressed by the use of dimension reduction techniques like the Stimson's Dyad Ratios (DR) algorithm (Stimson, 1991, 2018). The latter allows for the assessment of the dimensionality of EU support and the extraction of latent public attitudes towards European integration from a set of different indicators (Anderson and Hecht, 2018; Guinaudeau and Schnatterer, 2019). Over the last three decades, the DR algo-

rithm has established itself as a reliable and elegant technique to extract the population's latent "policy mood" in many domestic contexts.[1] Its underlying logic is very similar to that of principal component analysis, and it has been designed to circumvent the problems posed by interrupted public opinion series. In the EU context, it represents the most advanced approach that scholars have so far employed to estimate EU support starting from population-level data. Yet, although this technique represents a solution for reducing the reliance on single indicators, it has also been criticised for its lack of a grounding in coherent individual-level micro-foundations (McGann, 2014; Caughey and Warshaw, 2015).

Therefore, I propose the use of Bayesian item response theory (IRT) models with population-level data for the estimation of public support for Europe. I argue that IRT models have different advantages over the DR algorithm for the combination of single-question indicators when measuring public attitudes towards Europe, and they would also allow a more precise examination of support for European policy integration in specific areas, a domain currently characterised by lack of appropriate data (Zhelyazkova et al., 2019). IRT has recently been adapted to the estimation of aggregate-level public opinion in different national contexts (McGann, 2014; Caughey and Warshaw, 2015; Caughey et al., 2019). These models start from an explicit model of individual behaviour, they can deal with neutral responses, and have already been tested in situations where available public opinion data presents many interruptions preventing over time comparison. With the appropriate adjustments, these models produce measures of public preferences that are comparable both over time and between countries (McGann et al., 2019). The EU context represents a promising area of extension. Existing surveys, like the Eurobarometer, provide cross-nationally comparable opinion data for relatively long, yet irregular, time periods. IRT models allow the extraction of a measure of latent public attitudes towards Europe from these irregularly asked survey items, with virtually the same type of information required by the DR algorithm, and produce estimates of support with a sounder theoretical grounding.

In the following, I first give an overview of the existing approaches to measuring public preferences towards the EU, with their limitations. Next, I introduce the Bayesian IRT model along with the survey data used for the its estimation. Then, I present the estimates of public attitudes towards Europe from 1973 to 2019, assess their validity and compare the fit of the IRT model with those of the DR algorithm and possible alternative measures of EU support. I conclude by showing

[1] For the US context see, among others, Stimson (1991); Erikson et al. (2002). For a review of the use of the DR algorithm in other contexts see Stimson (2018, 202).

that IRT models produce valid estimates of public support for the EU, which are more precise than those obtained with alternative methodologies.

2 Existing approaches to capturing preferences for Europe

The end of the "permissive consensus" (Hooghe and Marks, 2009) and the increasing politicisation of the EU have led scholars to incorporate the role of the public into theories of supranational integration and to study the nature of support to the EU (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). In fact, the measurement of public attitudes towards the EU has typically faced a particular challenge: comparability across space and time. With regard to the former, we need surveys that (1) employ the same question wording across countries, and (2) are administered at more or less the same time in all EU member states or candidate countries. Cross-national surveys represent, therefore, the most obvious candidates as data sources. Among these, European Commission's Eurobarometer (EB) surveys have become by far the most often used ones. EB surveys have been conducted at least every semester, uninterruptedly, since September 1973 in all EU member states and most candidate countries, and EB questions tap into various aspects of public attitudes towards Europe, from preferences for regulatory harmonisation or establishment of common policies in specific domains, to general evaluations on one's country membership to the Union or trust in EU institutions.

Yet, this abundance of questions is counteracted by the fact that only few of them are asked in a comparable fashion across many surveys. There are hardly any items that cover the full period for which the surveys have been running and, even if one considers shorter time frames, changing question wordings and interruptions in the administration of a specific question make the use of particular items impractical and force the analyst to choose the "best" single-question indicator among the few available ones. This means that scholars relying on Eurobarometer or similar surveys "often [face] a trade-off between analyses that rely on a small set of items to measure support over time and a cross-sectional analysis of various dimensions of support captured at a single point in time" (Hobolt and De Vries, 2016, 416-417). Although variables are numerous, it is rarely the case that all time points are available for any of them. This has led public opinion studies to ultimately rely on the few single-question indicators available during longer periods, like questions on the evaluations of one's country membership to the EU (e.g., Eichenberg and Dalton 1993, 2007;

Toshkov 2011), on the attitudes towards the unification of Europe (e.g., Franklin and Wlezien 1997) or the "preferred speed" of integration (e.g., Hobolt and De Vries 2016). Such indicators, therefore, are often retrofitted to suit the needs of a particular analysis. Additionally, they tap into just one dimension of EU support, and are not ideal for properly capturing the possible multidimensionality of citizens' attitudes towards Europe (Anderson and Hecht, 2018, 618).

This multidimensionality has to do both with the *object* of support (the EU as a system of governance or EU policies) and with its *connotation* (utilitarian, affective or other). On the one hand, multidimensionality can be understood in terms of the object of EU support. De Vries (2018) two-dimensional approach, for instance, distinguishes between procedural and substantive elements of representative democratic systems (Dahl, 1998) and argues that citizens' support for the EU can be decomposed into regime and policy evaluations. The former relate to the assessment of the process and institutional settlement in place for the articulation of people's interest and its legitimacy. Policy evaluations, instead, refer to people's preferences about the content of EU policy decisions and actions, as well as the attribution of policy responsibilities to the supranational level (Hobolt and De Vries, 2016; De Vries, 2018). The appropriateness of this conceptualisation has found support in Boomgaarden et al. (2011) and, similarly, Guinaudeau and Schnatterer 2019 show that "in some countries a generally favorable climate towards the EU is not incompatible with a desire to keep certain competences on the national level", suggesting that a second dimension of policy specific support is present in some countries or for some domains (2019, 1192).

On the other hand, there is also little agreement on the different connotations of public support or opposition to the European Union (Hobolt and De Vries, 2016; De Vries, 2018). Cost-benefit and utilitarian dimensions (Gabel, 1998; McLaren, 2005), identitarian values (Hooghe and Marks, 2004; McLaren, 2005; Lubbers, 2008), social and policy-specific considerations (Hobolt and Brouard, 2011) are all elements that have been showed to characterise EU support, although their relative importance might differ from country to country. Other studies showed that support for EU integration entails five "genuinely distinct and independent" dimensions of performance, identity, affection, strengthening and utilitarianism (Boomgaarden et al., 2011, 259).

