Studying Individual-Level Attitudes Towards Important Problems Using Pairwise Comparisons *

Benjamin E Lauderdale *University College London*

How can we measure citizens' attitudes towards political problems in a way that generates data which are comparable across individuals, across countries, and across time? In this paper, I show how to measure these attitudes using pairwise comparisons, such that the assessments can be used as either individual-level dependent or independent variables for subsequent analyses. Even where it is infeasible for respondents to answer sufficient pairwise comparisons to recover a respondent-level ranked order, model-based multiple imputation enables the analyses that would be possible if all respondents compared all possible pairs of categories. I validate this approach using assessments of political problems in the UK according to the categories of the Comparative Agendas Project. By enabling individual-level analysis on researcher-defined categories with accessible questions, these methods make supplementing or replacing the typical most important problem and issue questions asked on national election studies with pairwise comparisons broadly advantageous for future research.

Funded by the European Union [ERC-2024-ADG 101198965]. Views and opinions expressed are however those of the author's only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the European Research Council can be held responsible for them.

*This version: 13 February 2026

Introduction

Citizens expect their governments to fix problems in their societies. When governments fail to address problems through a lack of effort, this can be understood to be a failure of representation; when governments fail to address problems despite effort, this can be understood to be a failure of competence. Sufficiently severe problems can lead to changes in voting behaviour in democratic countries, and undermine governments in non-democratic countries. Understanding citizens' views about problems is consequently important to understanding public opinion and political behaviour, to understanding representation and the quality of governance, and to understanding the stability of specific governments and entire political regimes, both democratic and non-democratic.

Despite this centrality of problem perceptions to politics, political scientists have struggled to reliably and comparably measure which problems citizens are concerned about. The multifarious nature of citizens' problem perceptions presents a methodological obstacle to researchers. There are many kinds of societal problems that may become the subject of politics, there are many potential categorizations of these, and their scale of assessment is difficult to specify. There is a long tradition of national election studies asking an open-ended "most important problem" or "most important issue" question, which skirts these challenges by letting respondents determine the form of their answer, at the expense of generating unstructured text data that is difficult to compare across individuals, across countries, and across time. As a consequence, there is limited cross-national or longitudinal political science research examining which problems citizens are most concerned about, where these views come from, or how these views shape their political activity (for a review, see Heffington, Park and Williams, 2019).

One avenue for generating more structured data on problem perceptions is through closed-ended survey questions that ask respondents to make pairwise comparisons of specific pairs of problems. Pairwise comparisons form an increasingly common strategy for measurement in political science. This is particularly the case when researchers want to learn *relative assessments* of a large number of

objects against a *criteria* which is not straightforward to quantify. For example, Zucco Jr, Batista and Power (2019) measure which of 37 ministerial portfolios are most desirable by asking Brazilian politicians to choose which they would prefer from randomly selected pairs. Hopkins and Noel (2022) measure the perceived ideology of 100 US Senators by asking activists to choose which Senator is more liberal/conservative from randomly selected pairs. Blumenau and Lauderdale (2024) measure the persuasiveness of 14 different categories of political rhetoric each applied to each of 12 different issues by asking citizens to choose between randomly selected pairs of political arguments. In each of these examples there is an assessment criteria (desirability of ministries, ideology of legislators, persuasiveness of arguments) communicated to a relevant population of assessors (politicians, activists, citizens) in the form of a survey question that requires only directional, comparative assessments against the criteria.

Pairwise comparisons are attractive because they are undemanding for respondents. In this paper, I measure citizen perceptions of the relative degree to which there are important problems in each of the 21 political domains defined by the Comparative Agendas Project (CAP) (Jones et al., 2023; Bevan, 2025). It would be tedious for an expert, let alone a survey respondent, to *rank* the relative importance of current problems in 21 areas. It would be extremely difficult for an expert or a survey respondent to *rate* the importance of problems in each of these 21 on an abstract interval or ordinal level scale in a consistent way. It is, by contrast, extremely easy for a respondent to indicate whether they think there are more important problems with the economy or with immigration. The essential insight of comparative judgement methods[1] is that for many criteria individuals cannot articulate interval level ratings, and they can only express ordinal level rankings through working these out from their relative assessments of specific pairs of objects. The best measurement strategy is therefore one that directly collects these pairwise assessments, and reserves the problem of aggregation to the analyst.

If you ask enough questions of this pairwise form to a set of respondents with the relevant assessment expertise, you can straightforwardly estimate *consensus* rankings or ratings. But what if one

[1]For a review of the use of comparative judgement methods in a range of research fields and meta-analysis with respect to aggregate scale reliability, see Kinnear, Jones and Davies (2025).

wants to describe how these assessments *vary* across different respondents at the individual level as a function of a variety of individual characteristics? What if one wants to study how individuals who make different assessments have other differing views or behave differently?

Pairwise comparisons appear poorly designed to answer these questions. Typical designs involve individuals assessing a small, randomly selected set of pairs constructed from the pool of objects. For even small numbers of objects n_j, it is very demanding to get any single respondent to make all $\frac{n_j(n_j-1)}{2}$ pairwise comparisons of n_j objects. It is small consolation that if respondents make deterministic and transitive assessments, one need only do the comparisons necessary to sort (i.e. rank) the objects being assessed. This form of adaptive testing reduces the rate at which the number of needed comparisons increases with n_j from $O(n_j^2)$ to $O(n_j \log n_j)$ (Knuth, 1998). However, adaptive testing can create inferential problems with non-deterministic comparisons (see e.g., Bramley and Vitello, 2019), and in any case this is still usually too many pairwise comparisons. If, for example, you have 20 categories, to guarantee recovery of a full individual-level ranking of these for all possible rank orderings–assuming deterministic and transitive responses for each respondent–can require as many as 62 pairwise comparisons , which will exhaust the patience of most survey respondents and the budgets of most survey researchers.[2] This appears to create a problem: if you cannot ask any single respondent enough questions to determine their rank ordering of problems, and if you cannot ask all respondents the same questions without radically curtailing the number of objects you study, how can you study variation in individuals' assessments of the full set of objects?

This paper proceeds from the observation that, even if one wishes to study variation in assessments *across* individuals, one is seldom interested in recovering a full rank ordering of all objects for any specific individual. Those individuals are, after all, just a sample from the population in which we want to study variation. Thus, this paper is focused on how the *partial information* that we can learn about each individual from limited numbers of pairwise comparisons can be used to enable analyses that relate variation in assessments to other quantities describing individuals in a population. As we

[2] Adaptive selection of pairs is also difficult to implement on many survey platforms.

ask someone more and more paired comparisons, we are gaining information about their personal rankings, up to the point where the pairs we have asked reveal the full ranking. The obstacle that this paper seeks to overcome is how we can use this partial information to analyse how assessments vary across different individuals and how these relate to other measured quantities.

The goals is to be able to complete all the kinds of analyses that survey researchers would with a fixed survey question asked of all respondents to a social or political survey. Researchers might want to describe citizens' problem assessments as a *dependent variable* as a function of measured covariates. For example, one might wish to study how problem perceptions differ by demographic variables such as gender, age, education, etc; or by variables capturing political attention or interest; or by political orientation variations such as past vote or party identification (Heffington, Park and Williams, 2019); or in response to an experimental or quasi-experimental treatment. Alternatively, researchers might want to to describe some other variable as a function of these problem assessments as an *independent variable*. For example, one might wish to use problem assessments as a predictor of how voters choose between candidates or parties (Green and Hobolt, 2008), in particular in understanding heterogeneity in how different voters assess candidate features, analogously to studies of issue salience (e.g. Fournier et al., 2003; Leeper and Robison, 2020).

To achieve the goal of facilitating all of these kinds of analyses, this paper introduces and validates methods for multivariate regression modelling of pairwise comparison data. The central contribution of this paper is to show that these methods straightforwardly enable us to study the dependent variable of citizens' assessments, and also that they also allow us to study citizens' assessments as independent variables, despite each respondent completing different pairwise comparisons and far fewer than is necessary to recover a full personal ranking. This is enabled through model-based individual-level measurement of latent ratings of all objects and model-based imputation of missing pairwise comparisons for all respondents in a data set. Feasible numbers of pairs per respondent can yield data capable of reliably answering substantive questions of interest. In my application, with 21 categories to be compared, 2065 respondents, and 8 pairwise comparisons per respondent, there is

more than enough data to reliably recover population variation in assessments across respondents, as well as to recover a meaningful signal about each respondent's latent assessments, and to run analyses based on imputing missing pairwise comparisons for all respondents.

This is an approach that is potentially valuable well-beyond this application to problem perceptions, and indeed beyond political science, to a broader class of assessment and measurement problems that are studied across the social sciences. However, the most immediate implication of this application is for the design of national election studies, where I urge a reconsideration of the questions we ask about political problems. These surveys provide a key facilitative role for political public opinion research. As I describe in more detail below, existing approaches to measuring problem perceptions are widely understood to be deficient in several respects which are directly redressed by the approach of collecting pairwise comparison data in conjunction with the modelling strategies proposed in this paper. One of these deficiencies with the most common existing approach is poor intertemporal comparability, which means that the usual arguments for maintaining sub-optimal historical survey question formats are weak. The sooner that political scientists start collecting interpersonally, intertemporally, and internationally comparable data on citizens' problem perceptions, the sooner the field can develop an appropriately rich understanding of this vital indicator of citizens' expectations for their governments.

Measuring the Problems That Citizens Think Are Important

The most common way that political scientists have measured citizen attitudes towards political problems is through open-ended "most important problem" (MIP) or "most important issue" (MII) questions, such as "What do you think is the most important problem facing the country today?". Variations on this question have been asked by Gallup since 1935 in the US and since 1947 in the UK (Jennings and Wlezien, 2011, p547), and by national election study series in the US starting in 1960 and in the UK starting in 1963. Variations on these questions are found on nearly all current national election studies, including the American National Election Study, the British Election Study, the Canadian

Election Study, the Dutch Parliamentary Election Study, the European Parliament Election Study, the French Electoral Study, the German Longitudinal Election Study, and so on.

These questions have come in for scholarly criticism along at least four dimensions. First, there are ambiguities of the "important problem" wording, and the question of whether citizens understand the word "important" to apply to the severity of problems in the domain or to the stakes of the domain more generally (Wlezien, 2005; Jennings and Wlezien, 2011). Second, there is the superficiality of (usually) only asking for the highest ("most") category among many, which yields limited information about individuals' portfolio of concerns, and means that fluctuations in the frequency of specific problems being mentioned are as sensitive to decreases in the importance of other problems as to increases in the specified problem (Enns, 2014; Pickett, 2019). Third, use of open-ended responses usually involves ex post category coding (e.g. Jones and Baumgartner, 2005; Mellon et al., 2024) that is constrained by what respondents spontaneously say, reducing researchers' ability to study categories for which we have other kinds of data that we wish to compare public opinion to, or which can be consistently applied across countries and across time. Fourth, respondents do not systematically consider all possible problems or issues, generating responses in the moment that are temporally unstable (Johns, 2010) and which can yield very different conclusions about the relative importance of issues by comparison to closed-ended formats (Fournier et al., 2003; Yeager et al., 2011)[3]

These methodological limitations appear to have been a meaningful obstacle to research related to citizens' assessments of political problems. By comparison to most national election study questions, MIP/MII questions are rarely used in research. There are very few studies that track citizens' problem perceptions over time, and aside from one three country study on a six category coding of responses (Hobolt and Klemmensen, 2008), these are limited to single countries (Pennings, 2005; Jones and Baumgartner, 2005; Heffington, Park and Williams, 2019). When Jones and Baumgartner

[3]The major advantage of these questions is that they are open-ended, and enable researchers to learn about the concerns of citizens in the language that those citizens understand those concerns. This is a worthwhile kind of evidence to collect, but is at odds with the reality of a typical political science survey composed almost entirely pre-specified, closed-ended survey questions designed to facilitate population inference. These open-ended questions more naturally act as the entrée to an interview or focus group, where one can further investigate how each citizen has expressed their political concerns.

(2005) code US MIP responses onto a specific set of categories for purposes of comparison to legislative attention, they conclude that "The public agenda space is constrained, and it is dominated by national and economic security.... with an occasional 'spike' of other concerns. These other concerns tend to occur together, because the absence of economics and defense opens somewhat of the 'window of opportunity' for these other concerns to move up in the collective attention of Americans. These windows are relatively rare." (p253). But before summarising the data this way, the authors also acknowledge that it is difficult to tell whether this is a substantive conclusion about public attention to problems, or whether it is more a result of the question wording, the "most" format, and the limitations of ex post categorization focusing responses onto only a few problem categories (Jones and Baumgartner, 2005, p251-252).

The most obvious alternatives are closed-ended assessments of researcher-defined categories. The French Electoral Study provides a closed-ended most important problem question with a long list of categories to choose from. Other election studies have supplemented the classic MII/MIP question with queries about specific problems of interest at the time of the study. For example, the 2024 American National Election Study (ANES), after asking a MIP question, asked a series of questions of the form "How important is the following issue in the country today?", with categorical response options from "Not at all important", "Slightly important", "Moderately Important", "Very Important", "Extremely Important". These were asked about "Illegal Immigration", "The cost of living and rising prices", "Climate change", "Gun policy", "Abortion", "Crime", "The war in Gaza", What's being taught in public schools", "Keeping the U.S. a democracy", and "Racial inequality". These questions have the benefit of gaining specific responses to researcher-defined categorisations. But single rating questions are time consuming, and expressed variation in attitudes towards the problems is potentially confounded by differential interpretations of the adjectives. As is the case for the French Electoral Study question, several of the categories are country and/or time-specific, and are consequently not a model for comparative research.

If we want to do comparative research, it is helpful to look to existing cross-national schemes for

categorising public policy activity by governments, which define policy domain categories to which we might want to link public attitudes. The question of whether governments are acting to try to address the areas of greater concern to the public is a fundamental question of the quality of representation, and one that is much easier to answer if one can ensure that there is consistency between how one measures public concern and government activity. Of course there are many possible categorisations of the "domains" or "areas" of politics and policy, and arguing for a particular one of these with respect to the problem perception measurement problem is not the aim of this paper. As my aim here is to illustrate methods for analysing data that would apply equally to data generated using any such categorisation, I have chosen to adapt a single existing categorisation for this study.

The Comparative Agendas Project (CAP) is "a network of many projects aimed at classifying political agendas according to the policies they address" (Jones et al., 2023; Bevan, 2025). This consists in a unified coding scheme anchoring country-specific coding schemes developed by different authors. There is a natural linkage between a problem domain typology and CAP's aim of providing a policy domain typology. Insofar as a political "problem" is a thing that "policy" aims to mitigate, categories that make sense for grouping the latter are likely to also work reasonably well for the former. CAP aims to be a "measurement system [that] allows the exploration of many different ideas grounded in different approaches, theories or empirically derived hypotheses" (Jones, 2016), and so closely aligning a measurement system for public assessments to this existing one for policy activity and attention is of obvious value.

