When Experts Disagree: Response Aggregation and Its Consequences in Expert Surveys

René Lindstädt, Sven-Oliver Proksch & Jonathan B. Slapin

September 22, 2018

Abstract

Political scientists use expert surveys to assess latent features of political actors. Experts, though, are unlikely to be equally informed and assess all actors equally well. The literature acknowledges variance in measurement quality, but pays little attention to the implications of uncertainty for aggregating responses. We discuss the nature of the measurement problem in expert surveys. We then propose methods to assess the ability of experts to judge where actors stand and to aggregate expert responses. We examine the effects of aggregation for a prominent survey in the literature on party politics and EU integration. Using a Monte Carlo simulation, we demonstrate that it is better to aggregate expert responses using the median or modal response, rather than the mean.

1 Introduction

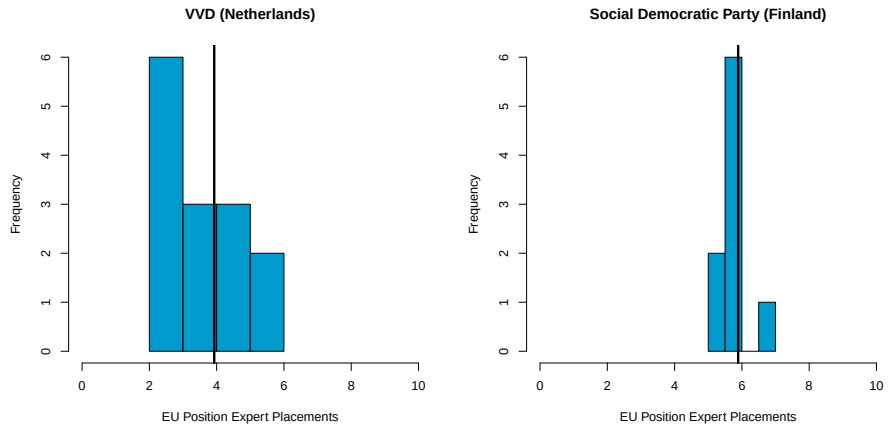
Political scientists rely on expert surveys to measure a wide array of variables — the positions of parties on policy dimensions (e.g. Benoit and Laver, 2006; Bakker et al., 2015), the importance of portfolios (Druckman and Warwick, 2005), the effectiveness of regional trade agreements (Gray and Slapin, 2012), the preferences of bureaucracies (Clinton and Lewis, 2008) and the quality of elections (Norris, Frank and Martínez i Coma, 2013). Yet, experts' ratings are rarely in perfect agreement. While scholars have explored the variation in expert placements (Hooghe et al., 2010; Martínez i Coma and Ham, 2015) and have proposed solutions to anchor experts on the scales (Bakker et al., 2014), we argue that these approaches are insufficient for understanding the nature of the measurement problem in data derived from expert assessments.

If a single expert were perfectly knowledgeable, the opinion of that expert may be sufficient and preferable to multiple opinions of lesser informed ones. But researchers do not know how knowledgeable experts are. The goal of an expert survey is thus to aggregate the responses from many experts, typically by taking the mean. We challenge this widely accepted form of response aggregation and demonstrate that mean expert responses may produce biased estimates of the latent concept researchers wish to measure. Confusion arises in part, we argue, because political scientists have not adequately distinguished between the tasks of *inference* and *aggregation* in expert surveys, leading to insufficient attention paid to problems surrounding aggregation. Taking the mean leads to bias due to scale truncation and central tendency biases among respondents in low information environments. We demonstrate the problem using Monte Carlo simulations, data from a prominent expert survey, and by conducting a replication of a prominent study. We provide an easy-to-implement solution to this aggregation problem — using the median or modal expert response, rather than the mean.

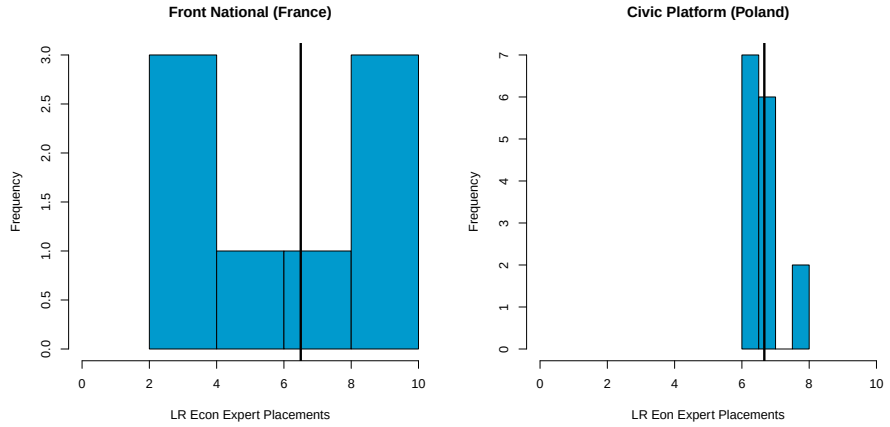
2 Example: Expert Surveys on Party Positions

