Available online at www.sciencedirect.com
ScienceDirect

Electoral Studies 26 (2007) 1–10

Electoral Studies

www.elsevier.com/locate/electstud

Editorial

Introduction: Triangulation and the square-root law

Abstract

This special issue of *Electoral Studies* evaluates the validity of expert, manifesto, and survey data on the positioning of national political parties. My purpose in this introduction is to examine the logic and limits of triangulation. In doing so, I make explicit the most important lesson of this special issue: improving the validity demands that one take full advantage of the information that is available. Even if observation is inherently biased, one can improve accuracy by comparing the observations that are biased in different ways.

Keywords: Expert surveys; Measurement; Political parties; Validity

> ‘We assume that all methods, including the experimental method that we so admire, are fallible; none provides a royal road to truth. The proper response to this inescapable predicament is to pursue research questions from a variety of methodological angles, all of them fallible, but fallible in different ways. Dependable knowledge has its base in no single method, but rather in triangulation across multiple methods’ (Kinder and Palfrey, 1993, 3).

> ‘Triangulation involves data collected at different places, sources, times, levels of analysis, or perspectives, data that might be quantitative, or might involve intensive interviews or thick historical description. The best method should be chosen for each data. But more data are better. Triangulation, then, is another word for referring to the practice of increasing the amount of information to bear on a theory or hypothesis’ (King et al., 1995, 479–480).

1. Introduction

Plato believed that knowledge of the empirical world is akin to interpreting flickering shadows cast on the wall of a cave by objects placed in front of a fire. The allegory of the cave is a frontal attack on empirical knowledge, yet it is based on an insight that no scientist can ignore. Empirical knowledge is inherently error-prone, and it is error-prone in ways that are impossible to specify with certainty. An observation cannot provide information about the extent to which it is biased; a set of observations cannot provide information about the extent to which they suffer a common bias. This is the dilemma of

0261-3794/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.electstud.2006.03.009

measurement. No science is immune to it, and social science is particularly susceptible.[1]

Social scientists are concerned with concepts that are far removed from objectively measurable facets of human existence. Power, identity, political preferences, including the topic of this special issue—the positions adopted by political parties—are typical social science concepts in that they cannot be measured directly. We perceive them indirectly, and our attempts to measure them depend on inferences.[2]

The most obvious way to increase accuracy is by developing better measurement tools. As a child, I remember doing calculations with a slide rule where accuracy depended on how carefully one aligned ruler and plastic casing and how carefully one read off the arithmetic result. There is no substitute for attentiveness in conducting experiments,[3] but no matter how intently I focused, I could not ascertain tenths of the distance between two consecutive ticks along the ruler. The accuracy of observation is constrained by the technology of measurement, in this case the slide rule.[4]

An alternative way to increase accuracy is to increase the *volume* of information, that is the number of observations of the phenomenon of interest, and this is what I shall discuss here. This can be done by replication, for example, by increasing the number of cases in a sample, or it can be done by comparing the datasets that contain observations of the same case or cases. A single logic underlies both strategies, as I argue below.

This special issue of *Electoral Studies* contains papers presented at a workshop that Liesbet Hooghe, Hans Keman, and I organized at the Free University of Amsterdam in 2004. We invited some two dozen researchers who have systematically measured policy positions of political parties and we asked each of them to compare their data to that collected by others. Our motive was practical: data on the positioning of political parties are vital in evaluating hypotheses on structures of democratic competition and conflict, on the interplay between electorates and political parties, or on how public policy is shaped by political parties with different agendas. But our enterprise also engages basic questions of reliability and validity. My purpose here is to introduce these issues in a general and accessible way. In doing so, I make explicit the most important lesson of this special issue: improving the validity demands