The CAP has 21 high level "topics", all but two of which have defined "sub-topics". The 21 topics are: "Macroeconomics", "Civil Rights", "Health", "Agriculture", "Labor", "Education", "Environment", "Energy", "Immigration", "Transportation", "Law and Crime", "Social Welfare", "Housing", "Domestic Commerce", "Defense", "Technology", "Foreign Trade", "International Affairs", "Government Operations", "Public Lands", and "Culture". While some of these topics have names which may not be legible to citizens (e.g. "Macroeconomics"), the codebook for the CAP and the definitions of the sub-topics within these facilitate presentation to respondents. For this study, I amend the names of four of

the CAP categories, in order to make them more accessible to UK survey respondents: changing "Macroeconomics" to "The Economy", "Civil Rights" to "Rights and Freedoms", "Labor" to "Work and Employment", and "Domestic Commerce" to "Regulation".

The prompt to survey respondents that I will use closely mirrors the typical language of a "most important problem" question. This choice should not be read as a disagreement with the arguments raised by Wlezien (2005) and Jennings and Wlezien (2011) regarding ambiguities of this question wording. The methods following would all apply if the wording were altered further, such as by asking about the "severity" of problems or about which problems are "worse" (as in Lauderdale and Blumenau, 2025). The prompt, for a pairwise comparison of two categories A and B, reads as follows:

- "In which of these two areas do you think this country faces more important problems today?"
  - [area A] ([area A description])
  - [area B] ([area B description])
  - "The problems in both areas are similarly important"

I use the high level category names (with the four amendments described above) as [area A] and [area B], and then provide parenthetical descriptions of these based on the sub-topic categories, full details of which are provided in the appendix. For example, the "The Economy" category is described "(including inflation, jobs and unemployment, interest rates, the government budget, and taxes)". For "Immigration" and "Culture", where the CAP does not have agreed sub-categories, I have written similar descriptions: "(including refugees, unauthorised migrants, skilled worker visas, family visas, and pathways to permanent status and citizenship)" and "(including language, arts and music, libraries, heritage and preservation, cultural industries and exchange, and diversity and inclusion)".

Any substantive conclusions will be sensitive to how categories are described, and so these choices regarding amending the category names and describing what those categories contain are consequential. Descriptions encourage respondents to think about the range of sub-topics that the CAP coding scheme actually puts within that topic, and so this presentation may help ensure that the responses

reflect that categorization. Or it may be that the descriptions are an unnecessary complication, and the topic/domain names would suffice. The focus of this paper is on statistical methods for analysis of pairwise comparison data. I am not aiming to argue for either the prompt or these categories worded in these specific ways, but rather for using a measurement instrument of this pairwise form. The requirement for present purposes is that these are plausible presentations of the categories, which are likely to generate pairwise comparison data with the kinds of statistical relationships that would be found also with different wordings and categorizations.

The survey prompt provides respondents with a neutral response category ("The problems in both areas are similarly important"), which reflects weighing several important design considerations. One can provide this kind of neutral response allowing the respondent to report approximate equality between the two objects, one can provide a non-response option (e.g. "I do not know"), neither of these, or both of these. Providing neither a neutral nor a non-response option yields "forced choice" data that is much easier to analyse, but at the cost of potentially annoying respondents and/or inducing them to drop out of the survey. It is important to avoid inducing differential non-response, either by respondent characteristics or for particular objects of evaluation that might be more sensitive to some respondents, as it undermines our ability to make reliable population-level inferences from the data. A ternary format with a neutral response has the benefit of recovering meaningful information about which objects are judged similar with respect to the criteria. This approach also gives respondents a mechanism to not take a view between the two alternatives if this is the reason they might be inclined towards a "don't know" option, and so I include a neutral response option but not a non-response option.

A Multivariate Regression Model for Pairwise Comparison Data

Paired comparison data is typically analysed using models from a broad family associated with work by Bradley and Terry (1952), as well as Thurstone (1927) and Mosteller (1951). These *Bradley-Terry* models can be understood as latent variable models, where the observable choice between two al-

ternatives is a function of the difference between two unobserved interval level "scores" or "ratings" associated with the two alternatives on a common, continuous, interval-level scale. These models can be motivated from a wide variety of perspectives and are consequently ubiquitous in the modelling of pairwise choice data (Hamilton, Tawn and Firth, forthcoming). One relatively inconsequential detail is the scale on which the response is linked to the latent variable, so for mathematical convenience I adopt a latent normal scale.

In the case where there are no "ties" (i.e. no neutral outcomes in the pairwise comparisons), the simplest version of this model is:

$$p\left(j \text{ defeats } j'\right) = \Phi\left(\psi_j - \psi_{j'}\right)$$

where $\Phi()$ is the standard normal CDF. This is equivalent to

$$\begin{aligned} j \text{ defeats } j' &\iff y^*_{jj'} > 0 \\ y^*_{jj'} &= \psi_j - \psi_{j'} + \epsilon \\ \epsilon &\sim N\left(0, 1\right) \end{aligned}$$

The latent variable ψ_j is a parameter describing the relative propensity of j to win in pairwise comparisons against potential competitors $j' \in 1, \ldots, n_j$. The errors of all pairwise comparisons are treated as fully independent of one another, and the focus is on the estimation of a single ψ_j parameter for each object j. This approach makes the most sense in an application where the source of unpredictability in the outcomes of the comparisons neither has any multi-level structure, nor is of any interest in itself.

The context being considered here, however, is a survey context in which a large number of different respondents are each being asked to make *multiple* comparisons between a set of alternatives. Here, the unpredictability in specific comparison outcomes partially arises from different individuals i having different assessments ψ_{ij} about the various objects j, with respect to the criteria that they are

being asked to apply. While consensus assessments across the set of respondents remain of interest, the goal is to characterise how these assessments ψ_{ij} vary across individuals i.

Because ψ_{ij} for a vector quantity for each respondent, the natural framework for modelling this latent scale of ψ is a general linear model (Mardia, Kent and Taylor, 2024). I define ψ_{ij} as respondent i's individual latent rating of problem j. I assume that these have an average value in the population and a covariation structure around this average. I assume that the vector ψ_i for respondent i for all problems $j \in 1, \ldots, n_j$ follows a multivariate normal distribution:

$$\psi_i \sim MVN\left(\alpha + \beta X_i, \Sigma\right)$$

for some $1 \times n_k$ vector of covariates X_i describing the respondent i and estimate a $n_j \times 1$ vector of intercepts α and a $n_j \times n_k$ matrix of coefficients β that describe variation in latent score as a function of each of the n_k covariates on each of the n_j categories.

This model captures variation in latent assessments first by identifying measurable features X_i of individuals i that are associated with systematically different assessments of particular js. I standardize each variable k in X to have mean zero and standard deviation to maintain the α as the overall mean for each problem j across the sample/population. I impose a sum-to-zero constraint across problems j on the intercept α and coefficients β for each variable k to achieve numerical identifiability of the latent scale. In order to avoid over-fitting, given the large number of coefficients in the model, the α and β_k are assumed to have normal prior distributions with mean zero and standard deviation σ_α and σ_{β_k}, with these standard deviations of effect magnitudes themselves having a half normal prior with standard deviation 3.

This model also captures variation in latent assessments by identifying correlations in individual-level assessments of objects j and j' that exist across the set of individuals i. The covariance matrix Σ describes variation in individual respondents' latent ratings around the conditional average, which can be decomposed $\Sigma_{jj'} = \rho_{jj'}\sigma_j\sigma_{j'}$ such that the $\rho_{jj'}$ are correlation coefficients.

Having defined this general linear model for latent assessments, we can now specify a response

model that links the latent assessments ψ to observable choices. When a respondent i chooses between two alternative categories j and j', generating the observable ordered choice $Y_{ijj'} \in \{1, 2, 3\}$—corresponding respectively to selecting category j, selecting the neutral response, and selecting category j'—I assume they are doing this based on the difference $\psi_{ij'} - \psi_{ij}$ between their personal latent ratings of the two categories. This latent difference is probabilistically mapped onto a three level ordered response with a neutral option, according to the following ordered probit framework:

$$p\left(Y_{ijj'} = 1\right) = 1 - \Phi\left(\frac{\delta + \left(\psi_{ij'} - \psi_{ij}\right) - \gamma_{i1}}{\omega}\right)$$

$$p\left(Y_{ijj'} = 2\right) = \Phi\left(\frac{\delta + \left(\psi_{ij'} - \psi_{ij}\right) - \gamma_{i1}}{\omega}\right) - \Phi\left(\frac{\delta + \left(\psi_{ij'} - \psi_{ij}\right) - \gamma_{i2}}{\omega}\right)$$

$$p\left(Y_{ijj'} = 3\right) = \Phi\left(\frac{\delta + \left(\psi_{ij'} - \psi_{ij}\right) - \gamma_{i2}}{\omega}\right)$$

The parameters δ, ω and γ each correspond to assumptions about how respondents engage with the question, which I describe in turn.

First, the parameter δ is an estimated order effect bias across all respondents towards selecting whichever category is listed first, regardless of which category it is. Second, the model assumes a test-retest error that is independent normal with mean zero and standard deviation ω. This helps to distinguish between variation in responses to the same pair jj' across respondents–which is captured through different respondents having different ψ–and variation in responses to the same pair jj' within respondents if asked the same pair multiple times. Thus, the model does not assume that the same respondent will give consistent transitive responses, or that they will give the same response every time to the same paired comparison if asked about it multiple times.[4]

Third, the model assumes that each respondent has their own cutpoints γ_{i1} and γ_{i2} which determine their propensity to use the neutral category.[5] I restrict these to be $\gamma_{i1} = -\frac{e^{\xi_i}}{2}$ and $\gamma_{i2} = \frac{e^{\xi_i}}{2}$, where

[4]In order to determine how much of the variability in responses is due to test-retest instability within individuals, as opposed to being due to variation in individual attitudes across individuals, it is necessary to have opportunities for respondents to reveal instability and/or intransitivity. This will occur if respondents sometimes see the same pair twice or answer closed loops of connected pairs.

[5]With forced choice binary data $Y_{ijj'} \in \{0, 1\}$, all respondents can be treated as having a single cutpoint at zero, and

$\xi \sim N(0, \sigma_\xi)$. The logic for this arrangement is that for the the latent scale to be identified in terms of its mean and variance, it is necessary to impose restrictions on the cutpoints γ_{i1} and γ_{i2} that link the latent scale to the observable responses. At the same time, there is reason to believe (and the data ultimately confirm) that different respondents use the neutral category much more or less frequently than others, in ways that are not consistent with variation in ψ. The prior for ξ implies that the median respondent is is assumed to have cutpoints of $-\frac{1}{2}$ and $\frac{1}{2}$, which fixes both the centre and variance of the latent scale, but through their response patterns individual respondents may reveal themselves to have wider (positive ξ) or narrower (negative ξ) neutral categories than the average respondent.

Estimation details for the model are provided in the appendix, including code for specifying the model in Stan (Carpenter et al., 2017) and details on estimation run times. As noted previously, there are two main types of applications of the model specified above. Different quantities from the model are the primary quantities of interest in these different applications.

Quantities of Interest for Assessments as Dependent Variables

Where problem assessments are the *dependent variable*, the primary quantities of interest in the model are to be found in α, β, and Σ. The n_j α parameters describe the population-level average relative assessments of the of the alternatives j. These summarise the overall propensity of respondents to select different alternatives (the categories of political problems) relative to one another. These are analogous to the parameters in a standard Bradley-Terry model.

The β and Σ parameters describe how evaluations of the alternatives j vary across individuals i around the averages defined by α. The $n_j \times n_k$ matrix of coefficients β describes variation in individual-level latent scores as a function of the n_k predictive individual-level variables in X_i. These coefficients are interpretable in the familiar terms of a regression model for the latent variable ψ_{ij}, which is individual respondent i's latent rating of problem category j. Each variable k in X_i could be observational, it could be a randomly assigned treatment, or it could be variation that can be argued to be quasi-

the response model simplifies to the single equation $p\left(Y_{ijj'} = 1\right) = \Phi\left(\frac{\delta + \left(\psi_{ij'} - \psi_{ij}\right)}{\omega}\right)$

experimental. The usual differences in causal interpretation of the coefficients apply, depending on how the X_i are assigned to individuals i.

In addition to the βs, there is further information to be gained by examining the Σ matrix which describes the covariance of individual-level assessments of different problems. This is particularly true if the model is fit *without* X_i, such that this estimated covariance matrix fully captures the covariation in respondents' latent scale evaluations of the categories. The diagonal elements of this matrix σ_j^2 provide information about the relative degree of variation across individuals in evaluations of different categories. For example, it could be the case that one problem category is highly variable in how many opposing categories it would be selected over by different respondents, while other problem categories would be selected against nearly the same opposing categories by everyone. If this were the case, the diagonal elements σ_j^2 for the former would be larger in magnitude than for the latter.

The off-diagonal elements of Σ reveal information about which problems tend to be evaluated more or less highly by the same people. For example, if people who are relatively concerned about Health j also tend to be relative concerned about Education j', these categories would have a positive $\rho_{jj'}$. If those who are more concerned about Health j tend to be less concerned about Immigration j', then these would have a negative $\rho_{jj'}$. Note, however, that if the model is fit with X_i, $\rho_{jj'}$ only contains the *residual* covariance excluding that captured by the relative values of β_j and $\beta_{j'}$.

Quantities of Interest for Assessments as Independent Variables

Where problem assessments are not the outcome variable of interest, but are an *independent variable* to be used in predicting another variable, the model described can be used as a measurement model for an individual's latent rating of each problem ψ_{ij} or as an imputation model for that individual's unobserved pairwise comparisons $Y_{ijj'}$. As noted in the introduction, realistic amounts of pairwise comparison data will yield only limited information about ψ_{ij} for each respondent, and thus estimates of ψ_{ij} and imputations of $Y_{ijj'}$ will be subject to substantial measurement error. Fortunately, this mea-

surement error arises primarily from unobserved pairwise comparison data $Y_{ijj'}$ which are *missing completely at random* (MCAR). The pattern of missingness is independent both of observable variables and of unobservable quantities of interest, arising as it does from randomly sampling pairs for each respondent. The regression model described above can straightforwardly be applied as an imputation model for any unobserved $Y_{ijj'}$ for the respondents i in the data set, by simulating responses from the predicted probabilities for those observations. Rubin (1987) describes rules for propagating uncertainty when using multiply imputed data sets which apply straightforwardly.