This multi-dimensional understanding of EU support has practical implications on how we decide to measure it. As Hobolt and De Vries (2016) argue, some common single-question indicators are more appropriate to tap on regime support, while others on policy support. Therefore, with the use of single-question indicators we can, at best, capture either one or the other dimension of EU support, but not both at the same time. To the extent that such indicators present different

trends (Hobolt and De Vries, 2016, 417-418), the measurement of public attitudes towards the EU should rely on a combination of different indicators so that both dimensions are properly accounted for. Moreover, single-question indicators possesses also unappealing empirical features. If compared to composite indexes, these indicators are subject to higher idiosyncratic variability due to measurement errors (Stimson, 2018), and conclusions drawn from unique series are more likely to be subject to different forms of bias due to a particular framing, ordering or wording of the question (Guinaudeau and Schnatterer, 2019, 1189).

To address the shortcomings of single-question indicators, political scientists have started to employ measures combining different survey items related to the latent concept of interest (i.e., EU support). Each item on its own might not offer a comprehensive indicator of the concept and, moreover, it is sometimes just not available for all time points the analyst wishes to consider. Yet, the combination of these items allows for the creation of a composite measure capturing a multi-faceted concept lacking a single empirical counterpart (Anderson and Hecht, 2018). In particular, the Dyad Ratios algorithm, introduced by Stimson in the US context in the early 1990s (Stimson, 1991, 2018), has recently been applied to the estimation of cross-national series of public preferences towards European integration (Anderson and Hecht, 2018; Guinaudeau and Schnatterer, 2019).

The DR algorithm was developed primarily to deal with the practical problem of data series interruption (Stimson, 1991). It builds on the assumption that the various available indicators (survey questions) "are influenced by latent dispositions of respondents so that some portion of their variation reflects something more general" than attitudes towards the particular item asked (Stimson, 2018, 203). The DR algorithm estimates this latent construct, calculates the amount of the indicators variance explained by the estimated dimension and tests whether a possible second latent dimension can be identified. In practice, the algorithm imitates principal component analysis, a statistical technique often used to estimate latent structures. However, principal components estimation requires as a starting point a covariance matrix of all the survey questions for each time point. Yet, in a situation where questions are asked over different time frames, this covariance matrix cannot be estimated in the first place (Stimson, 2018, 203), and the DR algorithm offers a solid and ingenious procedure to get around this problem posed by interrupted survey data.

To apply this technique to the estimation of attitudes towards Europe, we have to select a set of questions capturing favourable or critical attitudes towards the EU or European integration (Guinaudeau and Schnatterer, 2019, 1190), record the percentage of citizens answering in a favourable

way at each administration and pass this information on to the algorithm.[2] The algorithm will then calculate the rate of change of these percentages between two consecutive surveys. These ratios represent a first estimate of the relative change of the underlying latent variable over time and are used in an iterative estimation process where, by setting an arbitrary value (say 100) both at the final and at the initial points of the series, ratios are used recursively to estimate the scores of the other time points. After this first estimation, the iterative process weighs the contribution of each question by its "validity" (the squared correlation between the item and the estimated latent construct), and estimates new scores for each time point until a convergence criterion is met (Stimson, 2018). At the end of the process, means and standard deviations of each question are averaged, using items' validity as weight, and the results are imposed on the estimated construct to rescale it to a meaningful metric.

The DR approach has recenlty been implemented in the EU context and, at the moment, it represents the most advanced technique used in the estimation of EU support starting from population-level data. Anderson and Hecht (2018) used the DR algorithm to measure attitudes towards Europe in France, Germany, Italy and the United Kingdom from 1952 to 2017. They selected 14 frequently asked items, about various aspects of the integration process and EU support, tabled between 1952 and 2017 (2018, 624).[3] The estimated series show that support for Europe is captured by a broad set of indicators and that a single underlying dimension can explain most of the variance of citizens' attitudes towards Europe, although a second dimension emerges in all countries from the 1990s. Consequently, the authors conclude that single-question indicators should be "treated as incomplete measures of a larger underlying syndrome" (2018, 633).

Subsequently, Guinaudeau and Schnatterer (2019) extended the use of the algorithm to 27 member states from 1973 to 2014. They use 71 distinct Eurobarometer questions to create a bi-annual and country-specific measure of "European mood". Again, although "the first dimension estimated always captures a considerable share of variance", in some cases a generally favorable climate towards the EU is not incompatible with skeptics attitudes towards further competence delegation (2019, 1191-1192). This effectively indicates that there are countries where a second

[2] In fact, it can be also the ratio of positive responses over non-neutral ones or the net difference between positive and negative responses. The key requirement is that it must be a single number.

[3] The questions ask about affective attitudes towards the Community, trust in its institutions, evaluation of membership and preferred speed of integration.

dimension does emerge, and in which support for policy integration is not linked to the general support for Europe.

Both studies demonstrate the feasibility of a new approach to the examination of the "relationship between public opinion and the process of integration over the long historical run and across member states - a question that previously has been investigated with shorter time series and single item support indicators" (Anderson and Hecht, 2018, 633). The DR algorithm, therefore, has allowed to overcome the problem of the interruptions in the collection of other indicators (no item is available in all the years covered by these studies) and to assess to what extent any indicator on its own represents an accurate measure of EU support.

Nonetheless, the Dyad Ratios algorithm also possesses some features making it unappealing under both a theoretical and an empirical point of view, which are related to (1) the lack of grounding in an individual-level model of response, (2) the treatment of ordinal or neutral responses, and (3) the methodology used to impose a meaningful metric on the estimated series.

First, the algorithm assumes that the ratio of the percentage of positive responses to the same question asked in different years provides a measure of the relative mood of the two years. This linearity assumption means that, for any question, both a change in the percentage of favourable answers from 10% to 20% and one from 30% to 60% generate the similar expectation that mood between the two years has changed by a ratio of 1:2 (McGann, 2014, 117). However, this assumption does not find support in an individual-level model of response and, in fact, psychometric theory suggests that the relationship between a latent trait and an indicator measure is better represented by a curvilinear relationship (Nunnally and Bernstein, 1994).

Secondly, as the input information for the algorithm has to be a single value for each question-time dyad, there is no way for explicitly modelling ordinal or neutral responses. Either the information coming from neutral responses is dropped altogether (see the discussion in section 5.1) or it is incorporated with that of positive or negative ones. Same applies for all types of graded responses, which need to be aggregated into a single number. Since the decision on how to handle these responses is completely up to the analyst's discretion, the comparability of the estimated series is potentially hampered. Furthermore, when the share of neutral responses is relatively high, the decision to discard it leads to the loss of potentially relevant information on citizens' attitudes.