Studies using expert surveys largely rely on mean expert placement, using standard deviations or standard errors to assess uncertainty. Yet, the shape of expert placement distributions can vary drastically across the items in a survey. Political parties, for example, can have similar estimated mean party positions based on very different distributions of expert placements.[1] Figure 1 shows distributions of expert responses for selected parties on an *EU Integration* dimension and a *Left-Right Economic Policy* dimension, two commonly used scales from the Chapel Hill Expert Survey (CHES) (Bakker et al., 2015), a widely used dataset. On the *EU Integration* dimension, the mean expert placement suggests that the Dutch VVD, an economically liberal party, and the Finnish SDP, a center-left party, are moderately

[1] Other fields take different approaches to similar problems. In medicine, expert panels rate the severity of disease on scales and use consensus or median expert opinion to reach diagnoses (Bertens et al., 2013).



(a) EU Integration



(b) Left-Right Economic Policy

Figure 1: Distribution of expert responses on two common policy dimensions.

Euroskeptic.[2] However, whereas experts in the Netherlands strongly disagree about the position of the VVD, experts in Finland strongly agree on the position of the SDP.

The *Left-Right Economic Policy* dimension, which typically shows smaller variation in expert placements, uncovers similar problems. Figure 1 shows expert placements for *Left-Right Economic Policy* for the French National Front and the Polish Civic Platform. The Front National has a bimodal distribution; on average, experts judge it to be a center-right party. But the contrast to the Polish Civic Platform — a party with a similar average position — is striking. Experts use almost the entire scale to place the French party, while experts agree that the Polish party is center-right. These illustrations underscore that distributions of expert placements can vary drastically despite having similar means. Moreover, in the example of the National Front, the mean provides an answer that is likely wrong. Assuming all experts are equally well-informed, our best guess about the party position ought to

[2]The *EU Integration* dimension is problematic as experts tend not to place parties in the middle of the dimension (Proksch and Lo, 2012).

be somewhere near 2–3 or 8–10, the regions where most expert assessments lie; it should not be a value in the middle where the fewest experts locate the party. We explore this problem more systematically by using a Monte Carlo simulation. We demonstrate that when experts do not assess positions perfectly, the mean leads to biased estimates of true positions. Researchers are better off using the median or modal response.

3 Expert Surveys and Statistical Inference

The statistical inference problem in expert surveys differs substantially from the frequentist notions of inference researchers typically apply when analyzing public opinion surveys (see Benoit and Laver, 2006, ch. 4). In public opinion surveys, researchers wish to measure a population parameter by randomly sampling observations from that population. In contrast, the primary objective of expert surveys is *not* to learn about the experts, who are *not* chosen at random from a population. Rather, researchers wish to glean information from experts on a topic on which they have expertise. Because researchers do not necessarily know how knowledgeable their experts are, they ask many experts and aggregate their responses, hoping the aggregate response is closer to the truth. In asking for and aggregating multiple experts' responses, two problems arise: the first results from experts' differing perceptions and the second from the nature of scale truncation.

Formally, assume that an object to be rated has a true, latent position γ on some continuous scale which researchers ask n experts to assess.[3] Typically, researchers ask these experts to make their assessment on a (Likert) scale with a limited number of response options. Each expert assessment x_i, where $i = 1, \ldots, n$, forms part of a vector of expert responses $X = (x_1, x_2, \ldots, x_n)'$. Let there be an expert-specific function, $g_i(\cdot)$, where $i = 1, \ldots, n$, which maps the true party position γ to the expert assessments, X. We wish to infer γ from X, and our ability to do so rests on the nature of $g_i(\cdot)$. If expert assessments were continuous, we might assume $g_i(\cdot)$ to be a linear function such that:

$$\begin{aligned} x_i &= g_i(\gamma, \alpha, \beta, \epsilon) \\ &= \alpha_i + \beta_i \gamma + \epsilon_i, \end{aligned} \tag{1}$$

where α is a shift parameter, β is a stretch parameter, and ϵ is noise. If all experts are perfectly informed and have no biases ($\alpha_i = 0$, $\beta_i = 1$, $\epsilon_i = 0$ $\forall i$), then each $x_i = \gamma$. Having one expert is as good as having hundreds. However, if the experts