[1] The argument in this introduction complements Lakatos' proposed solution to the dilemma of scientific falsification. Lakatos' philosophy of scientific method is an attempt to bridge the following claims: (1) science is falsifiable knowledge; (2) a theory is a set of logically consistent guesses about the world which can be falsified by some empirically conceivable set of facts (Feynman, 1965); (3) an experiment is a procedure that creates a fact; and (4) no experiment can plausibly disconfirm a given theory. Lakatos' solution is to draw on the history of physics to recognize that (dis)confirmation rarely takes place as a result of a single experiment, but results, instead, from a process of experiment and counter-experiment in which observations are interpreted and reinterpreted. Even an experiment as simple as testing the tensile strength of a thread by placing an iron weight on it cannot produce observations capable of irrefutably disconfirming a hypothesis (1970, 184ff). Perhaps, Lakatos asks, a magnet or some hithero unknown force in the ceiling exerted effected the pull of the iron weight; perhaps the tensile strength of the thread depends on how moist it is; perhaps the scale for the iron weight was wrong; perhaps the thread did not break, but was only observed to break; perhaps the thread was not a thread, but a 'superthread' with special properties. The scope for theoretical adjustment to cope with new evidence is endless. Facts to not speak objectively to theories, but are themselves theoretically impregnated. Rational (dis)confirmation involves mutual cross-examination between experiments and a scientific program. What is decisive, according to Lakatos, is *how* a scientific program predicts, interprets, or adjusts to new facts, or fails to do so. The argument in this introduction probes the inherent uncertainty in observation and elaborates a strategy—triangulation—that is logically consistent with Lakatos' theory of scientific method.

[2] This implies variation in the degree to which facts are theoretically embedded (and therefore disputable). Some observations, while in principle disputable, are theoretically robust. My observation that 'there are now 38 chairs in this coffee bar' could be disputed by questioning my eyesight, my arithmetic or my definition of a chair, but this observation is far closer to the ground than my observation that in 2002, 38 political parties in EU countries were skeptical of European integration or that the British Liberal party is today to the left of the Labour party. Dispute about the number of chairs would focus on reliability: 'How carefully did you count?' 'Did you check your result?' Debate about the positioning of political parties engages issues of conceptualization, operationalization, and bias, in addition to reliability. In this introduction I am concerned with conceptualization and operationalization only insofar as they are manifested in bias or systematic error (on conceptualization and operationalization see Adcock and Collier, 2001).

[3] I use the term 'experiment' in its broad sense to describe the process by which an observation is generated (Wackerly et al., 1996; Jacoby, 1999; Lakatos, 1970). The experiment may be an attempt to measure the association between temperature and the expansion of a copper rod in a laboratory (classical experiment), an attempt to measure the positioning of a political party (natural experiment), or the response of a randomly selected citizen to an item on a mass survey (quasi-experiment). The observations resulting from both controlled experiments and natural or quasi-experiments have fundamental commonalities: they are in principle (and often in practice) scientifically *questionable*, *theory-driven*, and, most importantly, they produce *imprecise* estimates of true values. The Latin root of 'experiment' is experiri, 'from trying.'

[4] Technology of measurement is an engine of scientific progress. One need to only consider the influence of the microscope or telescope to appreciate this for physical science. Arguably, the same is true for social science. Advances in political science have followed breakthroughs in the measurement of public opinion (the social survey), economic indicators, and, most recently, documents (computerized coding).

that one take full advantage of the information that is available. Even if observation is inherently biased, one can improve accuracy by comparing observations that are biased in different ways.[5]

2. Replication and triangulation

There are two ways to increase the volume of information: one can repeat an observation that one has already made, trying to keep all relevant conditions the same, or one can observe from a different angle, using a different method. The first is replication; the second, triangulation.[6]

Replication is efficient in gaining precision to the extent that error is random. Randomness is the most delicate of qualities, and scientists go to extreme lengths to approximate it when, for example, they design surveys, samples, or coding algorithms. But an experimental design, no matter how sophisticated, is an *n* of one. How can one be sure that a method is valid? In other words, what if measurement error is not random but systematic? We are back to Plato's dilemma. A strategy to minimize inaccuracy due to systematic error—and the one taken in this special issue—is to triangulate, to compare observations derived from *different* experimental designs. The virtue of triangulation is that it does not require that systematic error be eliminated, but only that systematic errors differ across measurement instruments.

The argument rests on one absolutely essential claim: diverse methods produce diverse biases. The strong version of this claim is that these biases will behave as if they were random. That is to say, systematic error will tend toward zero when averaged across methods. The weak version of this claim is that observations generated by diverse methods will not be biased in the same way, and hence, one can reduce, if not eliminate, systematic error by triangulating. This assumption really is weak, for if systematic error was uniform across our methods of observation, we would live in a ''dupes' world,'' a world in which we were not merely error-prone, but prone to making the same error.