Finally, the algorithm produces a standardised measure of mood whose metric is not directly interpretable. The final series, therefore, needs to be rescaled so that it reflects the metric of the items used as raw data. This metric restoration is performed by averaging all questions' means and

standard deviations, weighted by items' validity, to produce a mean and a standard deviation that are imposed on the estimated construct. However, this procedure ignores number of occurrences of each question and, all things being equal, very rare items contribute to the determination of the new metric to the same extent of more frequent ones. As the new metric shifts the whole series as it is, this can affect the plausibility of the whole measure and hamper the comparability between of the mood of two countries with different longitudinal extension of the raw data.

For these reasons, more robust approaches have been proposed to aggregate sparse survey data and produce estimates of latent public preferences (Voeten and Brewer, 2006; McGann, 2014; Caughey and Warshaw, 2015). In the following, I build on McGann et al. (2019) specification to extend the use IRT models to the estimation of comparable country-specific measures of public support for Europe. I argue that (1) these models offer a valid approach to address two issues faced by studies of EU public opinion - namely the fact that single-question indicators represent only incomplete and irregular measures of EU support and the lack of policy-specific measures of support. Also, (2) IRT models are superior to the DR algorithm in fulfulling this task because the former reflect an individual-level model of response, allow for the explicit treatment of neutral responses and, use question-specific features to recover the level and dispersion of the estimated latent dimension.

3 An Item Response Theory model of support for Europe

IRT models were developed in psychometric theory to measure latent attributes of individual respondents, like ability or ideology. These individual-level models have been applied in political science to the estimation of legislators' ideal points (e.g., Clinton et al. 2004) or mass-opinion data (Levendusky and Pope, 2010). The IRT approach has recently been adapted to deal with population-level data and can be used to estimate latent public preferences from the percentages of responses to certain survey items (McGann, 2014; McGann et al., 2019), which is precisely the kind of information required by the DR algorithm as well.

However, contrary to the DR algorithm, 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 few questions related to different aspects of EU support. The probability that she answers in a pro-European way ("correct" response) is modelled as a normal cumulative distrubution function and is determined

by three factors. One is individual i's attribute θ_i, our ultimate object of interest, although at the population-level. All things being equal, the higher the individual's attribute the more likely she is to answer correctly to the question asked. Furthermore, each question q has its intrinsic difficulty b_q. So, we would expect that it is more "difficult" to answer in a positive way to a question asking whether the EU should become a federal entity rather than to a question asking whether more decisions should be taken at the EU-level on matters of research and development. Whenever respondent's attribute exceeds the difficulty of the question, she is more likely to answer correctly than incorrectly. Finally, changes in the attribute possessed by the respondent affect the likelihood of answering correctly, but the effect of these changes is not constant across items. Item discrimination, a_q, determines how rapidly the probability of a correct response changes at different levels of the attribute θ_i.

Bayesian IRT models with aggregate data incorporate this individual-level model into a population-level one. By knowing each question's difficulty and discrimination, as well as the average latent EU support in the population and its dispersion, it would be possible to predict the percentage of pro-European responses to question q at time t. In fact, we find ourselves precisely in the opposite situation. The number of pro-EU responses is a known quantity, whereas the aim is to estimate the latent EU support in the population. A Bayesian approach, therefore, is used to determine the most likely value of the unknown quantities (question difficulty and discrimination and latent EU support), given the observed data (Kruschke, 2014, 98). These quantities are called model "parameters". Bayesian IRT starts from some prior information about the probability distribution of the different parameters, then looks at the observed data (actual responses) and estimates the probability (the likelihood) that the data could be generated by a model with a given set of parameter values.

The formalisation of this model following McGann et al. (2019) is reported in section A of the Appendix. In practice, a Bayesian model "tries" all the possible values of the parameters of interest and selects the combination that is most likely to generate the data that is actually observed. 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 value of the population's latent support for Europe for that time point. In this way, IRT models can reduce the reliance on a particular item and overcome the problem posed by the interruptions in the series of the various questions used as input.

We can now contrast the properties of the IRT model with those of the DR algorithm, recalling the three disavantages highlighted with regard to the latter. First, by starting from an explicit individual-level model of response, IRT addresses the shortcomings caused by the assumption that the responses to a given indicator and the latent attribute of interest are linarly related. In the IRT model, this relationship is curvilinear: if latent support is already quite pro-Europe, we would expect a large movement towards even more positive attitudes to have little effect (virtually everyone is manifesting support already) on the proportion of observed pro-European responses. If, however, latent support is more ambivalent, a change in latent mood could have a much greater effect (McGann, 2014). Moreover, questions with different discrimination parameters react differently to changes in the latent mood, a nuance that cannot be modelled with the DR algorithm.

Secondly, the IRT model allows a sounder treatment of those neutral attitudes that represent a non-trivial share of the responses to some Eurobarometer questions and, in general, can be applied to the treatment of any item with ordinal response categories. Contrary to the DR algorithm, IRT models do not require the input to be a single number for each item. Thus, it is possible to pass on to the model both the data on the share of positive responses *and* the share of neutral and positive responses combined. In effect, two difficulty parameters will be estimated for this kind of items: there will be a b_{q2} measuring the amount of θ_i required to answer the question in a positive way and also a $b_{q1} < b_{q2}$ representing the difficulty of answering q at least in a neutral way (McGann et al., 2019, 51).

In fact, this property can represent an important advantage of IRT models over the DR algorithm. It has been suggested that, in the EU context, ignoring neutral responses is more problematic than it would be, for instance, in the US case, where the DR algorithm has been conceived. De Vries and Steenbergen (2013) show that 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 *viceversa*) attitudes, but in fact public opinion towards European integration is not so clear-cut. Rather, expressed EU support is sometimes characterised by a certain degree of indefference (Van Ingelgom, 2014) or ambivalence (De Vries and Steenbergen, 2013) which calls for a more careful empirical treatment.

Finally, IRT models address the issues raised with regard to the interpretation and comparability of the series estimated with the DR algorithm. IRT models use the information coming from questions' difficulty and discrimination parameters to estimate a measure of latent support with an interpretable metric. Questions with very high levels of positive response and a low difficulty

or discrimination should not pull the estimated support upwards, as these questions are recognised as those for which a high share of positive responses should not come as a suprise or it is not that reflective of the underlying mood. Also, by constraining question parameters to be the same across countries, it is possible to produce a measure of EU support which is comparable across EU member states (McGann et al., 2019). On the contrary, achieving cross-country comparability was not an aim that shaped the design of the DR algorithm in the first place.

Therefore, compared to the DR algorithm, Bayesian IRT models produce measures of the population latent support for Europe with a grounding in a model of individual-level behaviour and without the problematic assumption of a linear relationship between changes in expressed preferences and changes in the latent attribute. Furthermore, 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. Nonetheless, IRT models require precisely the same type of input information as the DR, namely population-level response shares and the number of respondents interviewed for each question used and, therefore, do not call for a more extensive data gathering process.