While the same regression model described above can be used for both the case where the assessments are the dependent variable of interest and also for these measurement and imputation applications where the assessments form the basis for an independent variable, the selection of X for these different types of models must reflect different considerations. If the goal is to study the assessments underlying the pairwise comparisons as an outcome variable, one should be operating from the perspective of regression specifications that will provide useful descriptive inferences and interpretability, or in the case where there is an experimental or quasi-experimental intervention, valid causal inference. However, when the goal is measurement and imputation: the aim is maximising predictive power, model flexibility, and ensuring *proper* imputation with respect to a subsequent analysis (Rubin, 1996, p478-479; Murray, 2018, p146-147).

A requirement of achieving proper imputation with respect to a subsequent analysis is the use of a *congenial* imputation model that captures all relevant conditional relationships involving the imputed variables and other variables in the subsequent analysis. In particular, this means the inclusion of the outcome variable(s) of that subsequent analysis in the imputation model. This means that the best imputation model, given data constraints, is likely to be application specific. "Unless the analyst is the imputer, congeniality is less a condition we should try to satisfy than one we should try to fail gracefully – uncongeniality is generally 'the rule not the exception' (Xie and Meng, 2017)" (Murray, 2018, p146). However, in practice it is still often possible to provide useful but improper imputations for a wide range of users (Rubin, 1996, p479; Murray, 2018, p147). In my applications below, I illustrate

both the use of a imputation model that omits the outcome variable of ultimate interest—simulating what might happen if an off-the-shelf set of imputed values were provided for subsequent analysis—as well as the use of an imputation model that includes the outcome variable of ultimate interest.

Describing Variation in UK Political Problem Assessments

The dataset for this study consists of nine pairwise comparison questions fielded to 2065 UK respondents by YouGov on 9-10 November 2025. This was delivered as part of YouGov's regular political omnibus survey, alongside current UK parliamentary vote intention and other questions. For this study, I have access to measures of standard demographics, 2024 UK general election vote, and self-reported attention to politics that YouGov holds on these panellists from previous surveys, plus a prospective vote intention question asked on the same survey as the pairwise comparisons.

Each respondent first received eight pairwise comparisons sampled independently with replacement from the 420 ordered pairs possible with 21 categories. The ninth question received by all respondents is the specific pairing of "The Economy" vs "Immigration", using the same format as in the randomly generated pairs.

I begin by estimating the model described earlier, using the eight randomly generated pairwise problems, omitting covariates X (Model 1). As shown in Figure 1, the coefficients α are very precisely estimated with this amount of data. These are the traditional quantities of interest from a Bradley-Terry model, that correspond to the population averages on the latent scale from this model. In these data, the areas of greatest concern are Health and The Economy, followed by Immigration, Law and Crime, Social Welfare, and Housing. Most of these differences are statistically significant because, in terms of the predicted probabilities for specific pairwise responses, the data indicate very large differences in the probabilities of selecting the alternatives. For example, if we compare Health to Education, examining only the limited number of direct comparisons in the raw data, we see that of all individuals who received this particular pairwise comparison, 37 (59%) chose Health, 25 (40%) chose the neutral option, and just 1 (2%) chose Education.

Table 1: Estimates of average latent rating (α_j), standard deviation of latent ratings (σ_j), and estimated proportion of population rating rating each problem most important (MIP_j), with central 95% intervals.

	α_j		σ_j		MIP_j	
Health	0.94	(0.84, 1.03)	0.54	(0.41, 0.68)	0.17	(0.11, 0.23)
The Economy	0.89	(0.81, 0.98)	0.52	(0.38, 0.65)	0.13	(0.08, 0.19)
Immigration	0.66	(0.55, 0.79)	1.33	(1.17, 1.50)	0.27	(0.23, 0.31)
Law and Crime	0.56	(0.48, 0.64)	0.69	(0.55, 0.82)	0.05	(0.02, 0.09)
Social Welfare	0.47	(0.39, 0.54)	0.66	(0.51, 0.80)	0.05	(0.02, 0.09)
Housing	0.42	(0.34, 0.50)	0.64	(0.50, 0.78)	0.05	(0.02, 0.08)
Energy	0.24	(0.16, 0.31)	0.57	(0.43, 0.72)	0.02	(0.01, 0.05)
Government Operations	0.18	(0.10, 0.25)	0.71	(0.56, 0.84)	0.03	(0.01, 0.06)
Work and Employment	0.14	(0.07, 0.21)	0.67	(0.53, 0.80)	0.02	(0.01, 0.05)
Environment	0.08	(0.00, 0.17)	0.83	(0.68, 0.97)	0.06	(0.03, 0.09)
Education	0.05	(-0.02, 0.12)	0.71	(0.58, 0.86)	0.03	(0.01, 0.04)
Agriculture	-0.02	(-0.09, 0.06)	0.72	(0.58, 0.87)	0.02	(0.01, 0.03)
Rights and Freedoms	-0.15	(-0.23,-0.06)	0.95	(0.79, 1.12)	0.04	(0.02, 0.06)
Defence	-0.24	(-0.32,-0.16)	1.00	(0.84, 1.14)	0.02	(0.01, 0.04)
Foreign Trade	-0.30	(-0.37,-0.24)	0.50	(0.33, 0.66)	0.00	(0.00, 0.00)
International Affairs	-0.37	(-0.45,-0.30)	0.72	(0.58, 0.86)	0.01	(0.00, 0.01)
Transportation	-0.39	(-0.47,-0.31)	0.80	(0.66, 0.96)	0.01	(0.00, 0.02)
Regulation	-0.69	(-0.77,-0.61)	0.64	(0.50, 0.77)	0.00	(0.00, 0.00)
Technology	-0.69	(-0.78,-0.60)	0.75	(0.61, 0.90)	0.00	(0.00, 0.01)
Public Land	-0.88	(-0.98,-0.79)	0.77	(0.62, 0.92)	0.00	(0.00, 0.01)
Culture	-0.90	(-1.00,-0.80)	0.95	(0.79, 1.10)	0.01	(0.00, 0.01)

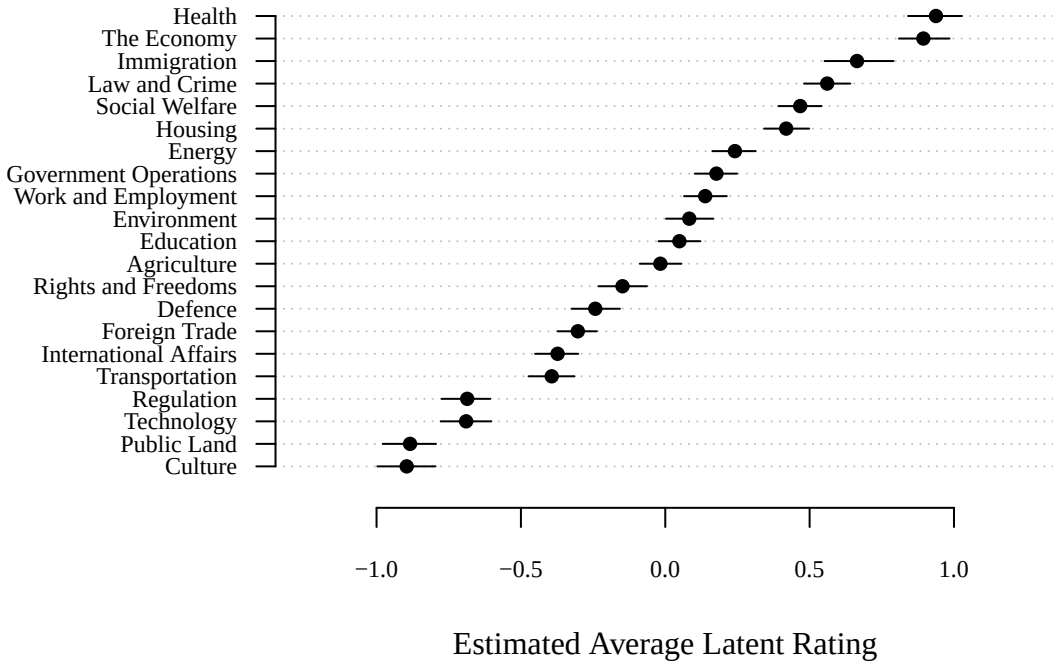

Figure 1: Estimated average latent ratings for the 21 problem categories across respondents, with 95% posterior intervals (Model 1).

The model estimates allow us to say something about variation at the individual level as well as the population averages. In Table 1, I show the estimated α_j from Figure 1 along with their 95% interval estimates. In addition, I show σ_j, the square root of the diagonal elements of Σ, which provide the estimated individual level variation around that population average expressed as a standard deviation on the same scale as α. These vary substantially across different problems, with the largest σ_j found on Immigration. One consequence of this high variation in where respondents rank Immigration is that, even though Immigration has only the third highest average rating α_j after Health and The Economy, the model estimates that 27% of respondents would Immigration above all other problems, which is higher than for either the Health (17%) or The Economy (13%) categories (a statistically significant difference).

The other piece of information in Σ is the extent to which individual ratings of different problems are correlated. Figure 2 shows the estimated pairwise correlations $\rho_{jj'}$ implied by Σ. There is clear evidence in the data that assessments of the different problem categories are correlated, and that they are correlated in a way that reflects contemporary UK politics. For example, the strongest positive

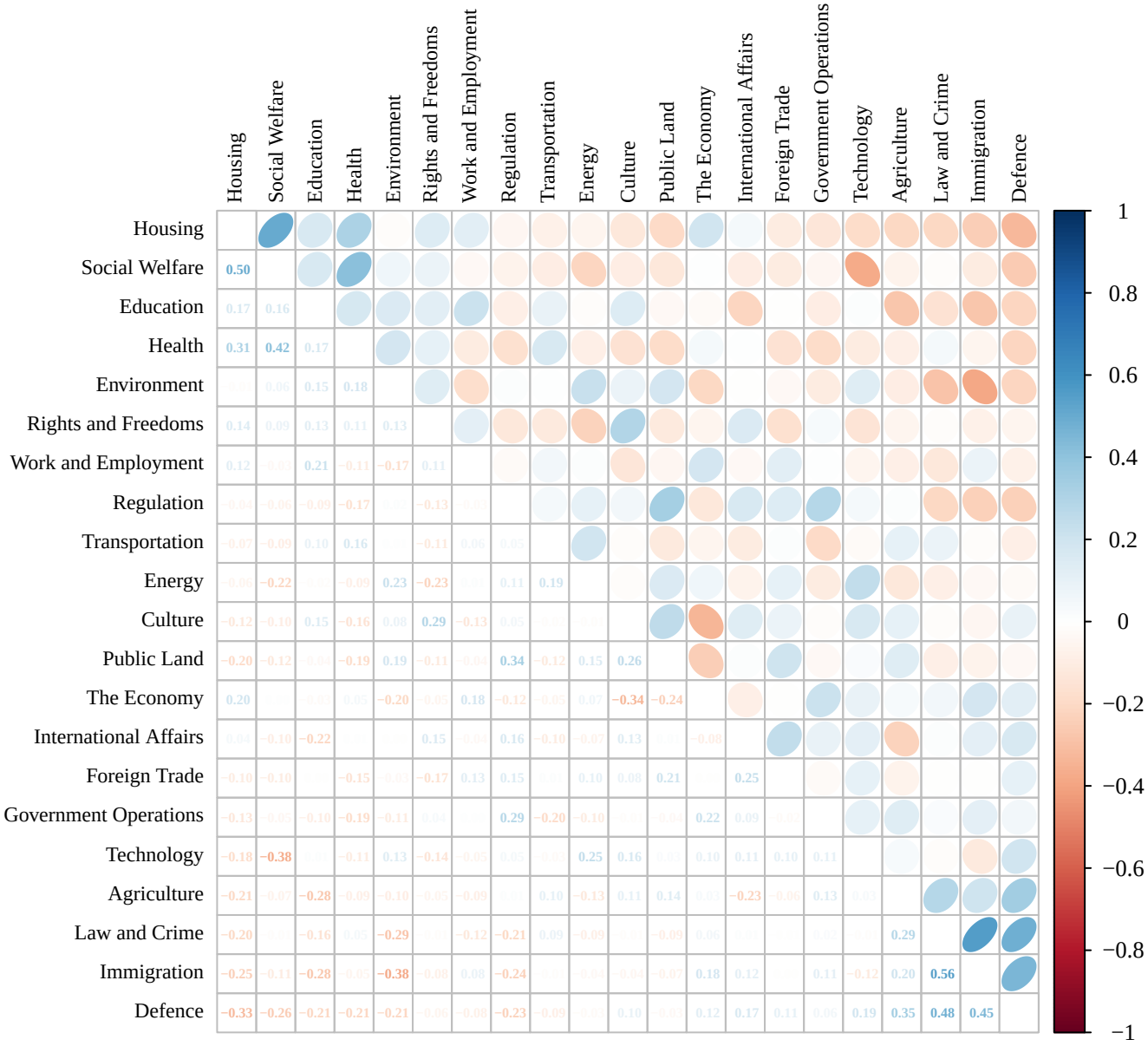

Figure 2: Estimated correlation in latent rating variation in the 21 problem categories across respondents. Categories ordered by first principle component.

correlation in the data (0.56) is between the tendency to view Law and Crime and Immigration as more important problems, with other strong positive correlations between these and Defence and Agriculture. The other strong cluster of positive correlations are between assessments of Housing, Health, and Social Welfare. The former group are concerns typically associated with the political right, the latter group are concerns typically associated with the political left. One of the strongest negative correlations is between Immigration and Environment (-0.38).

Figure 3 attempts to illustrate how individual-level deviations from the population average arise from an individual's raw responses, again under the model without covariates (Model 1). This figure shows two individuals, both of whom are white men in England, with high incomes and who own their homes. Respondent A (left) had three randomly generated comparisons involving Immigration, and all three times they chose the other option (Social Welfare, Environment, the Economy). As a result, the estimated value of $\psi_{\text{Respondent A,Immigration}}$ is substantially below the population average value of $\alpha_{\text{Immigration}}$. The model uses broader correlation patterns in the data, so this respondent's ratings for categories like Defence and Law and Crime—which this respondent never directly responded to—are estimated to be below the population mean as well because they are positively correlated with views on Immigration across respondents. Most of the respondent's other choices track the typical patterns in the data, and so their estimated ψ largely track α. In contrast, Respondent B (right) compared Immigration with Health and with The Economy, choosing Immigration both times. Immigration was consequently estimated as their highest rated category. They gave a neutral response on Defence versus Health, and selected Defense over Agriculture, boosting Defence up their personal ratings relative to typical respondents.