The next section of this article sets out a square-root law which specifies an outer limit for informational accuracy as a function of the volume of information. The square-root law is familiar in sampling theory. It applies to information in general, including that generated by triangulation, and I provide a formal basis for this in the following section.

3. The square-root law of information

How is the accuracy of information related to its volume? Let us for the moment assume pure efficiency of information generation (i.e. random sampling). Under this condition there is an outer limit to the relationship between informational accuracy and volume which is invariant and law-like: The accuracy of information is equal to the square-root of the volume of information.

$$\text{accuracy of information} = \sqrt{\text{volume of information}}$$

Accuracy is equivalent to validity, defined, for a single case, as the ratio of the estimated value to the actual value, or, for a set of cases, as the proportion of variance shared between the observed values and the dimension of interest to us. Volume of information refers to the number of times each case is observed.

The square-root law is grounded in the basic statement of sampling theory that the precision of the sample average improves with the square-root of the sample size.[7] The standard deviation of the normal distribution that describes the behavior of the mean of observations is equal to the standard deviation of the individual observations, σ, divided by the square-root of the sample size,

$$\text{the standard error of the mean} = \frac{\sigma}{\sqrt{n}}$$

Since σ is a constant, we can say that the standard error of the mean decreases with the square-root of n.[8]

[5] For an excellent example of cross-method collaboration, see Klandermans and Staggenborg (2002).

[6] The term triangulation often refers to the combined use of quantitative and qualitative analysis to achieve more plausible inference (Tarrow, 1995; Adcock and Collier, 2001). In this article, I use the term to refer to combining dissimilar sources of information to enhance validity of measurement. My comments and analysis focus on quantitative data, but the logic is perfectly compatible with qualitative analysis.

[7] I use the term precision to refer to reliability, i.e. the extent to which an observation is consistent in repeated trials, and I use the term accuracy to refer to validity, i.e. the extent to which an observation corresponds to the actual value of a case on the dimension one wishes to measure. Validity engages issues of conceptualization and operationalization—important issues that I set aside here (see Adcock and Collier, 2001).

[8] The square-root principle appears to underlie many sources of information. The extent to which an object shapes (i.e. informs) space at a particular location is inversely proportional to the square-root of the distance. Likewise, the intensity of a source of heat increases with the square-root of its proximity. The square-root law applies to the amplitude of noise and the intensity of light. The law operates at every known scale of these phenomena. For heat, light, and noise, the square-root law describes an outer limit that is approximated under the ideal condition of a vacuum.

The square-root law describes two imaginary worlds where error is characterized by randomness. The first is the world of classical test theory, in which all error is random error. Here, there is no distinction between reliability, the extent to which an observation is consistent in repeated trials, and validity, the extent to which an observation approximates the actual value of the case on the dimension that one wishes to measure. In such a world, replication diminishes random error which is the only error there is. Hence, replication produces precision, which is equivalent to accuracy.

The second world is one in which there is systematic error arising from the method by which one makes observations, but this systematic error is random across the range of possible methods. In this world, a distinction can be made between reliability and validity. Simply repeating the same experiment improves reliability, but does nothing to reduce the systematic error. However, using diverse methods to generate observations increases validity in the same way that replication increases reliability.

These scenarios are merely thought experiments that reveal how triangulation would work if method effects were random. On average, as the square-root law informs us, one would be able to double the accuracy of measurement if one relied on four methods for collecting data, instead of a single method.

Of course, one could do better by selecting a second method in which systematic error was the reverse of the first method, so that the errors cancel. However, this demands much of the world and our knowledge of it. First, the method effects would need to be mirror images of each other. Because method effects are rarely constant across the cases in a dataset, the method effects of the two datasets would have to line up, in opposite directions, case by case. Second, one would still have to be confident that average scores across the two datasets are valid scores. There is a paradox here: if one knew what the valid scores were, why would one combine two datasets in order to approximate them? In order to pick the two complementary datasets, one would need to estimate method effects across a minimum of three datasets on the assumption that the systematic component common to all the three is the closest one can come to validity.

The assumption that systematic errors arising from method effects are random is consistent with the lack of a criterion measure (Ray, this issue) which would allow one to accurately estimate and manipulate method effects. The square-root law describes a situation where the researcher has no control over the errors of successive observations, but where reliability (as a result of replication) or validity (as a result of triangulation) is improved as a diminishing function of the number of observations.