4 Estimates of public support for Europe

In this study, public support for Europe is defined following the two-dimensional conceptualisation proposed by De Vries (2018) and adopted by Guinaudeau and Schnatterer (2019) as well. In particular, I consider citizens' attitudes towards the EU as a political system (regime support) and citizen's preferences towards EU policy integration (policy support). Hence, to implement the Bayesian IRT model, I identified 121 Eurobarometer questions (list in Table A1 of the Appendix) tapping on either one or the other dimension of EU support, and asked at least four times between September 1973 and September 2019.[4] The first group of questions comprises items on trust in the EU or in its institutions, evaluation of one's country membership to the Union and dispositions towards the EU or the goal of European unification/integration. The second and bigger (77%) category of questions, instead, taps on 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, whether

[4] Other analysis have relied on a similar or even higher number of questions. Erikson et al. (2002), for instance, use 133 questions between 1952 and 1996 for the US "policy mood", whereas McGann (2014) employs 364 items for the study of UK public opinion from 1947 to 2005. The studies of Anderson and Hecht (2018) and Guinaudeau and Schnatterer (2019) use 14 and 71 items respectively.

the respondent favours or opposes EU-level harmonisation or the creation of an EU common policy in a given domain. Policy-related questions are included only if the answer options allow the respondent to express both favourable and hostile views on policy integration. Questions asking for evaluations of EU policy actions, therefore, are excluded as they ask respondents to evaluate past EU performances, but might not necessarily capture preferences towards EU integration *per se*.

The size of the dataset ranges from 591 question-semester dyads for Croatia to 1936 for the nine countries already member in 1973.[5] Data comes from 161 (Special or Standard) Eurobarometer surveys and if a question is asked more than once in the same semester, responses are averaged over the available surveys. For each observation, the percentage of responses showing 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). Percentages of refusal and "don't know" answers are excluded.

I implement a Bayesian IRT model using Gibbs sampling, a randomised algorithm commonly used to generate sequences of observations from approximations of the probability distributions of the desired parameters. The procedure is a type of random walk through the parameter space. It starts by giving an arbitrary value to one of the parameters of interest, and then the algorithm generates an instance of another parameter, conditional on the current values of all the other parameters and variables (Kruschke, 2014, 137). Progressively, this allows the algorithm to approximate the joint probability distribution of all the parameters of interest. Apart from the input data, therefore, the sampler requires parameter values from which the sampling process could start. These values are called priors. I assigned non-informative uniform priors to all parameters, reflecting the lack of information about their true values. The models have been estimated using the software JAGS (Plummer, 2003). In each EU member state, and for the EU as a whole, I run models with three independent chains, each with 30,000 iterations, 15,000 iteration burn-in and thinning parameter is set to 10. The JAGS code is provided in the replication material.

Estimated question parameters (discrimination and difficulty) are reported in Table A2 and A3 of the Appendix. As a way of example, we can see that questions related to "core state powers" (taxation, police, pensions, welfare and education) or to national identity proved to be

[5] 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. Yet, not all the selected question are asked in canidate countries (even when the latter are be covered in a survey) and, thus, Bulgaria, Romania and Croatia have less question-semester dyads if compared to member states joining the EU in 2004.

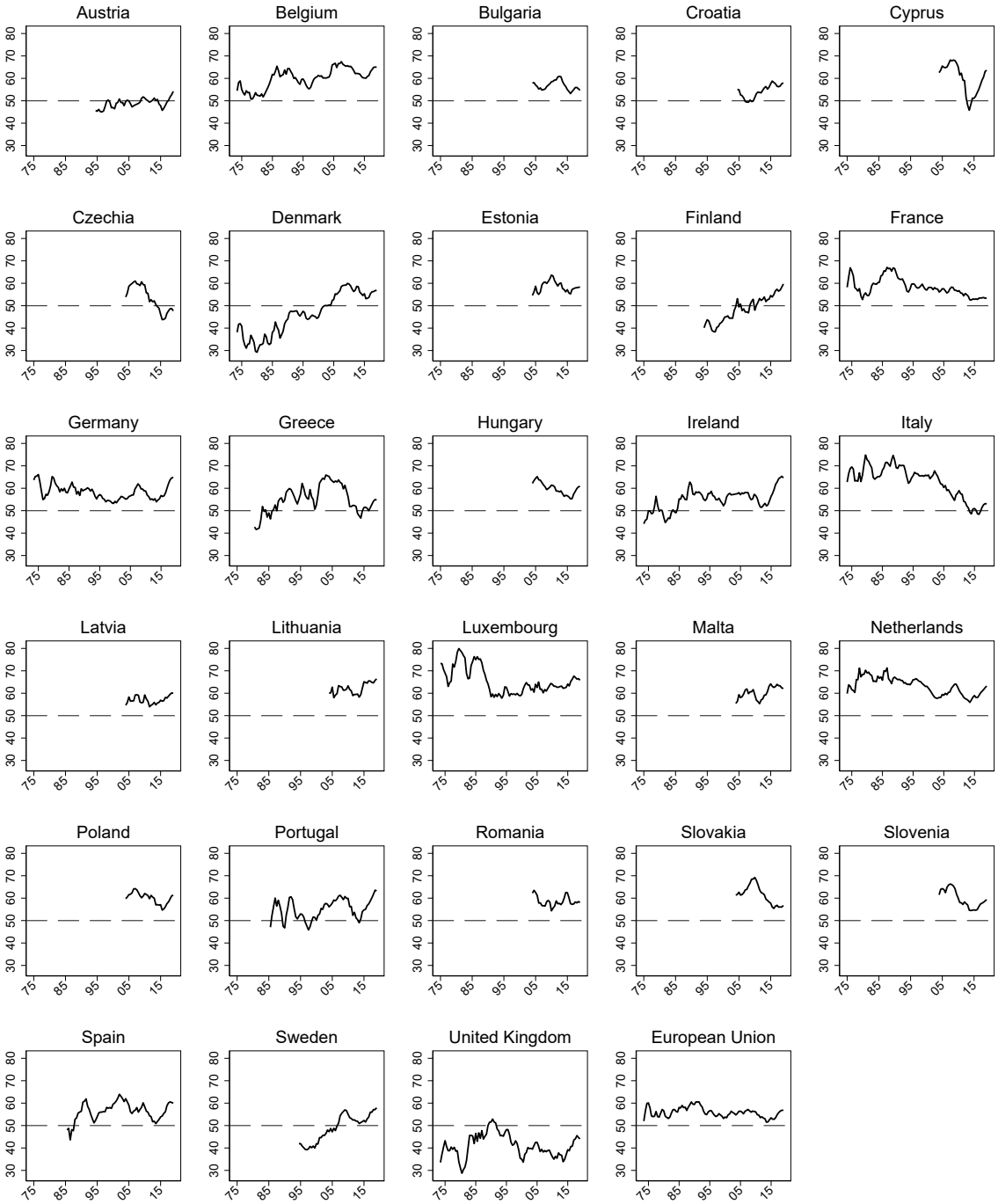

Figure 1: Public support for Europe estimated with a Bayesian IRT model.

more "difficult" than questions pertaining to established domains of EU action (free movement of people), general committments to the "unification of Western Europe" and to policy initiatives tackling transnational problems (terrorism, organised crime, development cooperation and human trafficking).

