European Journal of Political Research **46**: 347–366, 2007
doi: 10.1111/j.1475-6765.2006.00694.x

Evaluating expert judgments

MARCO R. STEENBERGEN[1] & GARY MARKS[2]
[1]*University of North Carolina, USA;* [2]*University of North Carolina, USA and Free University of Amsterdam, The Netherlands*

Abstract. Although expert surveys have gained a prominent place in comparative studies of party positions on issues, their validity has been called into question. In this article, some of the validity concerns are evaluated in the context of the authors' own expert survey on national party positions vis-à-vis European integration. One goal of the article is to demonstrate that this expert survey produces valid measures of party positions. An equally important goal, however, is to suggest some methods that can help in assessing the quality of expert survey data. These methods, which are rooted in psychometric theory, are applicable in a variety of contexts and are easily implemented.

Introduction

Expert survey data play an ever greater role in the measurement of policy positions of political parties (Castles & Mair 1984; Huber & Inglehart 1995; Laver & Hunt 1992; Laver & Mair 1999; Ray 1999). Such data form an alternative to party manifestos (Budge et al. 2001), voter and elite perceptions of party positions, or roll call data (Thomassen et al. 2004).[1] Expert surveys provide an economical way of measuring party positions. They can be administered at any time, unlike manifestoes which are tied to electoral calendars. As long as experts are willing to respond to surveys, the expert survey methodology may probe topics that do not surface in manifestos or other data sources such as internal dissent within a party. Thus, expert surveys offer a number of advantages, which may explain their popularity (for a general discussion of expert survey methodology, see Meyer & Booker 1991). The question is how valid is the information that one gets from these surveys.

The validity question was raised prominently by Ian Budge (2000) in a cogent assessment of the uses and limitations of expert surveys. Budge wonders what it is that experts evaluate – what aspect of a party do they judge, during what time frame, and with what criteria? These are important questions, which require an answer before we can be confident about relying on expert judgments for substantive research. Unfortunately, despite the growth in

Published by Blackwell Publishing Ltd., 9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

expert surveys in political research, not much explicit attention has been given to these questions. In this article, we hope to change that for our own data: a 1999 expert survey of national party positions on European integration, which is patterned after Ray (1999). The main goal of this article is to validate our expert survey data. We show that these data meet several basic validity criteria that speak to Budge's concerns. A second goal, however, is to lay out a framework that allows researchers to assess the validity of expert survey data more generally.

The article is organized as follows. First, we consider the validity concerns raised by Budge (2000). Next, we introduce the 1999 expert survey of national party positions on European integration. Third, we focus our analysis on the experts: how do experts answer questions in an expert survey? Fourth, we evaluate the party placements that our expert surveys produce. The focus here is on the behavior of the expert survey measure vis-à-vis other measures. We conclude by discussing the implications of this research.[2]

The validity of expert judgments

Framing survey questions demands careful preparation. How can one elicit valid responses from experts about party policy positions? (for a recent review of the complexities of survey research, see Tourangeau et al. 2000). What is the basis of the judgments that experts offer? How can we be sure that experts answer questions in the way they were intended? Budge (2000) has argued that problems can creep into expert judgments in four ways. First, what 'party' is being judged by the expert? Is it the party in the electorate, the party in government or the party organization, to borrow Key's (1964) tripartite distinction, or what? Second, what criteria do experts bring to bear when they judge party positions? For example, what do abstract labels like 'left' and 'right' mean to the expert? Third, do experts judge the intentions of parties or their behavior? Finally, what is the time frame for the judgments that we ask experts to make? Each of these questions addresses a more fundamental question: how do experts interpret the questions in expert surveys and how do they link substantive knowledge about parties to those questions?

An important concern is that experts may bring wildly varying considerations to bear when judging party positions. One expert may be judging the position of the electorate; another, the position of the party leadership; and yet another expert may be evaluating the views of party activists. Likewise, one expert may conceptualize left-right ideology in terms of the position of a party on economic issues, while another may concentrate on social issues. Some experts may consider what parties are doing in the government, while others may think about party rhetoric. Some experts may evaluate a party's position at the present moment, while others may consider a larger time span of perhaps several years. If the considerations that come into play were to vary this much, it would not be at all clear what expert judgments actually measured. Several of these concerns can be alleviated in the expert survey design. Any good questionnaire will attempt to avoid ambiguous terms such as 'party' and 'left-right', or will attempt to give them a more circumscribed meaning. Thus, it is quite common to specify that the judgment should concern the position of the party leadership and not activists or voters. Specifying time frames explicitly limits variation on this dimension. For example, experts may be asked to judge the position of the party leadership on issue X during the past year. Precise issue descriptions limit interpretation of what the issue covers. Where this is not possible, experts may be asked what criteria they used to evaluate a party. For example, in evaluating the left-right position of parties, experts may be asked to describe what 'left-right' means in a particular country or what criterion they used to define this dimension (Huber & Inglehart 1995).