[3]Increasing the number of experts does not increase certainty around the measurement of γ. Benoit and Laver (2006, ch. 4) claim otherwise. They calculate standard errors for party positions based on the standard deviation of expert placements as well as the number of expert placements. However, this approach has been almost unanimously rejected by the literature on interrater agreement (e.g., Kozlowski and Hattrup, 1992; LeBreton and Senter, 2008). If experts were drawn at random from the population of all experts, increasing the number of respondents would shrink the standard error of the mean expert perception of a party's position. But being increasingly confident about the mean expert perception does not imply that experts are actually good at assessing the latent party position.

are not all equally informed, uniformly poorly informed, or have different biases in their perceptions of γ, they will not respond in the same manner.

Existing rescaling techniques account for differences arising from individual respondent biases and perceptions, referred to as differential-item functioning (DIF) (Aldrich and McKelvey, 1977; Hare et al., 2015). These models estimate the expert-specific α and β parameters in Equation 1, but they do not account for another type of bias that arises as the result of rating items. When assessing objects that lie at the extremes of the scale, the truncated nature of the scale means that experts can only make mistakes in one direction, namely towards the middle. Experts with a tendency towards making centrist placements when uninformed (central tendency bias) are doubly susceptible to this bias.[4] Even when all experts perceive the scale identically, if any noise exists, truncation bias must exist as well. Mean expert ratings will estimate objects located near the extremes as increasingly centrist as random noise increases. And among summary statistics, the mean will be most affected by random centrist placements resulting from noise. Better and worse measured objects could even change rank positions when summary statistics more robust to centrist outlying placements are used for aggregation.

We could reduce truncation bias by using only the responses of better informed experts, except that we have no good way of identifying them. The best we can do is to observe the distribution of all expert responses to assess whether poorly informed experts may exist. Increasing the number of expert responses does not provide us with greater certainty about γ, but it does allow us to get a better sense of the shape of the distribution of expert responses. We can then determine the consequences of aggregation using different summary statistics in the presence of expert disagreement.

Existing robustness and validity checks applied to expert survey responses insufficiently assess the consequences of aggregation. Most analyses focus on the mean expert placement, and may, at best, examine disagreement using the standard deviation of placements (Hooghe et al., 2010). Although recent analyses take concerns about differing expert scale perceptions into account (Clinton and Lewis, 2008; Bakker et al., 2014), they do not consider truncation bias. In the next section, we explore these consequences using Monte Carlo simulations.

4 Monte Carlo Simulation

We simulate a data generating process underlying expert assessments of parties and determine when aggregate measures best capture true positions. Our latent dimension is continuous on a given interval. We generate 100 true positions by taking draws from a uniform distribution ranging from 0 to 10. We refer to this vector of true positions as γ. In the real world, researchers design the expert survey and ask experts to make an assessment of the truth on a discrete scale, often ranging from 0 to 10: $y \in [0, 1, \ldots, 10]$.

The simulation, which we run 1,000 times, begins with "experts" reporting their perceptions of the positions on the discretized scale y. In our simulation, we draw

[4]For an extensive discussion of central tendency bias, see James, Demaree and Wolf (1984).

expert j's assessment of party i as follows:

$$rating_{ij} = \alpha_j + \beta_j \gamma_i + \epsilon_{ij}, \tag{2}$$

with each $rating_{ij}$ rounded to the nearest integer and truncated to lie between 0 and 10. The quantity α is an expert-specific shift parameter and β is an expert-specific stretch parameter. Following the DIF setup, these parameters allow each expert to perceive the space differently. The error term means that experts assess some parties better than others. We run the simulation four times using 5, 10, 15 and 20 experts to examine the effect of consulting more experts. More information on the simulation parameter values is located in the supplemental appendix.