The estimation produces measures of public support for Europe for all EU member states in each semester from autumn 1973 (or from the semester when first EB data are available) to autumn 2019. The estimated series for all member states and the European Union are plotted in Figure 1, with the reference line representing the 50 percent threshold. Previous cross-national 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 (see also Eichenberg and Dalton 1993). This pattern emerges also in the first nine panels of Figure 1, with Denmark and especially the United Kingdom (but only to a minor extent Ireland) showing lower levels of support if compared to the first six panels. With regard to the development of public attitudes over time, the 1980s have been characterised by a steady rise in support for the EU, which eventually peaked at the moment of the adoption of the Maastricht treaty in 1992, and eventually fell during what has been labelled the "post-Maastricht blues" (Eichenberg and Dalton, 2007). Again, these trends can be seen, to a different extent, in most of the panels covering 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 support for European integration and are reassuring in terms of face validity of the measure.

5 Measurement validity

Apart from face validity, I conduct also more stringent validation procedures. First, I inspect the "convergent validity" (Adcock and Collier, 2001) of the measure, namely the association with other established indicators of the same phenomenon. In particular, I compare the IRT measure with the series produced by Anderson and Hecht (2018) and Guinaudeau and Schnatterer (2019) using the DR algorithm. Then, I assess the fit that these measures have with the observed data to determine which one can better capture latent public support for Europe. 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.

5.1 Convergent validity

If the product of the Bayesian estimation provides a valid measure of public preferences for Europe, then it should also be correlated to other valid measures of the same phenomenon. Anderson and Hecht (2018) and Guinaudeau and Schnatterer (2019) measures introduced above can be used as terms of comparison. Both measures should be reasonably correlated with the IRT estimates as all of them attempt the estimation of a measure of support for Europe by incorporating multiple single-question indicators, although they do so by using different sets of input indicators as well as a different methodology.

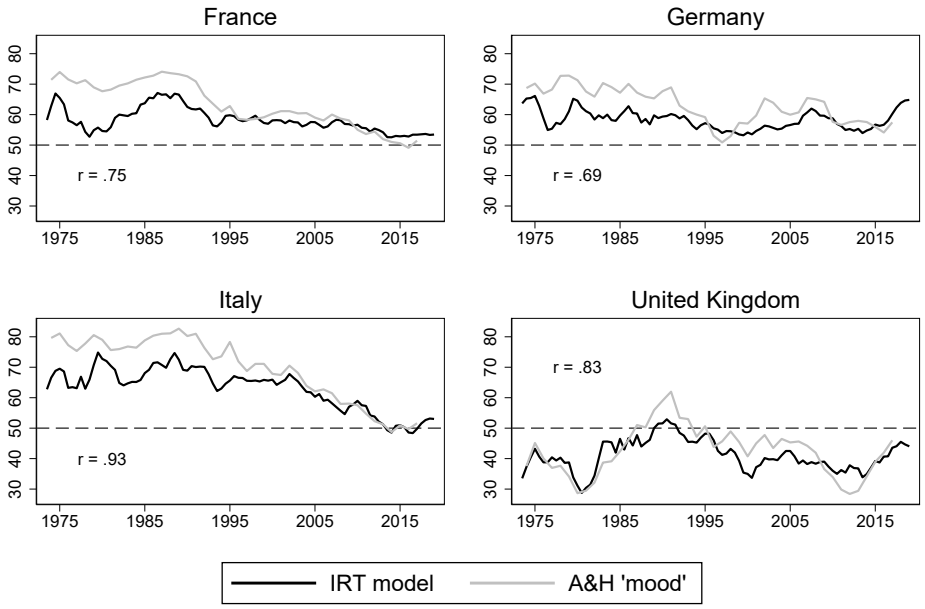

Figure 2: Comparison between estimates of the IRT model and Anderson and Hecht DR algorithm.

The comparison between the IRT-estimated series and the four series of public support measured by Anderson and Hecht (2018) produces very high correlations in Italy (0.94) and United Kingdom (0.84) and high ones in France (0.75) and Germany (0.69). Figure 2 shows that the estimated constructs experience relatively similar trends over time, but there are more or less marked differences in the measured levels of support. In general, the IRT model shows a higher stability of public preferences towards Europe as do the "preference for Europe" series. Differences in terms of

levels and overall trends of support are due both to the set of raw data used and, especially, to the different treatment that the two methods reserve to neutral response options. Indeed, the choice of focusing only on the relative ratio of positive and negative answers and to ignore neutral responses can make the overall levels of support more "extreme", if these responses have a non-negligible share. In addition, in these situations, the omission of neutral responses makes the raw items more prone to marked swings (De Vries and Steenbergen, 2013). Since the DR algorithm combines the changes, means and standard deviations of the raw indicators to calculate the latent public support, these features of the raw items are eventually transferred to the estimated mood.

Taking Germany as an example, it is possible to see the consequences of the points just raised. Figure 3 (left panel) shows that the DR estimate closely tracks the movements of the "membership" question, and the two have indeed a correlation of 0.95 (Anderson and Hecht, 2018, 627). The membership question represents the paradigmatic case of an item that can be better handled by the proposed IRT model. It has a neutral answer option ("EU membership is neither a good nor a bad thing") which, on average, represents the 26.5 percent of responses. Therefore, although only 58.7 percent of Germans have, on average, answered that membership is "a good thing", by ignoring neutral attitudes and "don't knows", and by focusing on the ratio of positive responses over non-neutral ones, the average share of positive responses is a more extreme 85.3 percent. Now, given the high "validity" attributed to the membership question and the fact that in the Anderson and Hecht analysis its relative importance is increased due to the small number of questions used, the estimated mood inherits a lot of the features of this indicator. Similarly, as the right panel of Figure 3 shows, the steep decline in the German "mood" corresponds to the years when the share of respondents saying that EU membership was "a good thing" decreased from 71 to 36 percent (its lowest level) between 1991 and 1997. However, in practice, respondents largely migrated to neutral, rather than negative, attitudes towards membership, as the proportion of "don't know" and neutral answers went from 23 to 48 percent in the same period, whereas the share of negative responses only increased from 6 to 15 percent.