In spite of these efforts, we cannot presume that better expert survey design alleviates all of the validity concerns raised by Budge (2000). For example, we know that retrospective judgments are very taxing on respondents, especially when we are asking judges to recall distant facts, as did Ray (1999) when he asked his experts to recall a party's position on European integration from twelve years earlier. Problems like telescoping (recalling facts for the wrong time period – i.e., either before they happened or afterwards) are bound to plague such judgments, even when we provide an explicit time frame (Tourangeau et al. 2000). Similarly, even the most carefully crafted question may still leave an interpretative space for experts that could distort their judgments. Fortunately, most of Budge's (2000) concerns lend themselves to empirical assessment. Rather than speculate that an expert survey may suffer from certain problems, we can actually evaluate the extent to which this is the case. Such assessments do not take the place of good expert survey design, but they can help us assess the extent to which a design has succeeded.

The 1999 expert survey of national party positions on European integration

One of the expert surveys evaluated by Budge (2000) was developed by Leonard Ray (1999) to assess the positions of national political parties vis-à-vis European integration. Ray's expert survey data run through 1996. In 1999, Gary Marks and Marco Steenbergen conducted a follow-up survey that incorporates Ray's template.[3] The 1999 survey sampled experts for the 15 European

Union (EU) Member States at that time. The list of experts consisted of political scientists and some others with expertise concerning party politics in a particular country. Many of the experts were the same as those used by Ray (1999). The number of experts for each country is listed in Table 1 along with the number of parties that were placed.[4]

The objective of the expert survey was to determine where the political parties in each EU Member State stood on issues arising from European integration. The key question for this purpose was: 'What was the overall orientation of the party leadership towards European integration in 1999?' This question had the following response options: 1 = strongly opposed to European integration; 2 = opposed to European integration; 3 = somewhat opposed to European integration; 4 = neutral; 5 = somewhat in favor of European integration; 6 = in favor of European integration; and 7 = strongly in favor of European integration. Note that the question specified the object that was to be evaluated (the party *leadership*) and the time frame for the evaluation (1999). Also note that all of the response options were explicitly labeled. This was done in order to minimize response variation due to differential scale interpretations by experts, although the efficacy of this approach is sometimes disputed (Andrews 1984). Thus, the question was designed to put the experts in a common frame of mind so that they would be judging the same object, on the same dimension, at the same point in time. We now turn to an analysis of the

Table 1. The 1999 expert survey on European integration

Country	Experts	Parties
Austria	5.0	5.0
Belgium	9.0	14.0
Denmark	7.0	12.0
Finland	5.0	11.0
France	7.0	15.0
Germany	15.0	8.0
Greece	6.0	6.0
Ireland	6.0	7.0
Italy	6.0	20.0
Netherlands	11.0	10.0
Portugal	5.0	5.0
Spain	12.0	15.0
Sweden	9.0	8.0
United Kingdom	13.0	7.0
Mean	8.4	10.2

response characteristics of our experts. We then discuss how the expert survey placement of parties corresponds to other data sources.

Experts as measurement instruments

What would happen if our worst fears about expert judgments were to materialize? We would find different experts judging different objects, on different dimensions, at different points in time. For instance, one expert might evaluate the rhetoric of party activists on dimension X at time t, while another expert might evaluate the actions of party leaders on dimension Y at time $t + j$. All of this would contribute to variation across experts, unless we experience the unlikely scenario that rhetoric and action are indistinguishable, activists and leaders hold identical views, dimensions X and Y are perfectly correlated, and the 'party' has not moved between t and $t + j$. Absent such conditions, high *variance* would be the necessary consequence of experts basing their judgments on different foundations. As a corollary, the correlation between expert judgments would be reduced. The key to assessing expert judgments, then, is to assess the variance in those judgments.

Variation across experts

How much variation is there across the experts? One approach is to compute the standard deviation of their placements of parties, which we shall do shortly. A more sophisticated approach is to perform a variance components analysis (Goldstein 1995; Searle et al. 1992; Steenbergen & Jones 2002). Let $y_{(ij)k}$ denote expert i's judgment of party j in country k. Experts and parties are cross-classified at the lowest level of analysis, as is indicated by the parentheses on their subscripts, since all parties in a country are assessed by the same set of experts. The cross-classified variance components model (Goldstein 1995) is now given by

$$y_{(ij)k} = \mu + \delta_k + \varepsilon_{ik} + \varepsilon_{jk}$$

Here μ is the grand mean – the mean party placement across experts, parties and countries – and δ_k, ε_{ik} and δ_{jk} capture country, expert and party effects, respectively, which can be thought of as displacements from the grand mean. We can treat μ as a fixed effect and δ_k, ε_{ik} and ε_{jk} as random components. As long as those components are uncorrelated with each other, the variance of y_{ijk} can be decomposed as

$$V\left[y_{(ij)k}\right] = \sigma^2_\delta + \sigma^2_{\varepsilon_i} + \sigma^2_{\varepsilon_j}$$

Here σ^2_δ is the cross-national variation in party placements, $\sigma^2_{\varepsilon_j}$ is the cross-party variance and $\sigma^2_{\varepsilon_i}$ is the cross-expert variance. It is this latter variance component that is of particular interest.