The application of the square-root law to triangulation demands some logical explication, which I provide below.[9]

4. Systematic error

All scientific observation is subject to systematic error. Measurement in social science is particularly prone on account of its conceptual abstractness and the corresponding difficulty of precise operationalization. As this special issue makes clear, the study of party positioning is no exception.

Replication produces accuracy only to the extent that error is random; it cannot diminish systematic error. The effect of systematic error can be understood intuitively. Imagine a dart player who aims at the bulls-eye of a dart board. Assuming that he misses randomly, the efficiency of each additional dart in producing a mean that hones in on the bulls-eye is given by the square-root law. However, to the extent that each throw is biased in one direction because of some external condition, say a biased set of darts or a down-draft, then the reduction of random error reduces the variance of the mean around the wrong center.

The point of departure of this special issue—and of recent advances in measurement theory (Bollen and Paxton, 2000; Saris and Andrews, 1991; Saris et al., 2004; Suen, 1990)—is the recognition that triangulation can be effective under conditions of bias. If bias varies across datasets, then adding a dataset can improve accuracy no less effectively than replication improves precision.

The simplest way of thinking about this is to view measurement from the standpoint of the individual case, let us say, the British Conservative party. Systematic error is produced by the method used to create a particular dataset, and therefore biases each observation in that dataset, including that of the Conservative party. When we add a second dataset, we are adding a second observation of the Conservative party, and when we add a third dataset, we add a third observation of the same

[9] Randomness is defined by mathematicians as *non-compressibility* of information (Chaitin, 1975). That is to say, a series of numbers is random when it cannot be generated by an algorithm with smaller informational content than the series itself. Randomness is difficult to achieve, even for applied mathematicians who design computer programs. But as elusive as randomness is, it is a fundamental goal of measurement. If all error in one's information is random, the non-random portion—the structured part—represents what one wishes to measure.

party. It is sensible to believe that each observation is biased, but the question is, in which direction? If the observations are derived from different methods, it is plausible to believe that the biases will be dissimilar. The following section formalizes this intuition and assimilates the logic of triangulation to the logic of replication.

5. The logic of triangulation[10]

5.1. The classical case

We begin with classical test theory, which assumes that an observation, X, is equal to its systematic component—the true score, T—plus random error, δ. Let X_i denote a particular measure of the position of party i

$$X_i = T_i + \delta_i \tag{1}$$

The measurement instrument which generates X_i can be any of those used in this special issue—an expert survey, or a measure based on party manifestos, roll-call data, or citizen responses to a survey of perceptions of party positioning.

A single dataset is composed of a series of observations, where X_{ik} is party i in dataset k

$$X_{1k}, X_{2k} \ldots X_{ik},$$

Each observation is equal to the true score, T, plus error.

$$\begin{aligned} X_{1k} &= T_1 + \delta_{1k} \\ X_{2k} &= T_2 + \delta_{2k} \\ &\vdots \\ X_{ik} &= T_i + \delta_{ik} \end{aligned} \tag{2}$$

Classical test theory makes the following assumptions about the structure of the error:

1. The expected (mean) error is zero,

$$E(\delta_{ik}) = 0 \quad \forall i$$

2. The correlation between true and error values is zero,

$$E[\delta_{ik} T_{ik}] = 0$$

3. The correlation between errors on distinct observations is zero,

$$E\left[\delta_{ik}\delta_{jk}\right] = 0 \quad \forall i \neq j$$

Under these assumptions measurement error is random. Hence, the expected value of each observation is the true score

$$E(X) = T \tag{3}$$

The critical assumption here is that the systematic component of X, which is the true score, T, is a valid measure of the actual position of the party. Hence, all that one must do to produce valid scores is to reduce random error by replicating observations. This can be achieved by using multiple indicators of the trait one wishes to measure, or tapping the expertise of multiple experts, or aggregating the responses of a random sample of citizens. If we assume that the error is random, the estimated score converges to the true score as a function of the error of each observation divided by the square-root of the number of observations (be they indicators, experts, or citizens sampled).

5.2. Systematic measurement error

It is sensible to assume, however, that measurement error has a systematic component alongside a random component. One can expect method effects because at least some of the observations in a dataset are systematically shaped by the particularities of the measurement instrument. For example, the response scale of an expert survey item, or the coding procedure of a party manifesto contribute to the variation in measured party positions, quite apart from the actual position that the party takes.