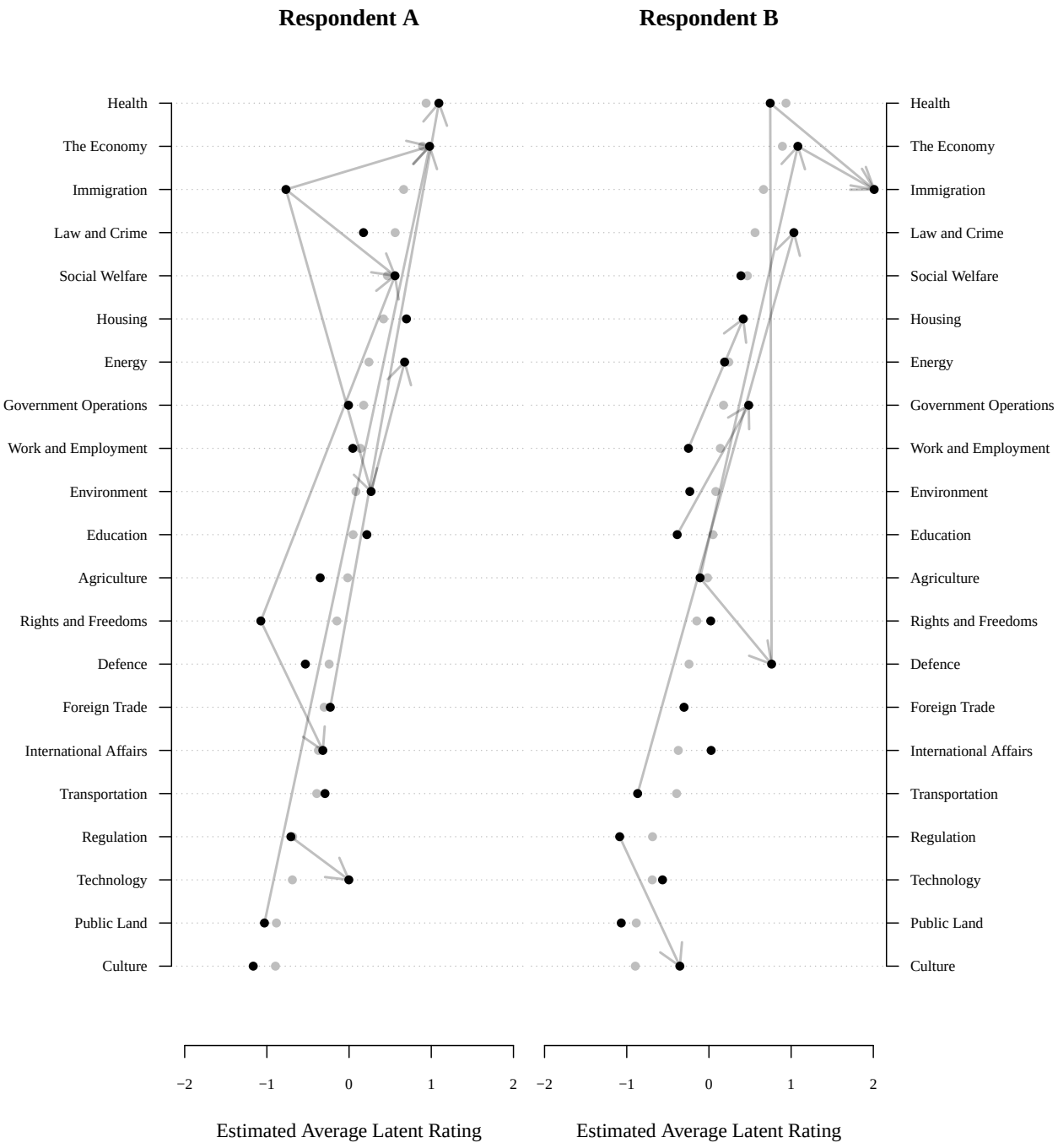

Figure 3: Two example respondent's estimated latent ratings (black dots), relative to estimated population distribution (grey points). Arrows depict the respondent's choices between the randomly selected pairings that they received, with the arrow head pointing towards the one they selected. Neutral responses are depicted without arrow heads.

Table 2: Coefficients predicting individual level latent ratings of problem categories from standardized predictors. Bold cells have central 95% intervals that exclude 0. Omitted categories are 35-49 years old, male, no university degree, living in England outside of London, not a home owner, medium or refused income, did not vote in 2024.

	Intercept	18-34	50-64	65+	Female	Degree	London	Scotland	Wales	Owner	Low Inc	High Inc	Attention	Con 24	Lab 24	Ref 24	LD 24	Other 24
Health	**0.95**	-0.01	-0.03	0.01	**0.08**	0.03	0.00	0.01	0.00	-0.01	-0.01	0.01	-0.03	-0.05	0.01	-0.07	0.00	0.05
The Economy	**0.91**	0.06	-0.03	-0.02	-0.05	0.03	0.04	0.01	0.00	0.04	-0.01	0.04	-0.01	0.05	-0.03	0.01	0.00	0.01
Immigration	**0.67**	-0.05	0.01	0.02	0.02	**-0.08**	0.01	0.01	-0.01	0.03	0.00	-0.03	**-0.21**	**0.29**	-0.09	**0.38**	-0.02	**-0.10**
Law and Crime	**0.57**	-0.04	0.03	0.01	0.03	0.01	-0.01	0.01	0.00	0.02	0.00	0.00	-0.05	0.07	0.00	**0.14**	-0.01	-0.03
Social Welfare	**0.48**	-0.05	0.04	0.01	**0.07**	0.00	0.00	0.00	0.00	-0.03	0.01	0.01	-0.02	**-0.14**	0.03	**-0.11**	0.01	0.02
Housing	**0.42**	0.01	0.01	-0.03	**0.07**	-0.03	0.02	0.00	-0.01	**-0.18**	-0.01	0.01	**0.07**	**-0.08**	0.06	**-0.10**	0.01	-0.01
Energy	**0.24**	-0.02	0.00	0.00	**-0.09**	0.00	0.02	0.00	0.00	0.06	0.01	0.00	0.03	-0.04	0.02	-0.06	0.00	0.01
Government Operations	**0.17**	0.04	-0.06	-0.02	-0.04	0.06	0.02	0.00	0.00	0.05	0.00	0.00	**-0.08**	**0.12**	-0.01	**0.10**	0.00	-0.04
Work and Employment	**0.14**	0.05	-0.02	-0.04	0.01	-0.02	0.06	0.01	0.00	-0.06	0.00	-0.01	-0.01	**-0.11**	-0.02	**-0.11**	0.00	-0.04
Environment	**0.08**	0.01	0.00	0.01	0.06	0.04	-0.03	0.00	0.01	0.00	-0.01	0.01	0.00	**-0.14**	0.01	**-0.24**	0.01	**0.10**
Education	0.05	-0.01	0.00	-0.02	0.03	0.04	-0.01	-0.01	0.00	-0.02	-0.01	0.00	-0.01	-0.03	0.04	**-0.13**	0.01	0.02
Agriculture	-0.02	-0.05	-0.01	0.03	0.04	-0.05	-0.05	-0.02	0.00	0.05	0.02	-0.02	0.01	0.06	-0.02	**0.08**	0.01	-0.03
Rights and Freedoms	**-0.14**	0.00	-0.01	-0.02	0.03	0.00	0.04	0.01	0.00	**-0.11**	0.00	-0.03	**0.08**	-0.04	0.00	**0.08**	0.01	0.01
Defence	**-0.24**	**-0.09**	0.05	0.08	-0.02	-0.03	**-0.08**	0.01	0.01	0.05	0.00	-0.01	0.03	**0.20**	0.00	**0.21**	0.00	-0.07
Foreign Trade	**-0.30**	0.02	0.00	0.01	-0.04	0.00	0.04	-0.01	0.01	0.04	-0.01	0.02	-0.01	0.00	0.04	-0.05	-0.01	-0.01
International Affairs	**-0.38**	0.02	-0.03	-0.02	0.05	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.05	0.00	-0.02	0.01	0.00	0.01
Transportation	**-0.40**	0.04	0.01	-0.01	-0.03	0.01	**-0.06**	0.00	0.00	-0.05	0.00	0.01	-0.06	-0.06	0.01	-0.07	0.00	-0.03
Regulation	**-0.69**	0.03	0.01	-0.01	-0.07	-0.03	0.04	-0.02	0.00	0.03	0.00	0.00	-0.01	**-0.09**	-0.01	0.00	0.00	0.03
Technology	**-0.70**	-0.01	0.01	-0.01	**-0.11**	0.05	-0.01	0.00	-0.01	**0.08**	0.00	0.01	0.06	0.01	0.00	-0.05	0.00	0.01
Public Land	**-0.90**	0.01	0.02	-0.01	0.02	-0.03	-0.04	-0.01	0.01	0.02	0.01	-0.03	-0.01	-0.05	0.02	-0.04	-0.02	0.05
Culture	**-0.91**	0.03	-0.03	0.01	-0.05	0.00	-0.01	0.00	0.00	-0.03	0.00	0.01	**0.17**	0.02	-0.02	0.03	0.01	0.03

Respondent A and Respondent B are demographically similar in some respects, but not all. Respondent A is aged 49, has a university degree, and voted Labour in 2024. Respondent B is aged 67, does not have a university degree and voted Conservative in 2024. The Model 1 estimates presented thus far do not use this information, but the apparent political alignment of these views with those respondent-level characteristics does suggest that further gains may be possible once we add covariates to the model. Accordingly, I now introduce Model 2, which adds demographic variables (age, gender, degree, home ownership, income, region) and political variables (self-reported attention to politics, 2024 general election vote). All of the variables in the model (including binary variables) are standardized to have mean zero and standard deviation one, so that the magnitudes of the coefficients are indicative of relative predictive power.

Predictors of Individual-Level Problem Assessments

Table 2 shows the estimated β coefficients from Model 2 predicting latent scale variation across respondents as a function of demographic and political variables. There are more significant coefficients (bold in the table) for the past vote variables than for most of the others. Consistent with the previous finding that assessments are more varied on Immigration than any other problem, assessments of Immigration have especially widely varying coefficients by 2024 vote choice. 2024 Conservative and Reform voters rate Immigration as a much more important problem, and Labour and Other[6] voters as a less important problem, relative to the baseline non-voters. Conservative and Reform voters rate the Environment and Social Welfare lower, Defence and Government Operations higher.[7] Labour voters are more concerned about Education and Housing than non-voters.

The demographic predictors are predictive for some specific problems. The most distinctive view of people of people who own their homes, relative to those who do not, is to rate Housing as a lesser problem. Londoners rate problems in Transportation lower than those in the rest of the country.

[6]The Other category here is predominantly a mixture of largely left-leaning voters for the Greens, the SNP, and Plaid Cymru, as well as assorted other independent and minor party candidates.

[7]Negative views of Government Operations by Conservative and Reform voters may reflect the fact that there was a Labour government at the time of survey or a more general anti-government disposition.

Women rate problems in Health, Social Welfare, Housing higher than men, and Energy, Regulation and Technology lower. Older voters rate problems in Defence higher than younger voters. Degree holders rate problems in Government Operations and Technology higher and Immigration lower. High income voters rate problems in the Economy higher than do low income voters. More politically attentive voters rate problems in Housing, Rights and Freedoms, and Culture higher, and those in Immigration and Government Operations lower. All of these are holding fixed all other variables in the model including 2024 vote choice, indicating within voting coalition variation. A demographic only model would exhibit more and stronger relationships among these variables, as demographics themselves strongly predict 2024 vote choice. For some applications it might make sense to not condition on past vote choice, it depends on the purpose of the analysis. Here, where the purpose is illustrative, I have included a relatively large set of variables. Whichever variables are included, exactly the same high level of caution associated with typical regression analyses on survey data should be taken before making any causal claims.

I will return later to the question of how much data is necessary to successfully conduct this kind of analysis, where the aim is primarily to understand variation in the underlying assessments. The main point to take from the above analysis is that the estimates recovered are easily interpretable, discover a range of interesting and plausible partial associations, and could be easily replicated at other times or in other places to make comparisons.

Measuring Individual Political Problem Assessments

We now turn to the question of whether the individual-level estimates from the model constitute useful *measurements* enabling analysis of how variation in these predicts other survey responses or behaviours. Beyond the face validity check from examining respondent profiles in Figure 3 above, there are several ways to ask this question about the usefulness of the model estimates. First, I assess what the model estimates imply about how precisely the ψ_{ij} are estimated. Second, I conduct a cross-validation exercise where I assess out-of-sample predication of all of the pairwise comparisons $Y_{ijj'}$

Table 3: Estimated R^2 of mean posterior estimates of individual-level problem category ratings relative to true individual-level variation across individuals of these ratings, by problem categories. Model 1 uses no covariates, Model 2 uses demographic and political variables.

	Model 1	Model 2
Health	0.15	0.20
The Economy	0.14	0.19
Immigration	0.36	0.45
Law and Crime	0.26	0.26
Social Welfare	0.21	0.25
Housing	0.22	0.33
Energy	0.15	0.19
Government Operations	0.17	0.22
Work and Employment	0.17	0.24
Environment	0.24	0.31
Education	0.21	0.24
Agriculture	0.21	0.24
Rights and Freedoms	0.24	0.26
Defence	0.32	0.39
Foreign Trade	0.11	0.15
International Affairs	0.19	0.20
Transportation	0.19	0.22
Regulation	0.18	0.21
Technology	0.19	0.23
Public Land	0.19	0.20
Culture	0.22	0.25

using a series of model fits that exclude all subsets of the data in turn. Third, I use the single pairwise comparison that I asked of all respondents at the end of the study, The Economy versus Immigration, to test out-of-sample prediction of a specific pairwise comparison. Finally, having completed these forms of validation of the key elements of the model as a basis fo imputing pairwise comparisons $Y_{ijj'}$, I illustrate an analysis of vote intention as a function of these imputations and validate this against an observed benchmark.

The model implicitly estimates how much of the true variation–the variation we would esti-

mate using this model if we were able to ask a very large number of pairwise comparisons to these individuals–is captured by the mean posterior estimates of the respondent-level ratings on each domain. We can calculate a R^2-like statistic describing the variance across individuals i on each domain j of the (mean posterior) point estimates $\hat{\psi}_j$ relative to the estimated variation in ψ across individuals i in that domain j:

$$\widehat{R^2}\left(\widehat{\psi_j}, \psi_j\right) = \frac{\text{Var}_i\left[\widehat{\psi_j}\right]}{\text{Var}_i\left[\widehat{\psi_j}\right] + \text{Mean}_i\left[\widehat{Var}\left[\psi_j\right]\right]} \tag{1}$$

This statistic is useful here, as well as later, for assessing the extent to which there are sufficient data to recover variation in the underlying quantities of interest given the amount of estimation uncertainty. For example, a $\widehat{R^2}\left(\widehat{\psi_j}, \psi_j\right) = 0.25$ implies a correlation $r = 0.5$ between the point estimates and the true values of ψ.