The cross-classified variance components model can be estimated using standard multilevel modeling software. Table 2 shows the full information maximum likelihood estimates obtained from MLwiN (Rasbash et al. 2004). As these estimates show, there is statistically significant variation across the experts. However, the size of this variation is quite limited: the estimated standard deviation is less than one unit on the 7-point scale. Apparently, the experts render rather similar judgments about the EU stances of the political parties. Indeed, the inter-expert correlation is quite high – namely $r = 0.788$.[5] Using the average number of experts as a foundation, this translates into a reliability of 0.969, as computed via the Spearman-Brown formula.[6] By all standards, this is a very high reliability.

An important question is whether the variance across experts depends on attributes of a party or party system. At the party system level, one could argue that experts will have a more difficult time placing specific parties when there is not much spread in the European integration stance across all parties in the country. At the party level, one could argue that party placement will become more difficult if parties are: first, smaller and perhaps covered less by the media; second, when the salience of European integration is low for a party; and third, when a party is internally divided (Steenbergen & Scott 2004; Marks et al. forthcoming). It may also be the case that parties of a particular ideological signature or party family are more difficult to place, but it is less clear in what direction those effects would flow.

In all of these scenarios, there may be considerable uncertainty about the location of political parties. At the level of a single expert, this should induce

Table 2. Variance components analysis of expert judgments

Parameter	Estimate	Standard Error
Fixed effects		
Grand mean (μ)	4.695**	0.184
Variance components		
National (σ^2_δ)	0.123	0.175
Party ($\sigma^2_{\varepsilon_j}$)	3.282**	0.423
Experts ($\sigma^2_{\varepsilon_i}$)	0.917**	0.041

Note: Table entries are full information maximum likelihood multilevel model estimates with their estimated standard errors. N = 1,127 country-party expert cases. –2 log likelihood = 3579.031. ** $p < 0.01$ (two-sided).

response variation. Rather than being able to pinpoint where a party stands, an expert may at best be able to provide a range within which the party stance is located. This would have two effects. First, even if a party has not changed its stance on European integration, the expert might give different placements of that party at different points in time (Alvarez 1997; Alvarez & Brehm 2002). Unfortunately, we lack data to test this implication, although we can say that it runs counter to the findings of McDonald and Mendes (2001) about expert judgments. We can test, however, a second effect of uncertainty: increased cross-expert variance. We would expect such variance because experts may use different criteria or heuristics in placing a party when they are uncertain about that party's true position. We measure this variance by considering the standard deviation across experts in the placement of political parties.

To assess the impact of party attributes and party system attributes on response variation across experts, we estimate another multilevel model. We measure party differentiation within a party system as the standard deviation of party positions in that country. At the party level, we include predictors for the vote percentage of the party in the previous election (a measure of party size), for issue salience (an expert judgment of the importance of the issue of European integration for the party leadership, with high scores indicating greater salience) and for internal dissent (an expert judgment of the amount of internal dissent about European integration, with high scores indicating greater dissent). We also include a measure of the left-right ideological position of a party (as determined by our experts) and dummies for eight different party families: the radical right, conservatives, liberals, Christian Democrats, Social Democrats, the radical left, greens and regional parties (parties without a clear family serve as the baseline category). Finally, we include three types of control variables. First, we include a dummy at the party system level to identify Austria, Finland and Sweden. These are the most recent entrants into the EU and party positions vis-à-vis integration might not yet have fully crystallized in those countries at the time of the survey, resulting in greater potential variability in the expert judgments. Second, we include a dummy to identify governing parties since their positions are particularly visible allowing for greater inter-expert consensus. Third, we include the effective number of experts for each party as a control since response variation may be partially a function of how many experts provided an answer.[7] As previously noted, our multilevel model predicts the standard deviation in expert judgments for a party *i* located in country *j*. The intercept of this model is allowed to vary across countries.

Table 3 shows the full information maximum likelihood MLwiN estimates of the model (Rasbash et al. 2004). We see that variation in expert judgments is a function of party differentiation, salience and dissent. As parties take more differentiated stances within a country, experts show more agreement in their placement of specific parties (i.e., the standard deviation is reduced). There is also greater consensus when European integration is salient to a party and when the party is relatively unified on the issue. Additionally, we see less variance across experts in the placement of liberal, Christian Democratic and Social Democratic parties. No other predictors attain statistical significance.[8]

In conclusion, the amount of cross-expert variation is small. It behaves predictably with variance increasing as the judgment task for experts becomes more difficult (i.e., few perceived differences across parties, low salience and high levels of internal dissent). These are encouraging findings. However, they

Table 3. Predicting the standard deviation in expert judgments

Parameter	Estimate	Standard Error
Fixed effects		
Constant	1.787**	0.327
Party Differentiation	–0.210*	0.104
Recent Member	0.054	0.101
Vote Percentage	–0.004	0.004
Salience	–0.251**	0.068
Dissent	0.306**	0.065
Left-Right Ideology	–0.027	0.028
Member of Government	–0.036	0.087
Radical Right	–0.241	0.173
Conservatives	–0.100	0.157
Liberals	–0.357**	0.128
Christian Democrats	–0.520**	0.142
Social Democrats	–0.508**	0.159
Radical Left	–0.266	0.164
Greens	–0.204	0.156
Regional Parties	–0.064	0.134
Number of Experts	0.010	0.013
Variance components		
National	0.005	0.007
Party	0.118**	0.015

Note: Table entries are full information maximum likelihood multilevel model estimates with their estimated standard errors. N = 143 country-party cases. –2 log likelihood = 105.895. ** $p < 0.01$, * $p < 0.05$ (two-sided).

speak to Budge's (2000) concerns only indirectly. It is time to consider experts in terms of an explicit measurement model.