Following Saris and Andrews (1991), we consider a true score as nothing but the systematic component of an observation.[11] While the value we wish to measure is one source of systematic variation, so are method effects. We can therefore measure the true score of party i in dataset k as

$$T_{ik} = \alpha_{ik} Q_i + \beta_{ik} M_k \tag{4}$$

where

- Q_i is the score of party i on the dimension of interest
- M_k is the method factor that captures the influence of the particular measurement instrument generating dataset k

[10] This section is informed by discussion with Marco Steenbergen and Georg Vanberg and draws on Marco Steenbergen's (2005) unpublished note, 'Multi-Method Approaches to Measuring Party Positions.'

[11] Zeller and Carmines (1980) arrive at a mathematically equivalent formulation by partitioning error into its systematic and random component.

- The α_{ik} coefficients, standardized, are validity coefficients
- The β_{ik} coefficients, standardized, represent method effects.

Substituting equation (4) into (2)

$$X_{ik} = \alpha_{ik}Q_i + \beta_{ik}M_k + \delta_{ik} \tag{5}$$

When we relax the assumption that the non-random component of an observation represents the dimension of interest, and model systematic method effects, a distinction can be made between reliability and validity.

Reliability is defined as the proportion of true score variance to total variance (i.e. the proportion of non-random variance in the observed score), once these variables are standardized.

$$\text{reliability} = \frac{V[T_i]}{V[X_{ik}]}$$

Validity takes into account the extent to which the non-random part of observed variance is due to the method effect, and is defined as the proportion of variance that the observed scores share with the dimension of interest, Q_i.

$$\text{validity} = \alpha_{ik}^2$$

Reliability can be assessed in the context of a single method. If observations using a particular measurement instrument are replicated, one of the two unknowns in equation (2), namely the random error, δ_{ik}, diminishes to zero.

Validity, by contrast, cannot be assessed in the context of a single method. This is the dilemma that Plato identifies: one cannot determine whether the shadows on the wall are accurate representations—except by gaining a different perspective. When we have only a single measure of a party position, it is impossible to tease apart Q and M. We know, however, that T reflects, to some degree, method effects since we have access to only one method of assessing party positions. Methodologists sometimes call this mono-method bias and it should be taken seriously. For example, when replicating a measurement instrument one may see a high degree of correlation between the test and retest simply because of their common M component. Replication is ineffective in separating Q and M.

5.3. The virtue (and limit) of triangulation

To estimate the method factor in equation (5) one must triangulate. In psychometrics this is typically done in a multi-trait, multi-method (MTMM) framework, where one compares three traits, each of which is measured with three different methods, yielding a nine by nine correlation matrix (Saris and Andrews, 1991). The data requirements for this framework are exacting, and methodologists have explored several ways to work around them (Bollen and Paxton, 1998; Saris et al., 2004).

The principle—and also the limits—of triangulation can be seen by extending the argument above to a multi-method setting. Consider a series of 1 to k datasets derived from k methods. Each dataset is composed of n cases, where X_{ik} is party i in dataset k.

$$\begin{array}{llcl}
\text{dataset 1} & \text{dataset 2} & & \text{dataset } k \\
X_{11} & X_{12} & \cdots & X_{1k} \\
X_{21} & X_{22} & \cdots & X_{2k} \\
\vdots & \vdots & & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nk}
\end{array}$$

We can re-arrange these datasets as a series of observations for the political parties the datasets have in common. The key to the argument is that systematic error due to method effects is a function of the error structure *within* dataset k. So when we re-arrange our datasets to focus on case 1, we are combining cases with *different* systematic errors. So for case 1 (say the British Conservative party) we have:

$$\begin{array}{l}
X_{11} = \alpha_{11}Q_1 + \beta_{11}M_1 + \delta_{11} \\
X_{12} = \alpha_{12}Q_1 + \beta_{12}M_2 + \delta_{12} \\
\vdots \\
X_{1k} = \alpha_{1k}Q_1 + \beta_{1k}M_k + \delta_{1k}
\end{array}$$

Given that the errors, $\delta_{11}\ldots\delta_{1k}$, are random, their expected value will approximate zero as the number of observations increases. The accuracy of our estimate of Q_1 depends on the structure of the method effects of the 1 to k measures, $M_1\ldots M_k$. To the extent that the method effects are correlated, we cannot partition them from Q_1. So $M_1 \cap M_2 \cap M_3 \ldots \cap M_k$ is a *hidden method factor* which is confounded with our estimate of the dimension we wish to observe. However, where there are diverse methods, as in the measurement of party positions, there is little reason to believe that method effects coincide. On the assumption that the population of possible method effects is distributed around zero, we can expect that the hidden method factor will diminish as one combines datasets based on different methods.