The DR algorithm and the IRT model employed here differ in the way these neutral responses are treated. As studies have shown that ambivalence and neutrality are key elements of public attitudes towards the EU (De Vries and Steenbergen, 2013; Van Ingelgom, 2014), the exclusion of these responses can lead to the loss of important information regarding citizens' support for the EU, particularly in cases where neutral attitudes represent an high proportion of the responses to a given indicator. IRT models, instead, can handle this information coming from the share of

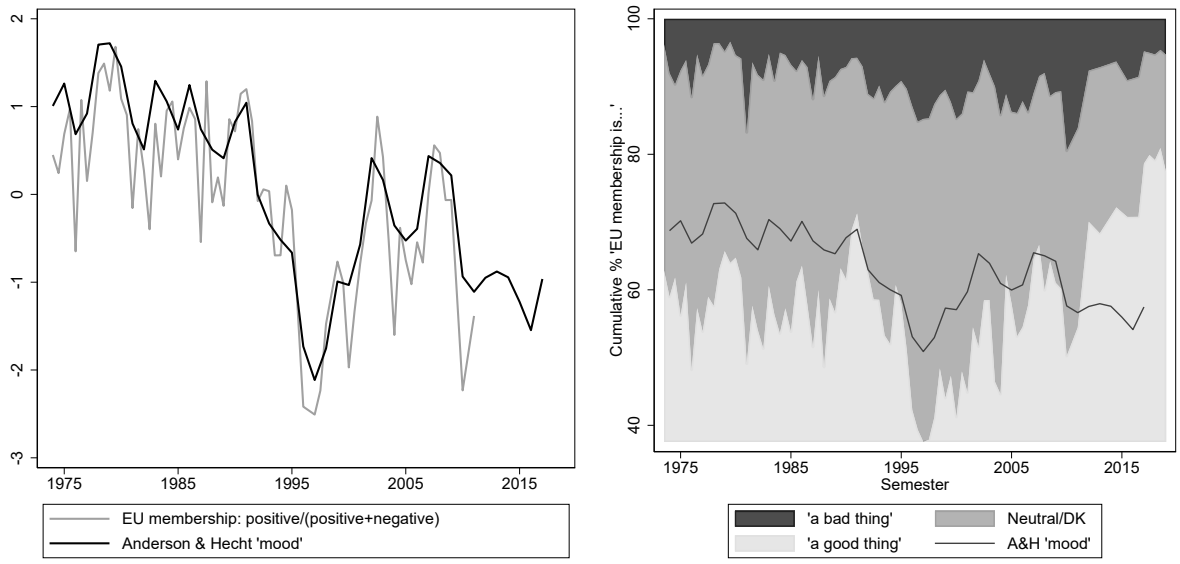

Figure 3: Anderson and Hecht "mood" compared to the "membership" question.

neutral responses, and estimate the most likely value of support for Europe in light of the fact that a certain number of respondents gives neutral answers despite having the possibility of manifesting more negative attitudes.

The work by Guinaudeau and Schnatterer offers even more opportunities to assess the reliability of the estimates as their measure covers 27 member states from 1973 to 2014. The correlation between the EU-average figures is 0.77 (Figure 4). However, there is a lot of cross-country variation as well. The average correlation across all countries of 0.63, with 16 countries showing above-average correlations. A closer inspection of less convergent cases suggests that, although some differences are due to the different raw items used, most of the discrepancies can be indeed explained by the different treatment of neutral answer options (particularly in those countries where they have higher shares) and to the choice of arranging the input to the DR estimation in the form of the ratio of positive responses over non-neutral one.. In a nutshell, the issues raised with regard to the membership question in the German case are by no means peculiar to that country. For instance, Belgium (a country where the two measures are negatively correlated) experienced a similar situation. The membership question, which is strongly correlated to the Guinaudeau and Schnatterer measure ($r = 0.93$), has an average share of neutral and "don't know" responses of 30 percent, peaking at 49 percent in the early 1980s. By designing the IRT models differently (i.e., deliberately discarding the data on neutral responses), the estimated support is more similar to the

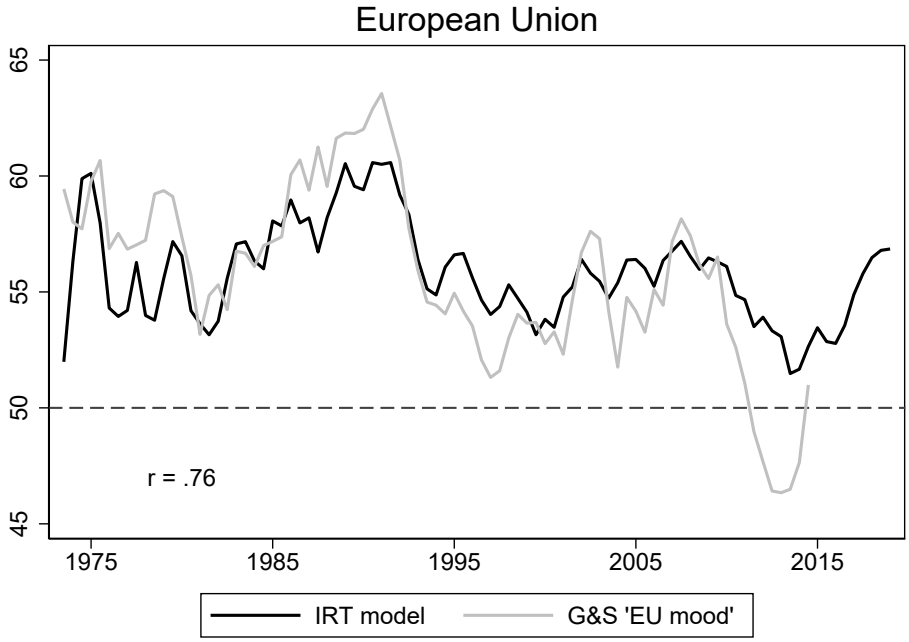

Figure 4: Comparison between the IRT measure and Guinaudeau and Schnatterer "EU mood".

Guinaudeau and Schnatterer "mood" (Figure A2 of the Appendix), suggesting that the differences are explained more by the different *treatment* of the raw items than by the differences in the *set* of items used in the first place.

Therefore, the overall picture shows that the different measures compared are reasonably associated. Differences can be explained by considering both the set of data used and, especially, the way the IRT model treats them. However, these differences also lead to the question of which measure offers a more precise estimate of support for Europe. To determine to what extent IRT models represent an improvement over available measures and techniques, I compare the fit of the different measures with real data. In practice, this means to use the estimated level of support to predict the expected share of pro-European responses of each item administration, a, and compare the latter with the real one to assess the fit of the model.