A measurement model

How should we conceive of expert judgments? We believe that it is best to conceptualize experts as measurement instruments, much like items in an attitudinal scale. As such, experts can be evaluated using the same psychometric theory that is applied to attitudinal items. Here, we focus on a psychometric framework that derives from classical test theory, although one could think of alternative frameworks. To develop our framework further, let x_i denote the placement of a given party, in a given country, on a given dimension, at a given point in time by expert i (where $i = 1 \ldots n$ and n is the total number of experts for the country). We assume this placement reflects the true party position, which we indicate by T.[9] However, x_i is not a perfect measurement instrument. There is therefore an error component, ε_i, which reflects the expert's perceptual distortion. Combining the true score and error components we have

$$x_i = \lambda_i T + \varepsilon_i$$

This model is known as the 'congeneric test model' (Crocker & Algina 1986; Jöreskog 1971). It is assumed that $E[\varepsilon_i] = 0$, $E[\varepsilon_i \varepsilon_j] = 0$ (for $i \neq j$) and $E[\varepsilon_i T] = 0$ (i.e., there is no systematic measurement error that replicates across experts or is related to the true party placement). In this case, the standardized λ_i^2 may be interpreted as the reliability of expert i as an indicator of the party position.

The congeneric model has important implications for the correlations between expert judgments that we should observe. Let us assume, without loss of generality, that x_i and T are standardized. Then the correlation between the judgments of experts i and j is given by $\rho_{ij} = \lambda_i \lambda_j$. This result does not mean much by itself, but it forms an important counter-point to a situation in which the experts wind up measuring different traits. For example, imagine that expert i measured T, as was intended by the researcher. Expert j, however, misinterpreted the question and judged another trait with true score M. Now the correlation between the two experts is given by $\rho_{ij} = \lambda_i \lambda_j \rho_{TM}$, where ρ_{TM} is the correlation between the two true scores. Unless the two traits are perfectly correlated it follows that $\lambda_i \lambda_j > \lambda_i \lambda_j \rho_{TM}$, assuming $\lambda_i > 0$ and $\lambda_j > 0$. In other words, the correlation between the expert judgments should be greater, perhaps much greater, when they are judging the same trait than when they are judging different traits.

The congeneric test framework allows us to consider each of Budge's (2000) concerns. Specifically, compared with T

- M may measure a completely different attribute (e.g., social as opposed to economic ideology).
- M may measure the same attribute but in a different segment of the party (e.g., the electorate instead of party elites).
- M may measure the same attribute but at a different point in time (e.g., economic ideology ten years ago as opposed to the present).
- M may measure rhetoric instead of behavior (or vice versa).

In each of these cases, the correlation between the judgments from two different experts should be attenuated and may even take a negative sign (depending on the sign of ρ_{TM}).

We can parlay the correlational implications of the congeneric test model into a variety of analytical strategies. First, we can obtain a reliability statistic such as the standardized item alpha. If the experts are indeed measuring the same trait for the same party segment at the same time, then the reliability should be comparatively high (unless measurement error is rampant). Second, we can compute similarity coefficients (Steenbergen 2000). These statistics compare the correlational patterns across experts. As Steenbergen has shown, the (pairwise) similarity between two items approaches one (the upper-limit) if those items measure the same true score. In the context of expert surveys, this implies that we should expect to see a high similarity coefficient between two experts if those experts indeed evaluate the same trait. Since the number of pairwise similarity coefficients increases rapidly with the number of experts, it is useful to summarize these coefficients. For this purpose, Steenbergen has developed a scalewise similarity coefficient, which is simply the average of all of the pairwise similarities. Since we expect the pairwise similarities to be high if experts evaluate the same trait, the scalewise similarity should also be high. Low values are thus an indication that the experts may not all be evaluating the same trait.[10]

Note that both the reliability analysis and the similarity coefficients require that the number of parties is greater or equal to the number of experts. Otherwise linear dependencies in the data matrix cause statistical artifacts in the correlations among experts, which may seriously distort the results.[11] In our dataset, there are eight countries for which meaningful reliability and similarity analyses can be performed: Austria, Belgium, Denmark, Finland, France, Greece, Ireland and Italy. The results of these analyses are shown in Table 4, using Pearson and Spearman rank correlations, respectively. The analysis based on Pearson correlations treats the expert judgments as cardinal data.

The analysis based on Spearman correlations considers these judgments to be ordinal only; the main concern here is whether experts rank parties in a similar order.