Table 3 shows the estimates of these quantities from Models 1 and 2. By issue, these vary from $R^2 = 0.11$ to $R^2 = 0.36$ under the model without covariates, and from $R^2 = 0.15$ to $R^2 = 0.45$ under the model with demographic and political covariates. The average R^2s are 0.21 for Model 1 and 0.25 for Model 2. Note that we are able to better measure ψ for categories which are better predicted by covariates and correlation with other categories, such as Immigration. While Table 3 implies significant measurement error is present in the estimated $\hat{\psi}_{ij}$ relative to the true ψ_{ij}, as noted previously and reinforced by the high correlation between the estimates across the two models,[8] this measurement error arises primarily from a controlled random process of sampling problem pairs for each respondent.

Cross and External Validation of Pairwise Comparisons

In order to facilitate multiple imputation, the model needs to be able to predict pairwise comparisons $Y_{ijj'}$ of problems j and j' for each respondent i. In order to confirm that the model is able to reliably

[8]The individual-level estimates of ψ recovered from Model 1 and Model 2 are correlated at 0.82 to 0.96, with an average of 0.89 across problem domains. The covariates are providing additional information, but they do not radically change the individual-level estimates of the latent ratings, as these are first and foremost shaped by the questions that the respondents answered.

predict out-of-sample—for pairs that a given respondent was not asked about—I conduct two kinds of validation of the pairwise comparison predictions. First, I conduct a cross-validation exercise in which I re-estimate the model 8 times, each time leaving out one round of response data, constructing the predicted probabilities for those omitted pairwise comparisons for the individuals who answered them. We can then examine the relationship between the predicted probabilities and the observed choices to assess whether the model provides well-calibrated predictions.

The top row of panels in Figure 4 uses a spline regression to assess how the frequency of choosing each option varies as a function of the predicted probability for choosing that option. While the model is not perfectly calibrated, the deviations are substantively small with observed probabilities tracking predicted probabilities from the model for held out data in cross-validation. The appendix provides these disaggregated by problem category, which shows that predictions are similarly well calibrated for all categories.

Second, the inclusion of the ninth pairwise comparison on The Economy versus Immigration for all respondents means that we can do an external validation exercise of predicting that specific pair using the model estimated on the preceding eight items.[9] The bottom row of panels in Figure 4 shows that while the out-of-sample predictions for the Economy versus Immigration pair are not perfectly calibrated, they track the predicted probability closely across the full range of predicted probabilities for all three response categories. The deviations are not substantively large and the predicted probabilities are highly informative regarding the unobserved pairwise comparison. Taken together, the cross-validation and external validation evidence illustrate that we can reliably generate predictions of the variation in unobserved pairwise comparison responses.

[9] Of the 95 respondents who had previously seen The Economy versus Immigration, in either order, when asked again 83 gave the same response, 9 switched between the neutral and a non-neutral category, and 3 switched between opposing substantive responses.

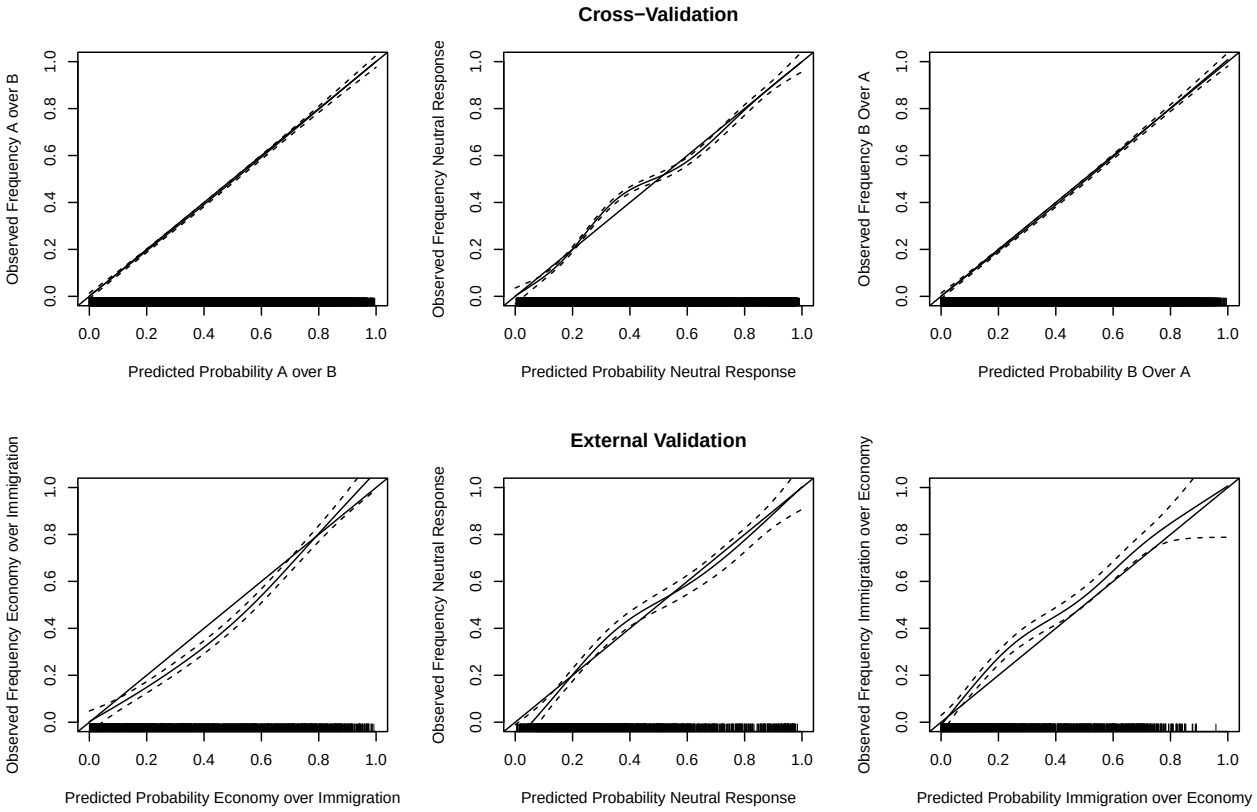

Figure 4: Top row: performance in out of sample prediction of held out responses, in 8-fold cross-validation. Bottom row: performance of out of sample predication of responses to Economy versus Immigration comparison completed by every respondent at the end of the survey.

Using Imputed Pairwise Comparisons

If we can reliably predict comparisons not used in the estimation of the model, we have a sound basis for using the model for missing data imputation of pairs not observed for a given individual, which can then be used in subsequent analysis. I focus on an example using the Economy versus Immigration pairing where we can validate analyses using the imputed pairwise comparisons to the gold standard analysis using the directly asked pairwise comparison. Figure 5 shows estimated vote intention, as of the time of survey, as a function of imputed and observed responses to the Economy versus Immigration pairwise comparison. The analyses using imputed data apply Rubin's rules to propagate imputation uncertainty. For this analysis, I add a new Model 3, which starts with Model 2 and adds the current vote intention variable to the model for the pairwise comparisons. This variable

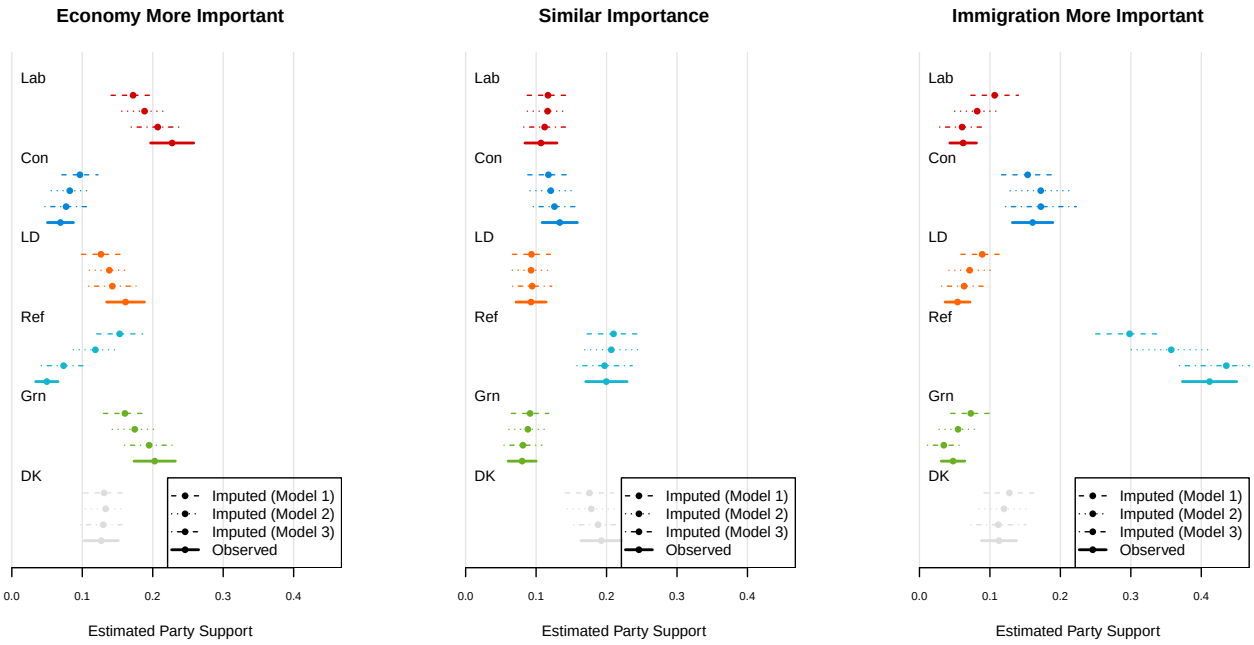

Figure 5: Estimates of support for parties at time of survey experiment, with 95% intervals, as a function of observed versus imputed assessments of the economy versus immigration as more important problem.

is the dependent variable for this application, and so forming a proper imputation model requires its inclusion for the reasons described earlier.

For Model 1 and Model 2, using the imputed data recovers the qualitative patterns found in the observed data regarding which parties have more versus less support, but with attenuation of the differences in vote intention as a function of the Economy versus Immigration response. This is especially clear in support for Reform, but a similar pattern holds for Labour, the Liberal Democrats, and the Greens. The degree of attenuation is greatest in the no covariate model which uses the least information in the imputation (Model 1), whereas the model with both demographic and political predictors (Model 2) comes closer to the estimates based on the observed data. Model 2 is the type of imputation model whose outputs might be provided with survey data for general use, including a range of relevant predictors but not the outcome variables of subsequent analyses.

Once we add the outcome variable for this analysis to the imputation model, we see that the analysis using imputations from Model 3 comes very close to recovering both the point estimates

from the observed data and also the relevant differences. For example, the difference in support for Labour between those who think the Economy is a more important problem and those who think Immigration is a more important problem is 16.6 percentage points based on the observed responses to that question. The estimated difference is smaller based on imputed responses under Model 1 (6.5) and under Model 2 (10.7), but very close to the value based on the observed data if we use imputed data from Model 3 (14.7). The differences between the proportion intending to vote Reform as a function of this Economy versus Immigration question are even larger. The corresponding differences are -36.2 percentage points (observed), -14.5 (Model 1), -23.9 (Model 2), and -36.2 (Model 3), again showing the attenuation bias dissipating as the imputation model is improved. Similar patterns hold for all of the other parties, with the more informative imputation models yielding estimates closer to the observed data benchmark. The uncertainty estimates associated with the analyses using imputation are larger than using the observed data, but as is typical for multiple imputation applications where the imputed variable can be predicted reasonably well, the additional uncertainty is not large.

The fact that this analysis strategy works is powerful. We could run this analysis for *any* of the 210 pairwise comparisons of categories. It may be that some other problem domains become more important to UK political alignments over the coming years, in which case having historical data from which analysis might be conducted for any pairing would become valuable as a reference to assess whether this reflected a change or whether it had always been true.

Beyond the Most Important Problem

The pairwise comparison methods illustrated above are general to the wide variety of applications where researchers are interested in studying individual-level variation in judgements/assessments/evaluations. The appendix for this paper provides a discussion of practical design considerations related to implementing pairwise comparison designs like the one considered here, for the study of problem assessment or for other applications. This includes considerations related to (a) the use of forced choice versus neutral response question formats, (b) different approaches to sampling pairs

for each respondent, (c) the number and selection of objects/categories to be assessed, (d) changing objects/categories over time given repeated use of this kind of instrument, and (e) the quantity of data required to recover reliable estimates, both in terms of the number of respondents and the number of pairs per respondent. This includes simulation studies on subsets of the data, which show that the estimates still perform reasonably well with half or even a quarter of the data collected in this instance. This is true both for reducing the number of respondents (e.g. from about 2000 to 1000 or 500) or the number of pairs (from 8 to 4 or 2), although the ability of the model to recover variation across respondents is reduced.

For the specific application of measuring citizen attitudes towards political problems, the viability of the approach gains importance from the well-known deficiencies of open-ended MIP/MII questions, which undermine the usual arguments against changing long-running survey questions. The methods developed in this paper show that we can use closed-ended, pairwise comparison questions which are easier and faster for respondents to answer than either open-ended questions or closed-ended ranking or single-rating questions. Unlike these formats, the responses to pairwise comparisons are more straightforwardly comparable interpersonally, internationally, and intertemporally. We can generate measurements on a set of researcher-defined categories. We can ask a small number of simple, clearly stated questions for each individual, yet generate data on a broader range of categories than we would want to ask any individual to respond regarding. We can still do all the kinds of complete data analyses we would be able to do if we asked everyone the same questions. There is a cost of some measurement precision, but this is a small price to pay if it opens up the possibility of genuinely comparable measurement of problem perceptions across individuals, across country, and across time. If a significant number of different national election studies started asking questions like these, ideally on the same categories,[10] this would quickly generate an incredibly valuable resource for studying the concerns that citizens have and want their governments to address. It is a shame that

[10] Agreeing common prompts and problem categories is a challenge on par with existing coordination of questions across national election studies. As discussed in the appendix, partially overlapping sets of categories will yield usable data for purposes of intertemporal and international comparisons.

we do not have that record historically, but that is no argument against starting to collect it now.

The approach described in this paper requires making a measurement choice that is atypical in election studies, which seldom include survey experiments or randomised choices of questions in their core question sets. Surveys experiments tend to answer one research question, whereas election studies are collective scholarly resources aiming to produce data that facilitates the answering of many questions. This is straightforwardly possible if every respondent answers the same prompts. The point of this paper is to illustrate that we can also achieve this aim using survey measurement instruments involving pairwise comparisons selected randomly from larger pools of objects of evaluation.[11]

Some scholars might be uneasy about relying on imputed quantities, as opposed to working with raw survey responses directly. This is an uneasiness with a pedigree: early criticisms of multiple imputation objected (1) to the use of simulation, (2) to the extra work required for the data user, (3) to the storage requirements of multiply imputed data sets, and (4) to the effort and challenge of creating (approximately) proper imputations (Rubin, 1996, p479-480). Over the past three decades, the widening use of simulation-based inference, improved statistical software, and the advance of computing technology have largely addressed the first three of these concerns. It would not be difficult for an election study to provide multiply imputed pairwise comparison data for download as supplementary data files, alongside original raw data. It is the concerns about the requirement to model the imputed variables *well* that give the most reason for pause.