6. Strategies of measurement

While authors of this special issue disagree about the relative virtues (and vices) of particular datasets, we implicitly agree that our disagreements can only be resolved by triangulation. Each article in this special issue takes some subset of the datasets on party positioning and attempts to make comparisons among them (as detailed in Appendix A). We make headway not because the datasets we have created are bias-free, but because the biases they suffer are diverse.

The logics of the square-root law and of triangulation have some general implications for measurement strategy:

(1) The smaller the volume of information for any set of cases, the greater the benefit of increasing it. This is a basic implication of sampling theory and the square-root law. When we are dealing with triangulation, we enter a world of small-*n* where the marginal increase in accuracy gained by an additional observation (i.e. an additional dataset based on a different method) is quite large. On average, the standard error of the mean will decrease by 29.3 percent with the addition of a second dataset of the same accuracy as the first. Four datasets based on different methods cut the mean standard error in half. Such gains in accuracy take place irrespective of the degree to which datasets are biased, so long as the bias varies.

(2) The more imprecise an observation, the greater the benefit of an additional observation, *even if it is no less imprecise*. The denominator for the square-root law is the standard deviation of an individual observation. The greater this standard deviation, the greater the scope for absolute improvement of the standard error of the mean.[12]

(3) The more biased a dataset, the greater the benefit of an additional dataset having a different bias, *even if the additional dataset is no less inaccurate*. The logic is the same as for (2) above. In short, additional information is most useful in information-poor environments.

(4) The greater the diversity of systematic error among datasets, the greater the benefit of triangulation. The fundamental challenge in measurement is that bias is undetectable when it is shared. This is Plato's conundrum. One should therefore *experiment* with the method underlying one's observations in order to evaluate method effects. The implication is that one should seek diverse sources of information—i.e. information derived from contrasting methods—and which, as a result, are likely to suffer different kinds of systematic error. The reason for this follows from all that has been written above: diversity of systematic error across datasets is the logical equivalent of random error across individual observations.

Replication increases validity to the extent that error is random; to the extent that error is systematic, the only path to validity is to reduce systematic error, and this demands that one use a more accurate method—or that one compare methods.

7. Structure of the special issue

The articles in this issue are concerned with measurement of the positioning of political parties on three variables:

- Left/right (McDonald, Mendes, Kim; Benoit and Laver; Budge and Pennings; Volkens; Keman).
- European integration (Marks, Hooghe, Steenbergen, and Bakker; Ray; Whitefield et al.).
- Salience of European integration (Netjes and Binnema).

Several kinds of data exist for each of these variables. In this issue we use the following:

- Expert data: information drawn from the responses of designated experts on the topic to a set of survey questions.
- Electoral manifesto data: information drawn from written statements of political parties summarizing their policy commitments prior to elections.
- Surveys of legislators: surveys of parliamentary representatives of political parties.
- Mass survey data: surveys of citizens with questions relating to the positions of political parties.
- Roll-call data aggregating the votes of legislators.

Appendix A lists the datasets used in this special issue and, most importantly, where readers can access these data.

The reader who wishes to know which dataset provides the most precise measurements of party positioning will be disappointed. Given the fundamental character of measurement, as outlined above, this should not be surprising. The greatest gains in accuracy are not reaped by predetermining the best dataset, but rather from triangulating datasets which suffer from different kinds of bias. We are condemned to live in Plato's cave, but perhaps we can combine half-truths to observe more accurately.

[12] To give a simple arithmetic illustration, if a single observation is 75 percent accurate, then four replicating observations will provide a mean that is 87.5 percent accurate, 16 observations will be 93.75 percent accurate, and 10,000 observations will be 99.75 percent accurate. If a single observation is 99 percent accurate, then the mean of four observations will be 99.5 percent accurate, and that of 10,000 observations will be 99.99 accurate.