Overall these results are encouraging. The correlations among experts are on average very high, both within particular countries and across the whole set. This results in impressive reliabilities (as computed via the Spearman-Brown formula), a finding consistent with that reported by McDonald and Mendes (2001) and our earlier results from the variance components analysis.[12] The scale-wise similarities, too, are quite good (considering that the upper-limit on those similarities is 1 – see Steenbergen 2000). Moreover, when inspecting the correlations and their corresponding similarities, there appear to be no experts that are clear outliers (with the exception of Italy, which we discuss below). Such outliers would occur if a particular expert is highly inaccurate in his or her assessments of party positions or, more importantly, if he or she evaluated parties based on a different standard than that of other experts. Thus, there seems to be substantial convergence in the judgment criteria that experts use.

There are exceptions to these general patterns. As Table 4 shows, while the reliability of expert judgments on Italian parties is not absolutely poor, it is worse than that for other countries surveyed. This is primarily due to one expert (labeled E4 in Figure 1). The correlations between the judgments of this expert and the remaining five experts are low, suggesting that he or she may have used different criteria for placing Italian parties. When we compute

Table 4. Reliabilities and similarities of expert judgments

	Pearson			Spearman		
Country	r	Reliability	Similarity	r	Reliability	Similarity
Austria	0.957	0.991	1.000	0.881	0.974	0.997
Belgium	0.794	0.972	0.983	0.734	0.961	0.972
Denmark	0.983	0.998	1.000	0.970	0.996	1.000
Finland	0.836	0.962	0.996	0.855	0.967	0.996
France	0.812	0.968	0.995	0.783	0.962	0.991
Greece	0.888	0.979	0.995	0.897	0.981	0.996
Ireland	0.942	0.990	0.999	0.941	0.990	0.999
Italy	0.652	0.918	0.977	0.634	0.912	0.962
Mean	0.858	0.972	0.994	0.837	0.968	0.990

Note: Reliabilities computed via the Spearman-Brown prophecy formula. See Steenbergen (2000) for details about the computation of similarity coefficients.

similarity coefficients based on the Italian expert data, the anomalous character of E4 becomes quite evident. This expert's similarity with the other experts is about 0.05 lower than the average across the remaining experts for Italy. When a multidimensional scale (MDS) analysis is performed on the similarity coefficients (see Steenbergen 2000), we see that E4 stands out on the first dimension. We therefore have good grounds for doubting the validity of E4's responses, although it is worth emphasizing that even after including this expert in the computations for Table 4, the final results for Italy look quite good.

Discussion

Our results reveal remarkable agreement among experts about the placement of parties. This suggests that the experts, for the most part, used the same

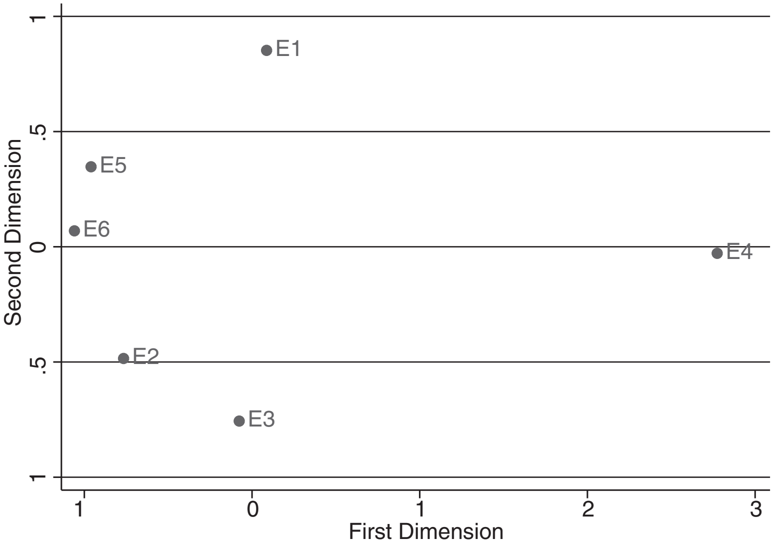

Figure 1. MDS of the similarities in expert judgments in Italy.
Note: Figure shows the derived stimulus configuration from the matrix of similarity coefficients of Italian experts. To obtain this figure, the similarity coefficients were first transformed into dissimilarities, which were then mapped onto distances using a classical multidimensional scaling model (for details, see Steenbergen 2000). The derived configuration in two dimensions had the best fit to the data; adding dimensions did not improve the fit sufficiently to warrant the loss of parsimony. The entries in the figure are the ratings of six Italian experts. The figure clearly shows that E4 (expert 4) is an outlier on the first dimension. The remaining experts cluster together, as they should if they measure the same party attribute.

criteria to judge parties. It is, of course, possible that they used criteria that ran counter to those intended by the developers of the survey, but this seems highly unlikely; it would have required that the experts overrode the guidelines provided by the developers and then miraculously all settled on the same alternative criteria. Thus, we can have considerable confidence in the expert data.

Experts versus alternative measures

The analysis so far has demonstrated impressive internal consistency among experts. Before we accept expert data, however, we should also demonstrate convergent validity with other measures of party positions – that is, the experts should provide evaluations of the same political phenomenon that other measurement instruments pick up. The convergence among instruments does not have to be perfect – each has unique qualities (Marks et al. forthcoming) – but if expert judgments diverged starkly from other measurement instruments, we would still question their validity.