The validation exercises described above ought to be reassuring that it is feasible to do the imputation well-enough to get the right answers in subsequent analyses. There are theoretical grounds to expect imputation methods to work well for this application. Missingness by design makes this a more statistically reliable application of multiple imputation than the more typical cases where the missingness mechanism is out of the researcher's control. One still needs a suitably predictive model,

[11]This principle also applies to non-pairwise formats of object single ratings or of item batteries. One can randomly sample a subset of questions from the full pool per respondent and then impute the missing categories using a model similar to the one used here, or a model with stronger assumptions about the covariance of items such as a low-dimensional IRT model. This enables full sample analysis on a larger set of questions than any one respondent could be asked about.

but the missingness is independent of any quantities of interest by design. The need for a good predictive model asks little more of the analyst than is required to use open-ended MIP/MII questions, which already rely on a coding or modelling exercise to translate open-ended responses into categories. Such text coding exercises are more opaque and difficult to reproduce than what is proposed here, they are however more familiar. Randomly asking different questions of different respondents within the sample and then using that data to make inferences about how all the respondents in the sample would have answered all of the questions involves unheroic assumptions by comparison, albeit ones we are less acclimatised to making.

References

Bevan, Shaun. 2025. "Comparative Agendas Project: Comparing Policies Worldwide.".

Blumenau, Jack and Benjamin E Lauderdale. 2024. "The variable persuasiveness of political rhetoric." *American Journal of Political Science* 68(1):255–270.

Bradley, Ralph Allan and Milton E Terry. 1952. "Rank analysis of incomplete block designs: I. The method of paired comparisons." *Biometrika* 39(3/4):324–345.

Bramley, Tom and Sylvia Vitello. 2019. "The effect of adaptivity on the reliability coefficient in adaptive comparative judgement." *Assessment in Education: Principles, Policy & Practice* 26(1):43–58.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan: A probabilistic programming language." *Journal of statistical software* 76(1).

Enns, Peter K. 2014. "The public's increasing punitiveness and its influence on mass incarceration in the United States." *American Journal of Political Science* 58(4):857–872.

Fournier, Patrick, André Blais, Richard Nadeau, Elisabeth Gidengil and Neil Nevitte. 2003. "Issue importance and performance voting." *Political Behavior* 25(1):51–67.

Green, Jane and Sara B Hobolt. 2008. "Owning the issue agenda: Party strategies and vote choices in British elections." *Electoral Studies* 27(3):460–476.

Hamilton, Ian, Nick Tawn and David Firth. forthcoming. "The many routes to the ubiquitous Bradley-Terry model." *Statistical Science* .

Heffington, Colton, Brandon Beomseob Park and Laron K Williams. 2019. "The "Most Important Problem" Dataset (MIPD): a new dataset on American issue importance." *Conflict Management and Peace Science* 36(3):312–335.

Hobolt, Sara Binzer and Robert Klemmensen. 2008. "Government responsiveness and political competition in comparative perspective." *Comparative Political Studies* 41(3):309–337.

Hopkins, Daniel J and Hans Noel. 2022. "Trump and the shifting meaning of "conservative": Using activists' pairwise comparisons to measure politicians' perceived ideologies." *American Political Science Review* 116(3):1133–1140.

Jennings, Will and Christopher Wlezien. 2011. "Distinguishing between most important problems and issues?" *Public Opinion Quarterly* 75(3):545–555.

Johns, Robert. 2010. "Measuring issue salience in British elections: competing interpretations of "most important issue"." *Political Research Quarterly* 63(1):143–158.

Jones, Bryan D. 2016. "The comparative policy agendas projects as measurement systems: Response to Dowding, Hindmoor and Martin." *Journal of Public Policy* 36(1):31–46.

Jones, Bryan D and Frank R Baumgartner. 2005. *The politics of attention: How government prioritizes problems*. University of Chicago Press.

Jones, Bryan D, Frank R Baumgartner, Sean M Theriault, Derek A Epp, Cheyenne Lee, Miranda E Sullivan and Chris Cassella. 2023. "Policy agendas project: codebook." *Comparative Agendas Project. URL: https://www. comparativeagendas. net/us* .

Kinnear, George, Ian Jones and Ben Davies. 2025. "Comparative judgement as a research tool: A meta-analysis of application and reliability." *Behavior Research Methods* 57(8):222.

Knuth, Donald E. 1998. *The art of computer programming: Sorting and searching, volume 3*. Addison-Wesley Professional.

Lauderdale, Benjamin E and Jack Blumenau. 2025. "Polarization over the priority of political problems." *American Journal of Political Science* .

Leeper, Thomas J and Joshua Robison. 2020. "More important, but for what exactly? The insignificant role of subjective issue importance in vote decisions." *Political Behavior* 42(1):239–259.

Mardia, Kanti V, John T Kent and Charles C Taylor. 2024. *Multivariate analysis*. John Wiley & Sons.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori and Phillip Schmedeman. 2024. "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale." *Research & Politics* 11(1):20531680241231468.

Mosteller, Frederick. 1951. "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations." *Psychometrika* 16(1):3–9.

Murray, Jared S. 2018. "Multiple Imputation: A Review of Practical and Theoretical Findings." *Statistical Science* 33(2):142–159.

Pennings, Paul. 2005. "Parties, voters and policy priorities in the Netherlands, 1971-2002." *Party Politics* 11(1):29–45.

Pickett, Justin T. 2019. "Public opinion and criminal justice policy: Theory and research." *Annual Review of Criminology* 2(1):405–428.

Rubin, Donald B. 1987. *Multiple imputation for survey nonresponse*. New York: Wiley.

Rubin, Donald B. 1996. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91(434):473–489.

Thurstone, L. L. 1927. "A Law of Comparative Judgment." *Psychological Review* 34:273–286.

Wlezien, Christopher. 2005. "On the salience of political issues: The problem with 'most important problem'." *Electoral studies* 24(4):555–579.

Xie, Xianchao and Xiao-Li Meng. 2017. "Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial?" *Statistica Sinica* pp. 1485–1545.

Yeager, David Scott, Samuel B Larson, Jon A Krosnick and Trevor Tompson. 2011. "Measuring Americans' issue priorities: A new version of the most important problem question reveals more concern about global warming and the environment." *Public Opinion Quarterly* 75(1):125–138.

Zucco Jr, Cesar, Mariana Batista and Timothy J Power. 2019. "Measuring portfolio salience using the Bradley–Terry model: An illustration with data from Brazil." *Research & Politics* 6(1):2053168019832089.

Appendix to Studying Individual-Level Attitudes Towards Important Problems Using Pairwise Comparisons *

Benjamin E Lauderdale *University College London*

Contents

*This version: 13 February 2026

Appendix A: Comparative Agendas Project Category Presentation

Table A1: Comparative Agenda Project (CAP) topic codes, category names, and presentation in this study.

CAP Code	CAP Name	Survey Presentation
1	Macroeconomics	The Economy (including inflation, jobs and unemployment, interest rates, the government budget, and taxes)
2	Civil Rights	Rights and Freedoms (including voting rights, freedom of speech, privacy, and discrimination by race, ethnicity, sex, gender, age or disability)
3	Health	Health (including the health care system, health care cost and availability, pharmaceuticals and medical device, insurance, doctors and nurses, public health, child health, mental health, and substance abuse)
4	Agriculture	Agriculture (including import and export of food, subsidies to farmers, food safety, food marketing, animal and crop disease, and fishing)
5	Labor	Work and Employment (including worker safety, employment training, employee benefits, labour unions, youth employment and migrant workers)
6	Education	Education (including primary and secondary education, higher education, vocational education and apprenticeships, and special educational needs)
7	Environment	Environment (including water, waste disposal, pollution, recycling, habitats and wildlife, conservation and climate.)
8	Energy	Energy (including electricity generation, gas production and supply, nuclear energy, wind and solar energy, and energy conservation)
9	Immigration	Immigration (including refugees, unauthorised migrants, skilled worker visas, family visas, and pathways to permanent status and citizenship)
10	Transportation	Transportation (including roads, railways, air travel, ferries and shipping, and transportation infrastructure)
12	Law and Crime	Law and Crime (including all types of crime, policing, prosecution, court administration, prisons, family law and the protection of children)
13	Social Welfare	Social Welfare (including benefits and support for the elderly, for those on low incomes, for those with disability, and for child care)

14	Housing	Housing (including all types of housing in urban and rural areas, and housing for those on low-incomes, the elderly, and the homeless)
15	Domestic Commerce	Regulation (including regulation of banking, investments, insurance, corporations, bankruptcy, consumer safety, and sport)
16	Defense	Defence (including all services of the military, intelligence services, nuclear weapons, military aid, personnel and recruitment, military bases and military operations)
17	Technology	Technology (including space and satellites, telecommunications, broadcasting, the computer industry and regulation of the internet, and research and development of new technology)
18	Foreign Trade	Foreign Trade (including trade agreements, imports, exports, tariffs, foreign investments, and foreign exchange rates)
19	International Affairs	International Affairs (including diplomacy, foreign aid, international development, international organisations, human rights, and terrorism)
20	Government Operations	Government Operations (including local government, government bureaucracy, government employment, government property, tax administration, government statistics, and scandals)
21	Public Lands	Public Land (including national parks and historic sites, natural resource management, disaster preparedness, and overseas territories and dependencies)
23	Culture	Culture (including language, arts and music, libraries, heritage and preservation, cultural industries and exchange, and diversity and inclusion)

Appendix B: Stan Code

```
data {

    int<lower = 1> N_obs;
    int<lower = 1> N_respondents;
    int<lower = 1> N_domains;
    int<lower = 1> N_covariates;

    array[N_obs] int<lower=1, upper = N_domains> ID_A;
    array[N_obs] int<lower=1, upper = N_domains> ID_B;
    array[N_obs] int<lower=1, upper = N_respondents> ID_respondent;
```

```
    array[N_obs] int<lower = 1, upper = 3> R_choice;
    array[N_obs] int<lower=0,upper=1> R_use;
    vector[N_obs] R_weight;

    matrix[N_respondents,N_covariates] X;

}

parameters {

    cholesky_factor_corr[N_domains] L_Sigma;
    vector<lower=0>[N_domains] Sigma_scale;

    matrix[N_domains,N_respondents] psi_std;

    real<lower=0> omega;
    real delta;

    vector[N_respondents] loggamma_respondent;
    real<lower=0> loggamma_scale;

    array[N_covariates] sum_to_zero_vector[N_domains] beta_std;
    real<lower=0> alpha_scale;
    real<lower=0> beta_scale;

}

transformed parameters {

    vector[N_obs] mu;
    matrix[N_respondents,2] gamma;
    matrix[N_domains,N_covariates] beta;
    matrix[N_respondents,N_domains] psi;
    array[N_obs] simplex[3] pi;

    // avg respondent has -1 to 1 neutral response width on latent scale
    for (i in 1:N_respondents){
        gamma[i,1] = -0.5*exp(loggamma_scale*loggamma_respondent[i]);
        gamma[i,2] = 0.5*exp(loggamma_scale*loggamma_respondent[i]);
    }

    for (j in 1:N_domains) {
```

```
        beta[j,1] = alpha_scale * beta_std[1,j];
        if (N_covariates > 1) for (l in 2:N_covariates) beta[j,l] =
            beta_scale * beta_std[l,j];
    }

    psi = X * beta' + (diag_pre_multiply(Sigma_scale,L_Sigma) * psi_std)' ;

    for(i in 1:N_obs) {
        mu[i] = delta + psi[ID_respondent[i],ID_B[i]] - psi[ID_respondent[i],ID_A[i]];
        pi[i,1] = 1 - Phi((mu[i]-gamma[ID_respondent[i],1])/omega);
        pi[i,2] = Phi((mu[i]-gamma[ID_respondent[i],1])/omega) -
            Phi((mu[i]-gamma[ID_respondent[i],2])/omega);
        pi[i,3] = Phi((mu[i]-gamma[ID_respondent[i],2])/omega);
    }

}

model {

    // weak LKJ prior on covariance of respondent-level rating errors
    Sigma_scale ~ normal(0, 3);
    L_Sigma ~ lkj_corr_cholesky(1);
    to_vector(psi_std) ~ std_normal();

    // priors over covariate coefficients
    for (l in 1:N_covariates) beta_std[l] ~ std_normal();
    alpha_scale ~ normal(0, 3);
    beta_scale ~ normal(0, 3);

    // prior over respondent-level deviations from avg neutral category halfwidth
    loggamma_respondent ~ std_normal();
    loggamma_scale ~ normal(0, 3);

    // weak (half-)normal priors on omega and delta
    omega ~ normal(0, 3);
    delta ~ normal(0, 3);

    // Weighted model for outcome
    for(n in 1:N_obs) if (R_use[n]) target +=
        categorical_lpmf(R_choice[n]|pi[n]) * R_weight[n];

}
```

```
generated quantities {
    cov_matrix[N_domains] Sigma;
    Sigma = diag_pre_multiply(Sigma_scale, L_Sigma) *
        diag_pre_multiply(Sigma_scale, L_Sigma)';
}
```

Appendix C: Estimation Details

The model estimates reported are based on 2 parallel chains of 500 iterations, saved after discarding 250 warmup iterations. The results are very similar with larger numbers of shorter changes (e.g. 4 x 250). Estimation time was roughly 20 minutes per model fit (2022 Apple Mac Studio, M1 Max). Saving 1000 iterations of all model parameters for each of the models estimated requires about 1.1GB of disk space.

Appendix D: Comparison to Contemporaneous Surveys

The estimates from the model generally track other data available from the same time in the UK. A survey fielded two months before this study using a different measurement approach involving ranking problems similarly shows evidence that views on Immigration had substantially more variation across different people in the UK when compared to variation in views of other problems (Brufatto and Edwicker, 2025, p9 & p13). YouGov regularly fields a closed-ended, MIP question in the UK, in which each respondent can choose up to three options (https://yougov.co.uk/topics/society/trackers/the-most-important-issues-facing-the-country). Many of the other categories are different than the ones I have used, but these top three are the same. In the week of 10 November 2025, they found "Immigration & Asylum" was selected by 53%, "The economy" by 53%, and "Health" by 33%, which were the three most frequently selected. Just as it is possible to derive from the model estimate what proportion of respondents have each problem with the highest latent rating, is possible to estimate from the model what proportion of respondents have each problem in their top three values of ψ_j. The model estimates are 41% for Immigration, 44% for The Economy, and 49% for Health. The relative proportions differ, but despite the very different survey questions, the top three are the same under both approaches.