Having drawn expert assessments, we calculate the mean, median, and mode response for each party.[5] Because we set γ, the true party positions, we can assess how well the mean, mode and median of the expert assessments recover the truth. We regress the truth, γ, on each of the summary statistics of the expert responses — the mean, median or mode — for each of the 1,000 simulated expert datasets. We expect an estimated regression slope of 1, representing a perfect relationship with the truth. We assess performance of the summary statistics using OLS to mirror how expert data are typically used — as an independent variable in a regression model. Although we examine the relationship between our aggregate measure and the truth, any bias we find would also be present in the relationship between the aggregate measure and a dependent variable causally related to the truth.

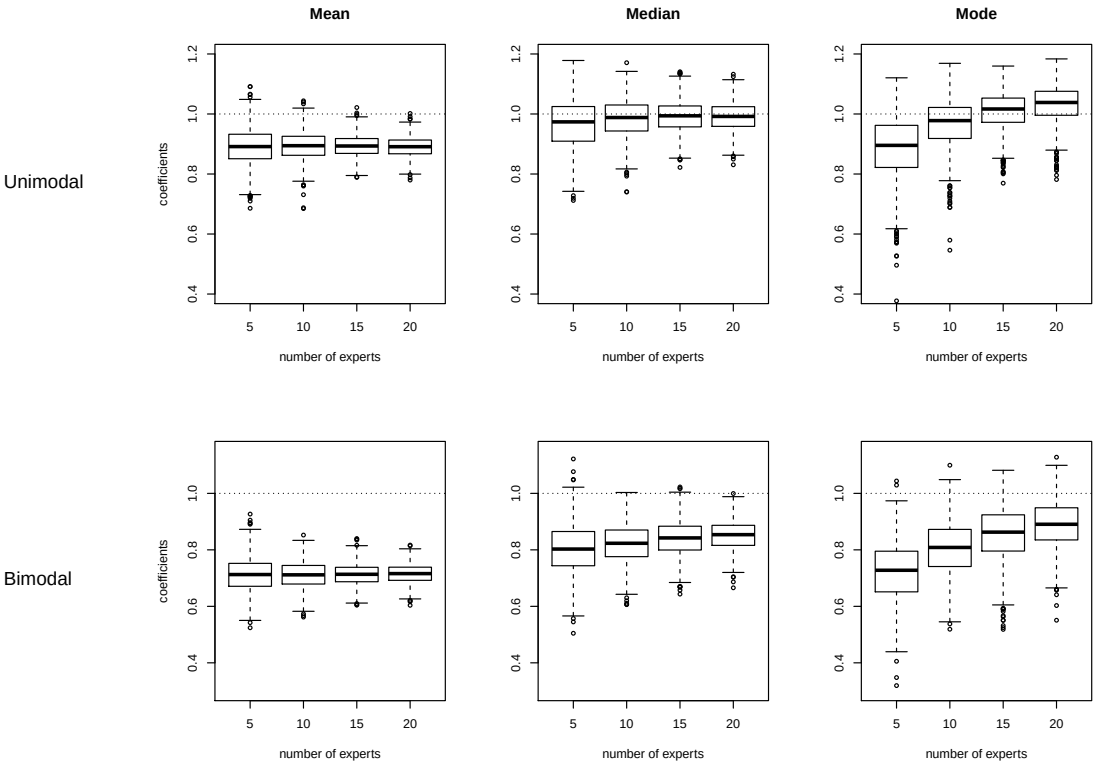
After setting the true positions, the simulation steps are as follows:

1. Draw n expert assessments for each of the 100 parties using Equation 2, where n is 5, 10, 15, or 20.
2. Round the expert assessments to the nearest integer and truncate them, so that all assessments lie between 0 and 10.
3. Aggregate the expert assessments using the mean, median and mode for each of the 100 parties.
4. Estimate three bivariate regressions of true positions on each of the aggregate measures and save the slope coefficients.
5. Repeat steps 1–4 1,000 times and generate a boxplot of the 1,000 slope coefficients.

We run a second set of simulations to capture the possibility that some experts perceive a party in a systematically different manner than other experts. This second simulation captures one way in which a bimodal pattern such as that seen for the *Front National* in Figure 1 may arise. Experts are selected at random (with probability 0.35) to view a subset of extreme parties (35% of parties with a position greater than 7.5 or less than 2.5) in mirror image. For example, while most experts would observe a party with a true position of 8, the randomly selected subset of experts would view a party with a true position of 2. This simulation is equivalent

[5]In secondary analysis, we account for uncertainty by using a non-parametric bootstrap of the expert responses, calculating the aggregation statistic in each of the simulated response datasets, and then accounting for that measurement error in all models.