Alternative measures

Several alternative measurement instruments exist for assessing party positions. First, we can extract party positions from party manifestos. The Manifesto Research Group has coded both favorable and unfavorable mentions of European integration in the manifestos of the most important parties in EU Member States (see Budge et al. 2001).[13] Second, we can use voter perceptions of party positions. The European Election Survey (EES) project asked respondents where various parties in their country stood on the issue of European integration (Eijk et al. 2002).[14] Such a question may be subject to projection effects – voters may project their own stance on European integration onto the party (Merrill et al. 2001). However, in the aggregate, such effects should cancel each other out so that average voter perceptions can be used to measure a party's position on European integration.[15] Finally, we can use members of parliament (MPs) and members of the European Parliament (MEPs) to gauge party support for European integration. Here, we rely on the survey conducted by Wessels et al. (1999) in 1996. We focus on MP and MEP perceptions of their party's position, rather than on their own position. In a sense, we are using the MPs and MEPs as experts to evaluate three components of integration: European currency, European employment program and national borders.[16] We average the evaluations of the MPs and MEPs where both are available.

Confirmatory factor analysis

The cross-validity of the expert survey can be gauged by the extent to which it measures the same underlying construct as the three alternative instruments. Thus, all measures should load on a single factor and the loading for the expert survey measure should not be significantly worse than that of other measures. If the expert survey measure were to load on a different factor, or if its loading was weak, then we would have grounds for concern about its convergent validity. A confirmatory factor model provides the best framework for studying the performance of the expert survey measure. The units of analysis are political parties. As measures of party positions, we use the positive and negative manifesto scores, as well as the average party placements by EES respondents, MP/MEP survey respondents and our own experts. In total, there are seven indicators: three MP/MEP indicators, two party manifesto indicators, and single indicators for the EES and our own expert survey. Because multiple indicators exist for two of the data sources, we decided to correlate the errors within each of those sources. The model was estimated using LISREL 8.54, using full information maximum likelihood estimation.

The standardized estimates of the factor model are shown in Table 5. The first thing to observe is that this model fits the data very well. The Chi^2 fit statistic is 11.064 with 10 degrees of freedom, resulting in a *p*-value that exceeds 0.05. In addition, the remaining fit statistics also suggest an excellent fit. Considering the loadings, we see that the expert survey loads heavily on the factor (the standardized loading is 0.970). If we consider the underlying factor as a true score, this loading translates into a true score reliability of 0.94, which is extremely high. The loading for the expert survey is comparable to that for the currency item from the MP/MEP survey and surpasses the other loadings. In all, these results suggest that the expert survey data converge very strongly with other measures of party positions vis-à-vis European integration.[17]

Discussion

The confirmatory factor analysis provides impressive evidence of convergent validity between our expert survey measures and alternative measures of party positions. These results further support the validity of our expert survey measure of party positions on European integration. While convergence with the other measurement instruments is not perfect – and should not be – these results corroborate the factor analytic results reported by Ray (1999). As was the case with his data, our expert survey measures seem to capture essentially the same information about party positions as other measures such as the party manifestos.

Conclusion

Reliance on expert judgments is an attractive option for measuring complex phenomena such as party positions about policies. Expert surveys are comparatively straightforward to conduct, but are they perhaps a little too straightforward to warrant confidence? How valid are expert judgments really? Ian Budge (2000) has raised a series of questions that cast doubt on the validity of expert surveys. Those who use expert surveys should take these questions seriously. In this article, we have evaluated one expert survey instrument: our own study of party positions on European integration. We have shown that there is good reason to trust our expert survey results. Not only did the experts

Table 5. Confirmatory factor model of alternative measures

Parameter	Estimate
Loadings	
Manifesto-Positive	0.439**
Manifesto-Negative	–0.734**
European Election Survey	0.690**
Expert Survey	0.970**
MP/MEP Survey-Currency	0.908**
MP/MEP Survey-Employment	0.369**
MP/MEP Survey-Borders	–0.651**
Error correlations	
Manifesto-Positive with Manifesto-Negative	–0.022
MP/MEP-Currency with MP/MEP-Employment	–0.064
MP/MEP-Currency with MP/MEP-Borders	–0.177**
MP/MEP-Employment with MP/MEP-Borders	0.300**
Fit statistics	
Chi^2	11.064
p	0.353
AGFI	0.865
NFI	0.964
CFI	0.996

Note: Table entries are standardized full information maximum likelihood confirmatory factor model estimates. N = 61. ** p < 0.01.

show remarkable consistency in their responses, but expert placements of political parties converge with other measures. Of course, we have evaluated only expert judgments of party *positions*; our results do not speak to judgments of internal dissent, salience (see Netjes & Binnema forthcoming) or other attributes of parties.

A second and related contribution of this article is to lay out some methodological tools that can help with the evaluation of expert surveys. The concerns that Budge (2000) has stated are to a considerable extent amenable to empirical investigation. We have illustrated the analysis of inter-expert variation, of the inter-correlation between expert judgments, and of the correlation of expert placements and other measures of party positions. Where possible we have rooted these analytical tools in an explicit measurement model that treats experts as indicators of true party positions. These tools are not unique to our expert survey; they may be applied to any expert survey instrument. Our hope is that these tools will be used in validity assessment – an exercise that should be a central part of expert survey methodology.