For the IRT measure, this means to plug in the Bayesian model the estimated support for each semester and the question parameters to predict the proportion of pro-European responses (the reader may refer to equations 2-4 in the Appendix). The measure of fit is then 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 1936 administrations, the RMSE equals 3.39. Using item means to guess the responses of all the questions would produce an average

error of 4.31. To avoid the fact that the predictions can be accurate just because some questions always produce few Eurosceptic (Europhile) answers, 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 $\overline{eu\%_a}$ (*ibidem*):

$$R^2 = 1 - \frac{\sum_{a=1}^{A}(eu\%_a - \widehat{eu\%_a})}{\sum_{a=1}^{A}(eu\%_a - \overline{eu\%_a})} \tag{1}$$

The IRT model has a by-item R^2 of 0.383. This fit can be compared with that of other measures of EU support like one estimated on the same data but with the dyad ratios algorithm, Guinaudeau and Schnatterer (2019) "EU mood", and the EB "membership" question.[6] As Table 1 shows, the "membership" question explains just 23% of the variance in response rates. The "EU mood" measure, instead, explains almost 1/3 of this variance. Of course, the differences between the latter and the IRT estimates are due to both the different set of input data and the different technique used. To properly compare the IRT model with the DR algorithm, I estimated a new measure of latent EU support useing the Dyad Ratios algorithm, but employing the same input items used for the IRT measure presented in this study. The R^2 shows that, although a minor portion of the improvement in model fit over the Guinaudeau and Schnatterer measure can be ascribed to a different set of input question used, the IRT model fits the data better than the DR algorithm even when the same input information is employed, with an increase of explained variance of 20%.

Table 1: Model fit using the IRT measure and alternative measures

	Root mean squared error	*By-item* R^2
IRT model	3.39	0.384
DR algorithm	3.56	0.319
"EU mood"	3.61	0.301
"Membership" question	3.78	0.229
Item means only	4.31	0 (by construction)

[6] As these measures do not rely on an individual-level model of response, I take their estimates and use them in a linear regression to predict the responses to the different items and to calculate items' difficulty and discrimination parameters (Voeten and Brewer, 2006; McGann, 2014).

5.2 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, 542-543). I here test whether the IRT estimates of public support for Europe are related to the share of votes that Eurosceptic parties received in 259 elections since 1999.[7] The dependent variable is, therefore, total vote share of Eurosceptic parties and the unit of analysis is the election. If the IRT estimate is a valid measure of support for European integration, it should be negatively related to the share of votes Eurosceptic parties receive in an election.

To identify Eurosceptic parties I use Chapel Hill Expert Survey data (Bakker et al., 2015; Polk et al., 2017; Bakker et al., 2020). The survey collects party positions on European 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 European 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). Given that the focus is on whether the total share of Eurosceptic vote is inversely related to the level of EU support, for each election, the total vote share of Eurosceptic parties is calculated as the share of votes received by all such parties. Tobit models are used to account for the fact that the dependent variable is bounded between 0 and 100.[8]

The bivariate model in Table 2 shows that Eurosceptic vote share is negatively associated with support for Europe. Model 2 introduces some political and economic controls commonly used in the analysis of support for Eurosceptic parties. First, it adds country fixed effects to deal with country-level confounders. Secondly, annual unemployment and GDP growth rates are included to account for the fact that economic insecurity might be an argument to cue voters against the EU (De Vries and Edwards, 2009). Similarly, both support for Eurosceptic parties and attitudes towards the EU 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 among the

[7] Data on vote shares are collected from the ParlGov database (Döring and Manow, 2019). Both national and European Parliament elections are considered.

[8] Alternative estimator are reported in Section B of the Appendix.

Table 2: Tobit regression models of public support for Europe and Eurosceptic vote share.

	DV: Eurosceptic vote share			
	Model 1		Model 2	
IRT-estimated EU support	-1.239***	(0.185)	-0.670**	(0.239)
Previous Eurosceptic vote share			0.286*	(0.111)
Election type (1=European)			2.731	(1.523)
% exclusive national identity			-0.116	(0.146)
% foreign-born population			-0.896	(0.904)
% foreign-born labour force			-0.0865	(0.877)
Net contribution to EU budget			-0.350	(0.461)
Annual GDP growth rate			0.214	(0.183)
Annual unemployment rate			0.0609	(0.262)
Year			0.877***	(0.247)
Constant	66.33***	(10.12)	-1708.9***	(493.5)
Observations	259		259	
Country fixed effects			Yes	
LR χ^2	48.82		255.14	
Prob. $> \chi^2$	0.000		0.000	
Log-likelihood	-507.552		-404.389	

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

labour force and in the population as a whole are used to measure the strength of identitarian concerns as well as the perceived threat to national identity.[9] 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., 2018). 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 (Schmidt, 2019) and for the effect of less salient EP elections (De Sio et al., 2019).

Recalling that the primary purpose of this analysis is to assess the construct validity of the Bayesian measure, the results of Model 2 are particularly encouraging. Only two of the nine controls added proved to be significantly associated (in the expected direction) with Eurosceptic vote share and, most importantly, the coefficient of the public preference variable remains correctly

[9] The first variable is the share of respondents answering "(NATIONALITY) only" to the Eurobarometer question "In the near future, do you see yourself as...?". The other two come from the ILO and OECD data, respectively.

signed and well below the conventional 5% threshold even with the inclusion of the control variables. The analysis, therefore, provides further evidence of the validity of the Bayesian estimation.

6 Conclusion and discussion

This study has showed that Bayesian IRT models allow the production of valid estimates of public preferences towards Europe, comparable across member states and over time. I argue that IRT models represent an improvement over existing approaches to the measurement of EU support, like single-question indicators and the Dyad Ratios 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 percentage of responses to 1936 Eurobarometer questions, I have generated estimates of public preferences for European integration in each member state of the European Union for every semester from autumn 1973 to autumn 2019. These measures have been validated both by inspecting their correlation with existing alternative measures of the same concept and by looking at their association with the vote share of Eurosceptic parties in 259 elections conducted over the last twenty years in all EU member states. Additionally, I have showed that, despite the relatively high association with DR measures, IRT models produce a more precise estimate of the underlying level of public support for Europe. 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 like the "membership" question is easier and more efficient to implement, it comes with different limitations. In fact, this measure explains just about half of the variance in expressed preferences explained by the IRT estimates, confirming that single-questions indicators represent only partial and incomplete measures of a latent phenomenon like EU support (Anderson and Hecht, 2018, 633). This consideration adds up to other more practical concerns related 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, IRT models can easily accommodate the interruption in a specific data series, as far as there are other indicators that can be used to measure the presence of the underlying attribute of interest. To some extent, this can be also achieved with the DR algorithm, the first dimension reduction technique used to measure EU support over a long time frame. DR and

IRT estimates show reasonably strong convergence in most cases, suggesting that both techniques allow the production of valid and consistent measures of EU support. Yet, apart for a sounder theoretical grounding, IRT models provides also what I argue to be a more precise measurement of the latent level of public support for Europe. The evidence for this is that, 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, if with single-question indicators there is an efficiency-precision trade-off, this is not the case when choosing between the DR algorithm and Bayesian IRT models, as they both require exactly the same type of input data: the population-level response shares, the date of the surveys and their the sample size.