Figure 2: *Simulations across different numbers of experts.*

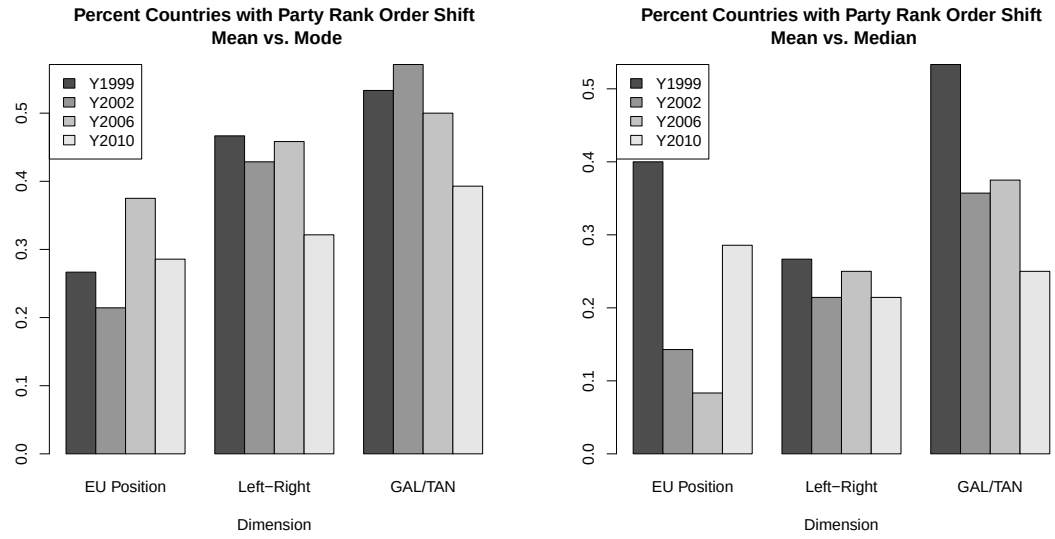


to a case in which experts are asked to rate a populist far right party on a traditional left-right economic dimension. The majority of experts view the populist economic policies of a far right party as right-wing policies, but a subset of experts recognise these policies (likely to include state intervention and subsidies) as matching notions of left-wing economic interventions. While neither is wrong, we take as the truth the scale mapping that the majority of experts see.

The results of both sets of simulations are presented in Figure 2. The top row of Figure 2 presents the results when all experts perceive the scales identically. The mean is biased and consistently underestimates the truth. Adding more experts reduces noise in the estimates but it does not reduce the bias. The median recovers the true relation nearly perfectly regardless of the number of experts asked, and increasing the number of experts reduces noise. Finally, the mode recovers the truth most accurately when we use 15 experts. Increasing the number of experts further reduces noise, but the mode starts to overestimate the truth. In the simulation where some experts view a mirror image of the truth, presented in the bottom row, the mean greatly underestimates the truth. The median and mode both perform much better, although they, too, underestimate the truth. The small subset of experts who view the truth differently have a greater impact on the mean than the median or mode.

These simulation results have important implications for the design and interpretation of expert surveys. If there is reason to think that experts may not be fully knowledgeable, possess varied biases, or perceive scales differently, then researchers

Figure 3: Party rank order changes resulting from aggregation.



should recognize that different summary statistics can provide different answers about the nature and impact of party positions. These simulations suggest that the median and mode recover the truth better than the mean. At a minimum, the results suggest that when faced with discrepancies in expert placements, researchers should check the robustness of their results to different means of aggregation.

The CHES party expert survey further demonstrates that researchers must carefully consider their approach to aggregating expert placements. Figure 3 presents the percentage of countries in each survey wave that experience at least one rank order shift among the parties as a result of using the mean versus the mode and the mean versus the median. Depending on survey and dimension, between 20% and 50% of countries have at least one rank order change in their party system as a result of using a different aggregation method.