Acknowledgements

The authors would like to thank the participants in the Amsterdam Workshop on 'Comparing Data Sets on the Positioning of National Political Parties', which was sponsored by the Free University of Amsterdam in October 2004. They would also like to thank George Leckie for his statistical advice on the cross-level variance component model and Liesbet Hooghe for her invaluable contributions to all aspects of this project.

Notes

1. Elsewhere we argue that expert surveys may be best viewed as complementary to other data sources rather than a substitute (Marks et al. forthcoming; see also Laver 2001; Mair 2001; Ray 1999).
2. The question of what experts are judging when they render expert judgments is not the only concern raised about expert surveys. Evaluating left-right placements, Mair (2001) wonders whether the fact that party interactions have become less deterministic has made party placements less relevant. He also argues that experts may artificially increase the distance between parties if they have to judge many of them in the same system, and that expert judgments should not be treated as an alternative to manifestoes and other data about party positioning. Laver (2001) cautions that expert surveys should not be used to predict party behavior since the expert judgments themselves may reflect this behavior (see also Budge 2000). McDonald and Mendes (2001) are concerned that experts may be guided more by a party's membership of a party family than by real shifts

in party positioning and, as a consequence, expert ratings may overestimate stability over time. While these are important concerns, our focus is exclusively on the concerns raised by Budge (2000).

[3] The survey was conducted with the help of Carole Wilson and David Scott. The enterprise was funded by the Center for European Studies at the University of North Carolina, Chapel Hill, and the questionnaire, codebook and dataset are downloadable from: http://www.unc.edu/~gwmarks

[4] We have excluded Luxembourg from the list because the experts for this country are somewhat different than those for other countries, consisting of journalists and politicians rather than political scientists.

[5] This statistic is computed as $(\sigma_{\delta}^2 + \sigma_{\varepsilon_j}^2)/(\sigma_{\delta}^2 + \sigma_{\varepsilon_j}^2 + \sigma_{\varepsilon_i}^2)\cdot$

[6] The Spearman-Brown formula is given by $nr/[1 + (n - 1)r]$, where n is the average number of experts and r is the inter-expert correlation that is generated from the variance components model. For precedent on the use of this procedure for assessing inter-coder reliability, see Jayasinghe et al. (2003).

[7] The numbers of experts listed in Table 1 reflect the number of completed surveys that we received from each country. However, not all experts provided estimates of the European integration stance for all parties. Thus, for certain parties, the effective sample size may be smaller than that given in Table 1.

[8] Ideology is not statistically significant. We re-estimated a model that included this predictor both in linear and quadratic form. This model sought to capture the possibility that experts may be least consensual for parties located at the ideological extremes. However, no significant effects were found for either ideology term, not even after the party family dummy variables were dropped from the model. Apparently, neither ideology nor ideological extremity played a role in inter-expert agreement.

[9] Technically, this equates a true score with a construct. This is perilous because true scores can reflect systematic method effects as well as traits (see Saris & Andrews 1991). While an ideal design would separate the trait and methods effects, we lack the data to do so. Instead we simply assume that any method effects are relatively weak and overshadowed by the trait component.

[10] A third approach would consist of performing a confirmatory factor analysis. However, since the number of parties is generally small, such an analysis will generally contain more parameters than cases. This could cause estimation problems.

[11] The technical reason for the requirement is that the analysis treats parties as the rows of the data matrix and experts as the columns. For this matrix to be full rank – a requirement for computing the cross-expert correlations – the number of parties should exceed the number of experts. Otherwise, the rank of the matrix at most will be equal to the number of parties, which means that some of the columns will be linearly dependent on other columns.

[12] The current analysis uses a different approach to estimating reliability – one that does not rely on an estimate of the inter-expert correlation from a variance component model, but instead evaluates this correlation within a classical test score model. Moreover, rather than estimating the correlation across all countries, the current estimates are country specific. However, to the extent that the classical test model cannot be applied because there are more experts than parties, the variance components approach is a valuable substitute.

[13] These measures are per108 and per110 (see Budge et al. 2001).

14. The precise question was: 'Some say European unification should be pushed further. Others say it already has gone too far. What is your opinion? Please indicate your views using a 10-point scale. On this scale, 1 means unification "has already gone too far" and 10 means it "should be pushed further". . . . And about where would you place the views of the following parties on this scale?' (Eijk et al. 2002).
15. The notion that perceptual biases cancel in the aggregate is a cornerstone of macro-political studies (see Erikson et al. 2002). It is also a common solution to projection effects in electoral research (see Macdonald et al. 1991).
16. The question wordings for these items are as follows: (1) 'Should [country] keep its [national currency] and make it more independent from the other European currencies, or should the aim be a new common European currency? Please indicate on the scale what you see as your national party's position.' (2) 'The former president of the European Commission, Jacques Delors, has proposed to raise funds for a massive programme to fight unemployment. Others argue that the completion of the Single European Market alone will be the best remedy for unemployment. Please indicate again on the scale what you see as your national party's position.' (3) 'Should the EU continue to remove national border controls or should tighter border controls be reintroduced to fight crime effectively? Please indicate on the scale what you see as your national party's position' (Wessels et al. 1999).
17. To verify this result, we also ran a two-factor model in which the manifesto, European Election Survey and MP-MEP measures loaded on one factor and the expert survey measure loaded on another. Allowing for correlated errors, this model is actually equivalent in fit to the model reported in Table 5, and a key result is the very high correlation (0.97) between the two factors. These results should not come as a surprise; the two-factor model is simply an alternative parameterization of the one-factor model. Instead of letting the loading for the expert survey vary, we have now fixed it to one (in order to identify the scale of the second factor). This loading is now absorbed into the factor correlation.