Appendix E: Dimensionality of Problem Assessment Covariance

The presence of correlation in the variation in how individual respondents rate these categories is statistically helpful in that it increases how much we learn about a given individual's latent ratings from a given amount of observed pairwise response data. One way to quantify this is through any of the several measures of effective dimensionality of a multivariate normal covariance matrix (Del Giudice, 2021). In the limiting case where all the components of the multivariate normal are uncorrelated, these measures equal k, the number of components; in the limiting case where all the components are perfectly correlated, these measures equal 1, because there is effectively only one independent random variate. Using a measure based on Shannon entropy, on the correlation matrix, I calculate

an effective dimensionality 16.6. This indicates that, due to the correlations in problem assessments across categories, the effective dimensionality of respondents' views is somewhat lower than the 21 problem domains initially suggests, but there is nonetheless a lot of variation in problem assessments which is idiosyncratic to individuals.

Appendix F: Cross-Validation by Problem Category

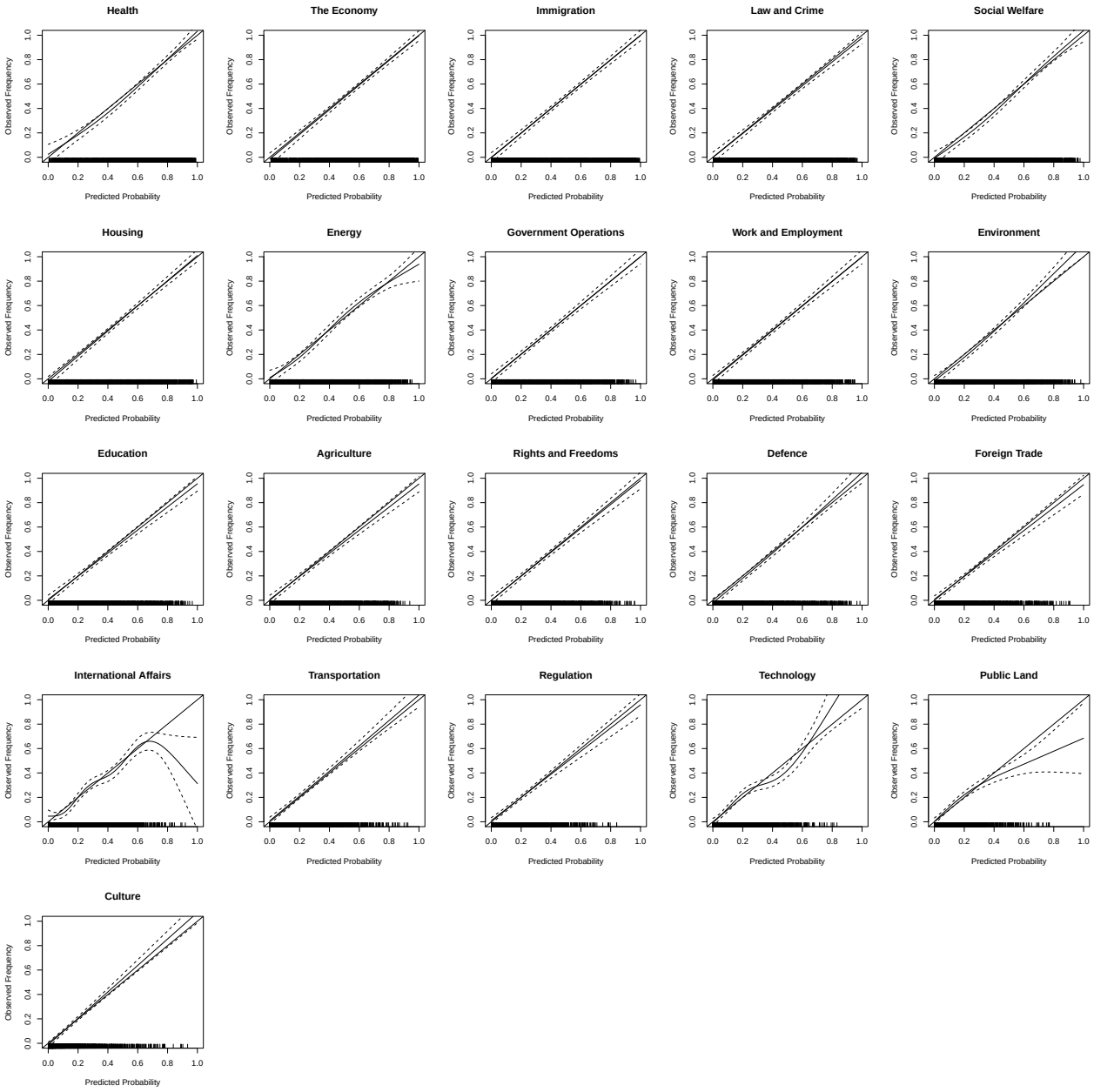

Figure A1: Performance in out of sample prediction of held out responses, in 8-fold cross-validation, aggregated by problem category.

Appendix G: Exploratory Research Example

To illustrate the exploratory value of this approach, I append an example of the kind of exploratory analysis that these data facilitate, this time using the estimated latent ratings ψ_j rather than the imputed $Y_{ijj'}$. As of the time of the survey in November 2025, a large fraction of those who had voted for Labour in the July 2024 general election no longer stated an intention to do so if there were to be an election now. In this data set, there are 586 respondents who voted Labour in 2024.[1]. At the point of this study, their vote intention had substantially fragmented, with 224 intending to vote Labour again, 101 saying they do not know how they would vote, 98 intending to vote Green, 58 intending to vote Liberal Democrat, 45 intending to vote Reform, and smaller numbers defecting to the Conservatives, Plaid Cymru and the Scottish National Party. What do each of these groups view as relatively important problems?

Figure A2 shows evidence of a few areas where problem assessments look very different across the 2024 Labour voter diaspora. The area with by far the largest numerical differences is views on Immigration, where those remaining with Labour, alongside those defecting to the Greens and Liberal Democrats, rate the problem far lower than those who are defecting to Reform, with those who do not know how they are going to vote in between. Those defecting to Reform differ from those remaining with Labour on a number of other issues, in ways we might expect given the relative policy focuses of the parties. Labour to Reform defectors are more concerned about Immigration, Law and Crime, Agriculture, and Defence, and less concerned about the Environment, Education, and Technology.

The relatively large groups of defectors to the Greens and the Liberal Democrats have modest disagreements with those remaining with Labour, with regards to which domains have more important problems. Those defecting to the Greens are somewhat more concerned about the Environment and Rights and Freedoms than those remaining with Labour, and somewhat less concerned about Immigration, Law and Crime, and Defence.

Various observers have noted that Labour finds itself in a challenging position, vis-a-vis the various directions in which it is losing voters relative to its 2024 election coalition. While the number of voters defecting to the Greens and Liberal Democrats is larger than the number defecting to Reform, there are suggestions in the data that the large group of those indicating they do not know how they will vote shares some of the same concerns on Immigration (and perhaps Law and Crime) as those already intending to vote Reform.

All of the preceding discussion is in terms of understanding the partial associations in the data, no causal statements are justified by the research design. The tendency of partisan commitments to shape how respondents answer evaluative questions related to current circumstances is well-established in a number of contexts (see, e.g. Conover, Feldman and Knight, 1987; Gerber and Huber, 2010; Ladner and Wlezien, 2007), and it is entirely possible that Labour supporters, for example, might wish to downplay the extent to which Immigration is a problem because it is a difficult issue for the Labour government, and those opposed to Labour might do the reverse. One should not look at these data alone and conclude that attitudes towards whether immigration is a problem have *caused* the changes in vote intention among those who voted Labour in 2024 versus behaviour in that election (one should

[1]Note that because YouGov maintains vote records for its panel, these respondents are identified by having said they voted Labour in the immediate aftermath of the election, not at the time of this study.

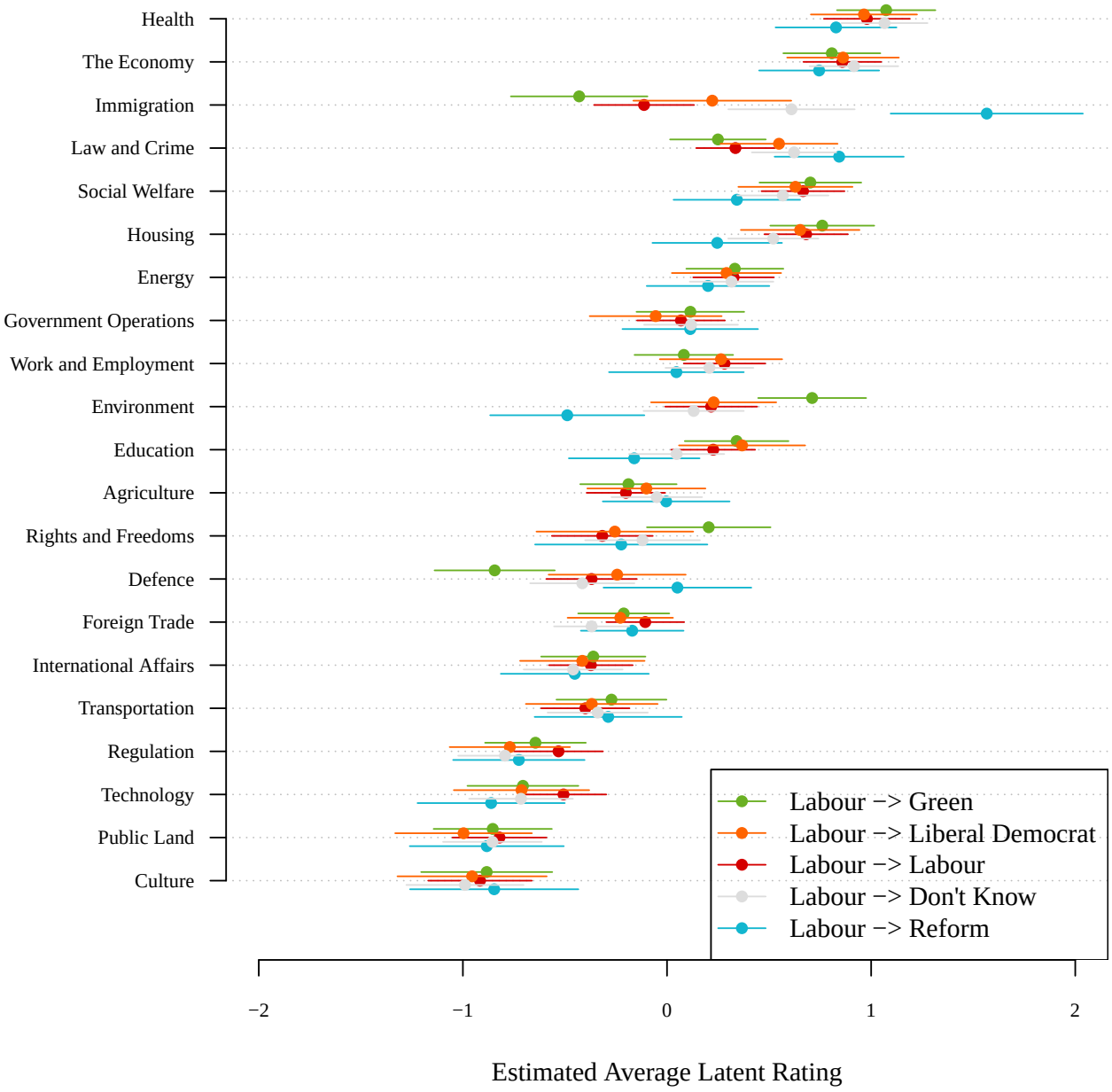

Figure A2: Estimated average latent ratings from Model 3 for the 21 problem categories across respondents by current vote intention among 2024 Labour voters, with 95% intervals accounting for measurement error in ψ calculated by Rubin's rules.

also not conclude that this is not true). But this is a familiar limitation of public opinion data, and not specific to the measurement strategy developed here.

Appendix H: Design Considerations for Future Applications

The analyses in the paper illustrate that it is possible to do several kinds of potentially valuable analyses, having collected a moderate number of pairwise comparisons (8) of a moderate number of problem categories (21) from each of a large number of survey respondents (2065), with a specific presentation of those categories and a specific prompt text and response format with a neutral option. A natural question to ask is what we could change about the data collection, and still be able to successfully do the kinds of analyses described above.

Forced choices versus neutral responses

The model presented above is made significantly more complex by the design choice to include a neutral response category. This means that one needs to worry about multiple cutpoints, about the possibility that people will use the neutral category more/less frequently for reasons that have nothing to do with the latent assessments of the categories, and the resulting model requires more parameters. If one did away with the neutral category, the data become simpler and the modelling simpler also. The tradeoff is that if you force people to make decisive choices that they find uncomfortable or difficult to make, they may refuse to respond to specific items or may drop out of your survey. This is a difficult design choice to navigate because these tradeoffs are difficult to quantify. Future work might directly compare binary and ternary data collections experimentally, however the underlying tradeoffs will vary across applications.

Sampling pairs per respondent

There are many ways that one could sample which pairs j, j' a given respondent is asked about, from the $n_j \times n_j$ matrix of possible ways to construct an ordered pair of two objects from n_j objects. The diagonal elements–comparisons of objects to themselves–are generally worth sampling, to avoid annoying respondents and collecting data of negligible value. One does, however want to sample from both the lower triangle and upper triangle of this matrix, presenting object pairs in both orders, to avoid confounding attitudes towards the objects with presentation order effects.

The study I have presented here samples independently with replacement from all non-diagonal elements in the j, j' matrix, using repeated pairs presented to the same individuals to describe response variation in terms of both a test-retest variability within respondent-pairs as distinct from variation in attitudes across respondents. The advantage of this approach is that it allows the model to isolate the stable variation in respondents' views, which is helpful for documenting that there is substantial variation of this kind, but it is not necessary for most of the subsequent analyses, which could be conducted with a slightly simpler model on data that did not include repeated pairs.

Sampling independently is not the only defensible choice however, and there are some potential gains to be made from sampling pairs that are more *connected* than will occur by chance. For example, one could sample an ordered sequence of m objects $j_1, j_2, j_3, \ldots, j_m$ construct m pairs consisting of

j_1 vs j_2, j_2 versus j_3, ..., and j_m versus j_1. If one then presented these pairings to respondents in a random order, every object would appear in two comparisons and all objects would be *connected* via comparisons to all others for that respondent. Greater connectedness of this kind increases our ability to estimate the covariance Σ of individual ratings. It yields fewer respondents seeing any particular object, but more information on where that object is situated for each of those respondents.[2]

Selecting problem categories

I have used the 21 categories of the Comparative Agendas Project as the basis for the experiment here. However, one might wish to study more categories than this, indeed the CAP sup-topic categories number more like 200. There is a straightforward tradeoff between the number of categories tested and the amount of data that can be collected on each of them. There are two quantities that are vital to consider in determining the number of categories that can be included in a data collection like this one.