5 Application — Understanding Party Position Shifts

Response aggregation can affect the results of studies relying on expert placements. We replicate a study by Adams, Ezrow and Somer-Topcu (2014), which examines how citizens update their views on parties' policy position shifts. In doing so, we also show how a simple bootstrap can help gauge the effects of uncertainty in the expert data on our inferences. Adams, Ezrow and Somer-Topcu (2014) argue that citizens, rather than relying on party manifestos to update their information on party policy position shifts — a view that has had a long tradition in the extant literature —, draw on a *variety* of information sources when updating their beliefs. Adams, Ezrow and Somer-Topcu use expert opinions from the CHES surveys as a proxy for broad information gathering. Focusing on European integration, their empirical analysis

Table 1: Citizens' perceptions of parties' policy shifts on European integration (Adams, Ezrow and Somer-Topcu, 2014): Replication and alternative models.

	W/O Clustered Mean	**Clustered Mean**	**Bootstrapped Median**	**Bootstrapped Mode**	**Mean** $r_{wg} > 0.7$
Party j's perceived shift using experts (t)	0.263 (0.092)	0.263 (0.082)	0.150 (0.087)	0.107 (0.071)	0.215 (0.117)
Party j's shift using Euromanifestos (t)	−0.192 (0.170)	−0.192 (0.137)	−0.150 (0.178)	−0.155 (0.179)	−0.191 (0.211)
Intercept	0.138 (0.069)	0.138 (0.062)	0.134 (0.072)	0.135 (0.073)	0.101 (0.086)
Adjusted R^2	0.085	0.085			0.036
N	78	78	78	78	59

confirms their hypothesis. The finding is an important contribution to the ongoing debate about political sophistication of citizens.

However, using the mean to aggregate divergent expert opinions when calculating party policy shifts may impact these results. The study uses one of the better measured items in the CHES data — party position with respect to European integration — and focuses on parties in Western Europe, where experts tend to display higher levels of agreement. Thus, if we find that expert disagreement creates problems in this case, it is likely to create issues in a large number of other studies, too.

First, we examine how robust the results are to different aggregation approaches. We also account for uncertainty in the aggregated expert responses resulting from disagreement among experts. If we were to simply rely on the median, modal or mean response, we would be assuming that our point estimate has no associated noise. To address this issue, we conduct a non-parametric bootstrap of the expert data. We create 100 bootstrapped expert data sets by sampling with replacement from the set of expert responses for each party on the European integration dimension. From the sampled expert responses, we calculate the modal response to construct the relevant variable and estimate the Adams, Ezrow and Somer-Topcu (2014) model 100 times — once in each sample. Finally, we summarize the results across the 100 samples using model averaging just as one would when imputing missing data (Blackwell, Honaker and King, 2015).

Table 1 presents the results. The first model replicates the *Multivariate Model (3)* in Adams, Ezrow and Somer-Topcu without clustered standard errors. The second model uses clustered standard errors and therefore is an exact replication of *Multivariate Model (3)*. Clustering has very little effect on the standard errors. We therefore proceed to estimate the other models without clustering. Using the bootstrapped median and mode, the results of Adams, Ezrow and Somer-Topcu become much weaker (columns 3 and 4). The coefficient on the variable of interest is only 57% of its former size when using the bootstrapped median and only 41% of its reported size when using the bootstrapped mode. Neither the median nor mode variable attains statistical significance. In the final model, we use the mean

responses, but estimate the model using only well measured parties.[6] The coefficient estimate using only better measured parties is still smaller than the original estimate using the mean and all parties, further indicating that poorly measured parties with high levels of expert disagreement are contributing to the authors' findings.

Our analysis suggests that the substantive effects presented by Adams, Ezrow and Somer-Topcu (2014) may not be as strong as they suggest. Their results are at least partly driven by disagreement in expert placements of parties and the choice to aggregate these responses using the mean. Nevertheless, we would not go so far as to say that the Adams, Ezrow and Somer-Topcu (2014) results are incorrect. We are accounting for measurement error in only one of the two variables. There is almost certainly measurement error in the variable based on Euromanifestos, as well, and accounting for that measurement error could impact the coefficient on the expert survey variable. Our point is simply that disagreement among experts in rating parties can lead to incorrect inferences about the impact of party positioning when using party position as an independent variable in regression analyses.

6 Discussion and Conclusion

Political scientists make frequent use of expert surveys, but they have not properly examined the consequences of lack of expert agreement on aggregation of responses. Our findings have implications for those who wish to collect new expert survey data and those using existing data. Those running new surveys must consider the degree to which experts can assess individual items. Within party position surveys in political science, the locations of some parties and on some dimensions are easier to assess than others. In the supplementary appendix, we apply common measures of agreement to the CHES data to underscore the problems of lack of agreement.