References

Alvarez, R.M. (1997). *Information and elections*. Ann Arbor, MI: University of Michigan Press.

Alvarez, R.M. & Brehm, J. (2002). *Hard choices, easy answers: Values, information and American public opinion*. Princeton, NJ: Princeton University Press.

Andrews, F.M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly* 48(2): 409–442.

Budge, I. (2000). Expert judgments of party policy positions: Uses and limitations in political research. *European Journal of Political Research* 37(1): 103–113.

Budge, I. et al. (eds) (2001). *Mapping policy preferences: Estimates for parties, electors and governments, 1945–1998*. Oxford: Oxford University Press.

Castles, F.G. & Mair, P. (1984). Left-right political scales: Some 'expert' judgments. *European Journal of Political Research* 12(2): 73–88.

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Eijk, C. van der, et al. (2002). *European election studies, 1999: Design, implementation and results*. Amsterdam: Steinmetz Archive.

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

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Huber, J. & Inglehart, R. (1995). Expert interpretations of party space and party locations in 42 societies. *Party Politics* 1(1): 73–111.

Jayasinghe, U.W., Marsh, H.W. & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society* 166: 279–300.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* 36(2): 109–133.

Key, V.O. (1964). *Politics, parties and pressure groups* (5th edn). New York: Crowell.

Laver, M. (2001). How should we estimate the policy positions of political actors? In M. Laver (ed.), *Estimating the policy positions of political actors*. London: Routledge.

Laver, M. & Hunt, W.B. (1992). *Policy and party competition*. New York: Routledge.

Laver, M. & Mair, P. (1999). Party policy and cabinet portfolios in the Netherlands, 1998: Results from an expert survey. *Acta Politica* 34(1): 49–66.

Macdonald, S.E., Listhaug, O. & Rabinowitz, G. (1991). Issues and party support in multiparty systems. *American Political Science Review* 85(4): 1110–1131.

Mair, P. (2001). Searching for the position of political actors: A review of approaches and a critical evaluation of expert surveys. In M. Laver (ed.), *Estimating the policy positions of political actors*. London: Routledge.

Marks, G. et al. (forthcoming). Cross-validating data on party positioning on European integration. *Electoral Studies*.

McDonald, M.D. & Mendes, S.M. (2001). The policy space of party manifestos. In M. Laver (ed.), *Estimating the policy positions of political actors*. London: Routledge.

Merrill, S., Grofman, B. & Adams, J. (2001). Assimilation and contrast effects in voter projections of party locations: Evidence from Norway, France and the USA. *European Journal of Political Research* 40(2): 199–223.

Meyer, M. & Booker, J. (1991). *Eliciting and analyzing expert judgment: A practical guide*. London: Academic Press.

Netjes, C.E. & Binnema, H. (forthcoming). The salience of the EU integration issue: Party manifesto and expert survey data compared. *Electoral Studies*.

Rasbash, J. et al. (2004). *A user's guide to MLwiN, version 2.0*. London: Centre for Multilevel Modelling, Institute of Education, University of London.

Ray, L. (1999). Measuring party orientations toward European integration: Results from an expert survey. *European Journal of Political Research* 36(2): 283–306.

Saris, W.E. & Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P.B. Biemer et al. (eds), *Measurement errors in surveys*. New York: Wiley.

Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance components*. New York: Wiley.

Steenbergen, M.R. (2000). Item similarity in scale analysis. *Political Analysis* 8(3): 261–283.

Steenbergen, M.R. & Jones, B.S. (2002). Modeling multilevel data structures. *American Journal of Political Science* 46(1): 218–237.

Steenbergen, M.R. & Scott, D.J. (2004). Contesting Europe? The salience of European integration as a party issue. In G. Marks & M.R. Steenbergen (eds), *European integration and political conflict*. Cambridge: Cambridge University Press.

Thomassen, J.J.A., Noury, A.G. & Voeten, E. (2004). Political competition in the European Parliament: Evidence from roll call and survey analyses. In G. Marks &

M.R. Steenbergen(eds), *European integration and political conflict*. Cambridge: Cambridge University Press.

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Wessels, B., Kielhorn, A. & Thomassen, J.J.A. (1999). *Political representation in Europe: European Members of Parliament study*. Berlin: Wissenschaftszentrum Berlin.

Correspondence to: Professor Marco Steenbergen, Department of Political Science, University of North Carolina, Chapel Hill, NC 27599, USA. E-mail: msteenbe@email.unc.edu