Therefore, this methodological novelty comes with both theoretical and empirical gains. On the one hand, Bayesian IRT models offer a theoretically-grounded alternative to the use of the DR algorithm for the estimation of latent opinion. Contrary to the DR algorithm, IRT models are based on an individual-level model of behaviour, allow a sounder treatment of neutral responses and are better suited to recover a more precise level of public support. On the other, 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 the analysis at hand, and sometimes may not even be available for a specific task.

In fact, the EU landscape is characterised by the lack of long trends of public preferences towards integration in particular policy fields (Zhelyazkova et al., 2019). Yet, what is available is a range of sporadically asked items capturing attitudes towards specific policy initiatives, which can be combined to produce a more homogeneous measure of public preference towards EU integration in a given domain. For example, there is no such a thing as a single-question indicator capturing preferences for economic policy integration comparable in terms of continuity and length to the "membership" question. Yet, there are different items administered with varying regularity asking about preferences on the creation of a direct EU taxation system, the common supervision of the national public debts, the EU being responsible for domestic redistributive policies or pensions, whether the European Parliament should be competent for taxation matters, and so on. This narrower set of items can be used in an estimation process analogous to the one validated above to produce an estimate of public support for EU integration in the economic domain, that combines the information coming from different items which, on their own, offer only a partial, issue-specific, information about public preferences for economic integration. To the extent that it is possible to identify an appropriate number of different indicators, this solution can of course be replicated for

different policy domain as well (see McGann et al. 2019, 51-54). An exercise of this sort is discussed in section C of the Appendix with regard to home affairs and economic policies.

Of course, the implementation of IRT models is more time-consuming than the employment of single-question indicators, as it requires the collection of more input data. Hence, in the event that appropriate indicators are available, they might be preferred to a computationally more demanding technique like Bayesian IRT. However, by comparing the explanatory performances of both measures, I have showed that this preference for more efficient techniques might come at the price of less precise estimates of public support.

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, C. J. and Hecht, J. D. (2018), 'The preference for Europe: Public opinion about European integration since 1952', *European Union Politics* **19**(4), 617–638.

Bakker, R., De Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M. and Vachudova, M. A. (2015), 'Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010', *Party Politics* **21**(1), 143–152.

Bakker, R., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M. and Vachudova, M. A. (2020), *Chapel Hill Expert Survey. Version 2019.1*, Available on chesdata.eu. University of North Carolina, Chapel Hill.

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

Boomgaarden, H. G., Schuck, A. R., Elenbaas, M. and De Vreese, C. H. (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. (2018), 'EU transfers and Euroscepticism: Can't buy me love?', *University of Zurich, Department of Economics, Working Paper* (289).

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* .

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

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

Clinton, J., Jackman, S. and Rivers, D. (2004), 'The statistical analysis of roll call data', *American Political Science Review* **98**(2), 355–370.

Dahl, R. A. (1998), *On democracy*, Yale University Press, New Haven.

De Sio, L., Franklin, M. N. and Russo, L. (2019), *The European Parliament Elections of 2019*, LUISS University Press, Rome.

De Vries, C. E. (2018), *Euroscepticism and the future of European integration*, Oxford University Press, Oxford.

De Vries, C. E. and Edwards, E. E. (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. (2019), 'Parliaments and governments database (ParlGov): Information on parties, elections and cabinets in modern democracies', *Development version* .

Eichenberg, R. C. and Dalton, R. J. (1993), 'Europeans and the European Community: The dynamics of public support for European integration', *International organization* **47**(4), 507–534.

Eichenberg, R. C. and Dalton, R. J. (2007), 'Post-Maastricht blues: The transformation of citizen support for European integration, 1973–2004', *Acta politica* **42**(2-3), 128–152.

Erikson, R. S., MacKuen, M. B. and Stimson, J. A. (2002), *The macro polity*, Cambridge University Press, Cambridge.

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

Gabel, M. J. (1998), *Interests and integration: Market liberalization, public opinion, and European Union*, University of Michigan Press, Ann Arbor.

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**(03), 1187–1197.

Hagemann, S., Hobolt, S. B. and Wratil, C. (2017), 'Government Responsiveness in the European Union: Evidence From Council Voting', *Comparative Political Studies* **50**(6), 850–876.

Hobolt, S. B. and De Vries, C. E. (2016), 'Public support for European integration', *Annual Review of Political Science* **19**, 413–432.

Hobolt, S. and Brouard, S. (2011), 'Contesting the European Union? Why the Dutch and the French rejected the European constitution', *Political Research Quarterly* **64**(2), 309–322.

Hoeglinger, D. (2016), 'The politicisation of European integration in domestic election campaigns', *West European Politics* **39**(1), 44–63.

Hooghe, L. and Marks, G. (2004), 'Does identity or economic rationality drive public opinion on European integration?', *PS: Political Science & Politics* **37**(3), 415–420.

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*, Academic Press, Cambridge.

Levendusky, M. S. and Pope, J. C. (2010), 'Measuring aggregate-Level ideological heterogeneity', *Legislative Studies Quarterly* **35**(2), 259–282.

Lubbers, M. (2008), 'Regarding the Dutch 'Nee' to the European Constitution: A test of the identity, utilitarian and political approaches to voting 'No'', *European Union Politics* **9**(1), 59–86.

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

McGann, A. J. (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.

McLaren, L. (2005), *Identity, interests and attitudes to European integration*, Palgrave Macmillan, London.

Nunnally, J. C. and Bernstein, I. H. (1994), *Psychometric theory*, McGraw-Hill, New York.

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)'.

Polk, J., Rovny, J., Bakker, R., Edwards, E., Hooghe, L., Jolly, S., Koedam, J., Kostelka, F., Marks, G., Schumacher, G. et al. (2017), 'Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data', *Research & Politics* **4**(1), 1–9.

Rauh, C., Bes, B. J. and Schoonvelde, M. (2020), 'Undermining, defusing, or defending European integration? Assessing public communication of European executives in times of EU politicization', *European Journal of Political Research* **59**(2), 397–423.

Schmidt, V. A. (2019), 'Politicization in the EU: between national politics and EU political dynamics', *Journal of European Public Policy* **26**(7), 1018–1036.

Stimson, J. A. (1991), *Public opinion in America: moods, cycles, and swings*, Westview Press, Boulder.

Stimson, J. A. (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.

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.

Van Ingelgom, V. (2014), *Integrating indifference: A comparative, qualitative and quantitative approach to the legitimacy of European integration*, ECPR Press, Colchester.

Voeten, E. and Brewer, P. R. (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.

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