It may also be useful to design items in expert surveys aimed at gauging expert knowledge. Researchers could give more weight to knowledgeable experts and determine whether disagreement results from heterogenous expert ability or a fundamental lack of agreement on where targets lie on the scale. Lastly, it might be useful to think about other types of survey designs, beyond Likert scales, that may lessen the cognitive burden placed on experts, resulting in higher levels of agreement (e.g. pairwise comparisons).

For those using existing data, we suggest that researchers examine expert agreement and reliability within the items they wish to use by drawing on well-known techniques (Finn, 1970; James, Demaree and Wolf, 1984; van der Eijk, 2001). If items are poorly measured, it may not be wise to use them in secondary analyses. And when disagreement among experts exists, researchers should check the robustness of their results to aggregation using the median and modal expert responses.

[6] We rely on a common measure of agreement, the r_{wg} score (Finn, 1970; James, Demaree and Wolf, 1984), which examines the dispersion of responses with reference to a null distribution. The supplementary appendix provides details on how it is calculated and applies it to the CHES data more broadly. The measure ranges from 0 (no agreement) to 1 (perfect agreement). Scores in excess of 0.7 are considered indicative of strong agreement.

References

Adams, James, Lawrence Ezrow and Zeynep Somer-Topcu. 2014. "Do Voters Respond to Party Manifestos or to a Wider Information Environment? An Analysis of Mass-Elite Linkages on European Integration." *American Journal of Political Science* 58(4):967–978.

Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(01):111–130.

Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hoogh, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Anna Vachudova. 2015. "Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999-2010." *Party Politics* 21(1):143–152.

Bakker, Ryan, Erica Edwards, Seth Jolly, Jonathan Polk, Jan Rovny and Marco Steenbergen. 2014. "Anchoring the experts: Using vignettes to compare party ideology across countries." *Research & Politics* 1(3):2053168014553502.

Benoit, Kenneth and Michael Laver. 2006. *Party Policy in Modern Democracies*. Routledge.

Bertens, LCM, BLD Broekhuizen, CA Naaktgeboren, FH Rutten, AW Hoes and Y van Mourik. 2013. "Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting." *PLoS Med* 10(10).

Blackwell, Matthew, James Honaker and Gary King. 2015. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods and Research* doi: 10.1177/0049124115589052.

Clinton, Joshua D and David E Lewis. 2008. "Expert opinion, agency characteristics, and agency preferences." *Political Analysis* 16(1):3–20.

Druckman, James N and Paul V Warwick. 2005. "The missing piece: Measuring portfolio salience in Western European parliamentary democracies." *European Journal of Political Research* 44(1):17–42.

Finn, R.H. 1970. "A note on estimating the reliability of categorical data." *Educational and Psychological Measurement* 30:71–76.

Gray, Julia and Jonathan B. Slapin. 2012. "How effective are preferential trade agreements? Ask the experts." *The Review of International Organizations* 7(3):309–333.

Hare, Christopher, David Armstrong, Ryan Bakker, Royce Carroll and Keith T. Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine De Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen and Milada Vachudova. 2010. "Reliability and Validity of the 2002 and 2006 Chapel Hill Expert Surveys on Party Positioning." *European Journal of Political Research* 49(5):687–703.

James, Lawrence R., Robert G. Demaree and Gerrit Wolf. 1984. "Assessing within-group interrater reliability with and without response bias." *Journal of Applied Psychology* 69(1):85–98.

Kozlowski, Steve W.J. and Keith Hattrup. 1992. "A Disagreement About Within-Group Agreement: Disentangling Issues of Consistency Versus Consensus." *Journal of Applied Psychology* 77(2):161–167.

LeBreton, James M. and Jenell L. Senter. 2008. "Answers to 20 Questions about Interrater Reliability and Interrater Agreement." *Organizational Research Methods* 11(4):815–852.

Martínez i Coma, Ferran and Carolien Ham. 2015. "Can experts judge elections? Testing the validity of expert judgments for measuring election integrity." *European journal of political research* 54(2):305–325.

Norris, Pippa, Richard W. Frank and Ferran Martínez i Coma. 2013. "Assessing the quality of elections." *Journal of Democracy* 24(4):124–135.

Proksch, Sven-Oliver and James Lo. 2012. "Reflections on the European integration dimension." *European Union Politics* 13(2):317–333.

van der Eijk, Cees. 2001. "Measuring Agreement in Ordered Rating Scales." *Quality and Quantity* 35(3):325–341.