First, the most relevant "sample size" quantity is the number of times each category appears in a choice. In this study, this was an average of 1573 (2 categories per prompt $\times$ 8 prompts per respondent $\times$ 2065 respondents / 21 categories. While the average number of times that each category appeared against each other category is much smaller (79) and declines rapidly with larger numbers of categories, the only parameters in the model that are specific to pairs are the $\rho_{jj'}$ in Σ governing the correlation in how respondents assess pairs of categories, which are informed not only by the direct comparisons of those categories but also by respondents whose comparisons bridge those via comparisons to other categories. Nonetheless, the more categories are present in the data, the less informative the category correlations (via Σ) are likely to be relative to covariates (via β), and so the predictiveness of the latter becomes more important.

Second, it is important to consider how decisively respondents will choose particular categories versus others. In the data set considered here, there were very large differences in the frequency of selecting particular categories over others. This meant that the differences in α were very large relative to the uncertainty in estimating these. If all of the comparisons are 60-40 and 55-45 propositions, much more data will be required to distinguish them. Similarly, the prospects for detecting differences in ψ across respondents as a function of covariates or category correlations depends very much on whether there are in fact sizeable differences between different individuals in the relevant population for the categories that are being investigated. The mere fact that not everyone gives the same pairwise response to a given pair typically indicates that there are in fact disagreements, the question is the extent to which they are predictable disagreements based on individuals' other assessments and individual characteristics.

Changing problem categories

An attractive feature of pairwise comparison designs is that they are accomodating of changing categories over time. If a new problem category became relevant in the future and it was added to the design, the vast majority of the pairwise comparisons would remain directly comparable to previous

[2]Note that this non-independent sampling of pairs is not *adaptive* testing, because the pairings presented do not depend on responses to previously presented pairings.

waves as they would involve the pre-existing categories. The new category would then be situated relative to these. Similarly, dropping categories that are no longer relevant is straightforward. So long as a large number of categories are maintained from wave to wave, there is a large amount of directly comparable data to enable analysis of those categories that have been changed.

Reducing the number of respondents or pairs per respondent

Both the number of respondents and the number of questions per respondent map directly and (typically) proportionately into survey costs. When can we get away with collecting even less data than this? Is maintaining a larger number of respondents or a larger number of questions per respondent more valuable than the other, at the margin?

My focus here will be on the range from 2 to 8 pairs per respondent, and the range from 500-2000 respondents, scaling each from 1/4 of the original data set to the full data set while holding the other fixed at its size.[3] These ranges can be assessed by taking subsets of the collected data, and correspond to feasible ranges for many data collections. The answer to how many responses we need will depend on precisely which information we are aiming to recover. First, I look at the estimation of α, ρ and ψ under Model 1 and α, β and ψ under Model 2. For all of these quantities, Figure A3 reports the model estimate of the R^2 of the point estimates for that quantity relative to overall variation in the true quantity, as defined in the paper. This is a statistic that is relevant to our ability to learn a given quantity from the data, as it directly relates the variation in the recovered statistics to the estimation uncertainty in those quantities.

Figure A3 illustrates how key parameters estimation precision relative to their overall level of variation declines with reduced data under both Model 1 and Model 2. All of the sample sizes considered are more than adequate to very precisely estimate α, the population-level averages. In general, the effects of reducing the number of questions per respondent and reducing the sample size by the same proportion tend to be very similar. The exception is that the ρ parameters estimation precision falls off more quickly with reductions in the number of pairs per respondent than with reductions in the sample size. This make sense, as reduction in the number of pairs per respondent do not just reduce the number of direct observations of a given category pair, but additionally reduce the number of categories linked through other categories for a given respondent. However, despite the increasing difficulty of estimating any particular ρ parameter with fewer pairs per respondent, the estimates of respondent-level ratings ψ nonetheless fall off in precision at similar rates as questions per respondent and sample size are reduced, even under Model 1.

Figure A4 shows a subset of the analyses using Model 3 imputed Economy versus Immigration responses as predictors of vote intention, for varying numbers of questions per respondent (top panels) and varying sample sizes (bottom panels). As we did comparing across Models 1, 2 and 3, we again see evidence of attenuation of differences in party support with the models using less data. The at-

[3]For the analyses with reduced questions per respondent, I use the first n waves in the order they were collected. For the analyses with reduced sample size, I assign each respondent to one of four sample subgroups, stratified by survey weight, in order to approximate four subsamples with roughly the same distributions of weights and respondent attributes that would occur had smaller samples been selected. I then fit models on only the first subgroup, on the first two subgroups, and the first three subgroups, so that we can compare estimates of nested samples of roughly 500, 1000, and 1500 to the full sample of about 2000 respondents. The necessarily arbitrary collection of respondent subsets means that specific parameter point estimates may differ across these in ways that are idiosyncratic to the particular subsets selected.

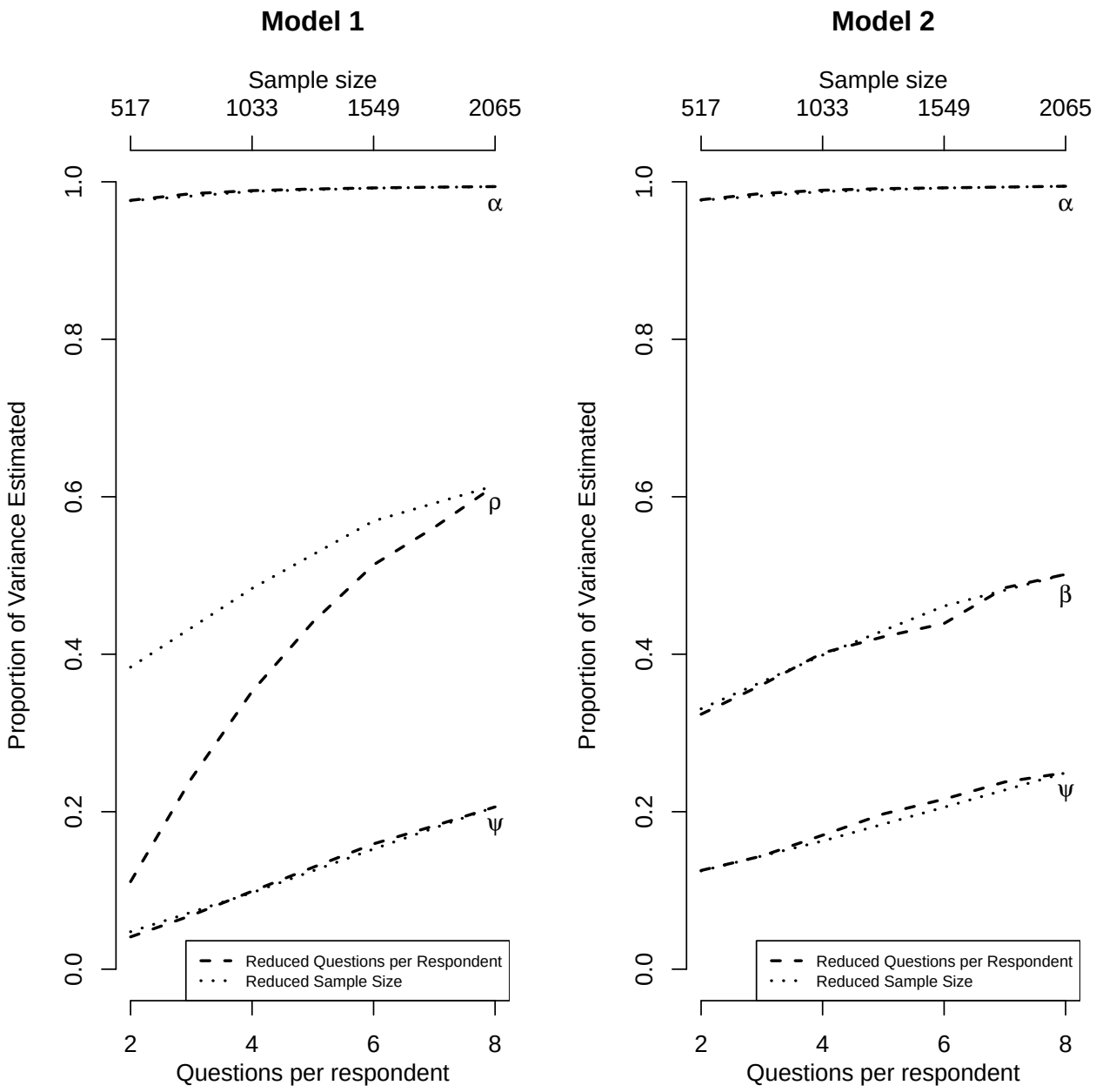

Figure A3: Selected estimates over a range of questions per respondent (bottom x axis labels) and number of respondents (top x axis labels), both scaled to be proportional to the total number of pairwise comparisons completed and holding the other at its maximum value. Left panel shows ratio of variance in parameter estimates to estimated true variation in those parameters for Model 1, right panel shows corresponding quantities for Model 2.

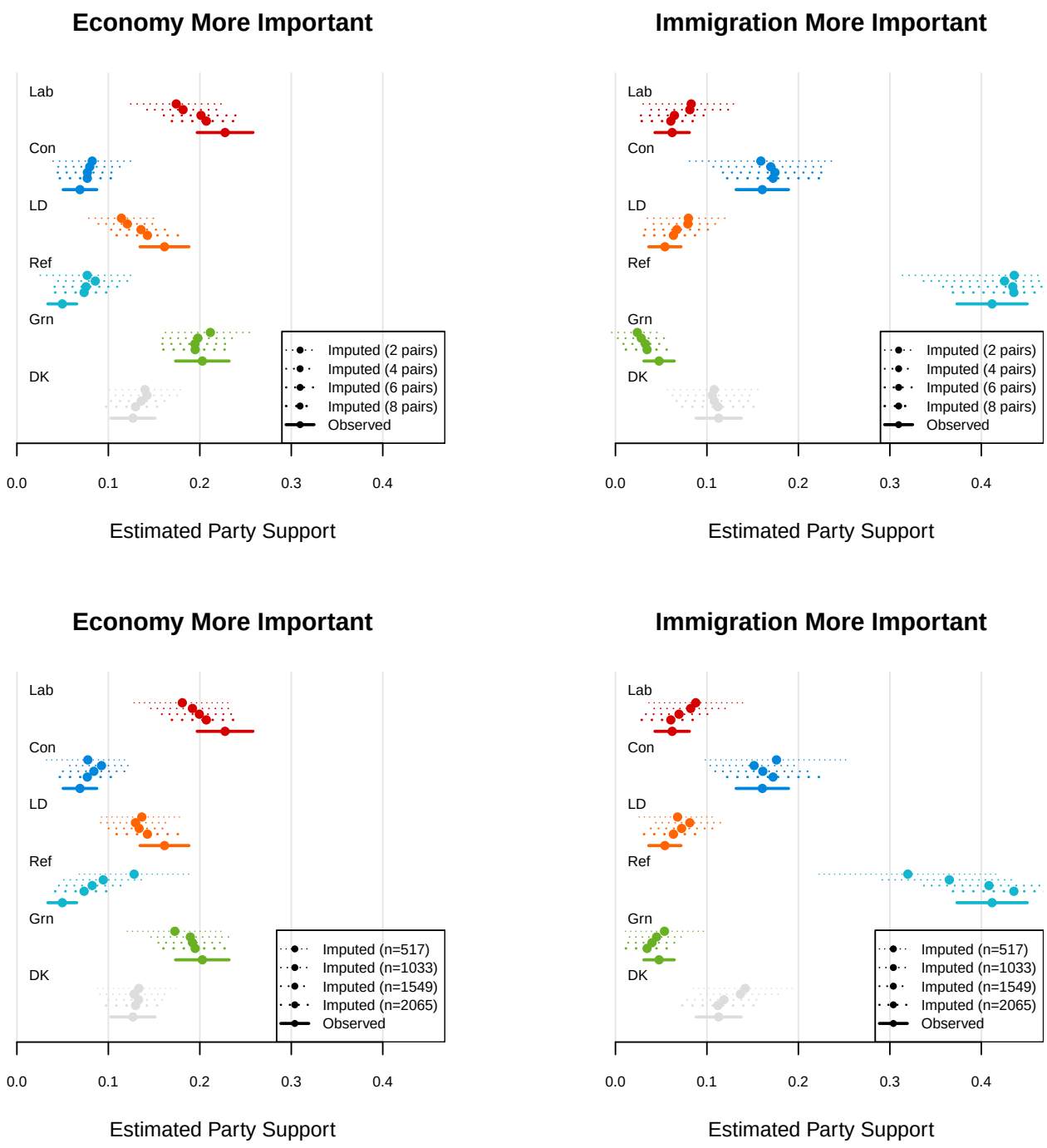

Figure A4: Analysis of vote intention by imputed and observed views on the Economy versus Immigration, using Model 3, based on varying the number of questions per respondent (top panels) and the number of respondents (bottom panels).

tenuation is generally smaller when we give up response data than when we used less rich imputation models previously, but this may not generalise. The analysis with 517 respondents, one quarter of the full data set, appears to perform substantially worse in terms of recovering the Reform party vote in particular, although it still recovers qualitatively correct differences between parties.

Ideally the interval estimates from the analyses using imputed data would entirely cover the smaller intervals from the observed data analysis, since that would imply that if the latter have good frequentist properties, so would the ones from the imputed data. This is not always the case for any of the imputation models, but it is closer to being the case when we reduce the number of pairs than when we reduce the sample size. This may be because the imputation model relies more on covariates (via β) than on correlation (via Σ), and so maintaining sample size is more important because recovering the coefficients precisely is particularly important to proper imputation.

It is difficult to provide a general quantification of the cost-precision tradeoffs here, given only one data set. However, with data like that collected on these categories in the UK in late 2025, it appears that this modelling approach works with 1000-2000 respondents and 4-8 questions per respondent, ideally with totals closer to the top of those ranges, and can recover some information with even less data than that.

References

Brufatto, Tom and Joshua Edwicker. 2025. "The UK's biggest challenges: Public attitudes towards the most important issues facing the UK and local communities". Technical report Best for Britain.

Conover, Pamela Johnston, Stanley Feldman and Kathleen Knight. 1987. "The personal and political underpinnings of economic forecasts." *American Journal of Political Science* pp. 559–583.

Del Giudice, Marco. 2021. "Effective dimensionality: A tutorial." *Multivariate behavioral research* 56(3):527–542.

Gerber, Alan S and Gregory A Huber. 2010. "Partisanship, political control, and economic assessments." *American Journal of Political Science* 54(1):153–173.

Ladner, Matthew and Christopher Wlezien. 2007. "Partisan preferences, electoral prospects, and economic expectations." *Comparative Political Studies* 40(5):571–596.