Acknowledgement

I would like to thank Elinor Scarborough for guiding this special issue to publication and Anna Brigevich and Jan Rovny for editorial assistance. I am grateful to Lawrence Ezrow, Liesbet Hooghe, Marco Steenbergen, Jim Stimson, and Georg Vanberg for detailed criticisms and advice without which I could not have pursued this topic. I alone am responsible for systematic (or random) errors.

Appendix A

Authors	Party manifesto data (hand- or computer coded)	Expert survey data	Public opinion data	Elite data
Ray	Comparative Manifesto Project, hand-coded, on European integration	Ray expert survey on European integration (Ray, 1999)	• Self-placement from Eurobarometer 29, 30 • Placement of party positions from Eurobarometer 30	Roll-call votes in European Parliament 1979–2001 (Nominate data: Hix et al., 2005)
Time and cases	Election nearest to 1988 for 15 EU countries	1988 for 15 EU countries	1988 for 15 EU countries	Parliament 1984–1989 for 15 EU countries
Marks et al.	Comparative Manifesto Project, hand-coded, on European integration	Marks and Steenbergen (2006) expert survey on European integration	Placement of party positions from 1999 European Election Survey (van der Eijk et al., 2002)	MP/MEP placement of their party on European integration (Katz et al.)
Time and cases	Latest election on CD-Rom (1998 or earlier) for 12 EU countries	1999 for 12 EU countries	1999 for 12 EU countries	1996 for 12 EU countries
Netjes and Binnema	Comparative Manifesto Project, hand-coded, on European integration	Marks and Steenbergen, 1999 expert survey on European integration	European Election Survey	
Time and cases	Latest election on CD-Rom (1998 or earlier) for 14 countries	1999 for 14 EU countries	1999 for 14 EU countries	
McDonald et al.	Comparative Manifesto Project 1945–1998, hand-coded (Budge et al., 2001), on left/right	Left/right surveys: • Castles and Mair (1984) • Laver and Hunt (1992) • Huber and Inglehart (1995)		
Time and cases	1972–1998 for 17 countries	1984, 1992, 1995 for 17 countries		
Keman	Comparative Manifesto Project, hand-coded, on left/right and progressive/conservative (Keman and Pennings, 2004)	• Castles and Mair (1984) • Huber and Inglehart (1995) • Marks and Steenbergen, on (economic) left/right and gal/tan		
Time and cases	1981–1998, for 18 western countries	1984, 1995, 1999		

Appenix A *(continued)*

Authors	Party manifesto data (hand- or computer coded)	Expert survey data	Public opinion data	Elite data
Benoit and Laver	Comparative Manifesto Project, hand-coded, on left/right	Benoit and Laver (in press) expert survey on left/right		
Time and cases	Latest election on CD-Rom (1998 or earlier) for 23 countries	2002 for 23 countries		
Whitefield et al.		• Rohrschneider and Whitefield, 2003 expert survey • 2002 Chapel Hill expert survey, on European integration		
Time and cases		2002–2003, for 9 CEE countries		
Volkens	• Comparative Manifesto project, hand-coded • Computerized word scores, on left/right	Castles and Mair (1984); Laver and Hunt (1992); Huber and Inglehart (1995)		
Time and cases	NA, literature review	NA, literature review		
Budge/ Pennings	• Comparative Manifesto Project, hand-coded • Computerized word-score estimates of manifestos (Laver et al., 2003), on left/right	Castles and Mair (1984)		
Time and cases	UK: 1979–1997; US: 1980–1996	1984		

References

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

Bollen, K.A., Paxton, P., 1998. Detection and determinants of bias in subjective measures. American Sociological Review 63 (3), 465–478.

Bollen, K.A., Paxton, P., 2000. Subjective measures of liberal democracy. Comparative Political Studies 33, 58–86.

Chaitin, G.J., 1975. Randomness and mathematical proof. Scientific American 232 (5), 47–52.

Feynman, R., 1965. The Character of Physical Law. MIT Press, Cambridge.

Jacoby, W.G., 1999. Levels of measurement and political research: an optimistic view. American Journal of Political Science 43 (1), 271–301.

Kinder, D.R., Palfrey, T.R. (Eds.), 1993. Experimental Foundations of Political Science. University of Michigan Press, Ann Arbor.

King, G., Keohane, R.O., Verba, S., 1995. The importance of research design in political science. American Political Science Review 89 (2), 475–481.

Klandermans, B., Staggenborg, S. (Eds.), 2002. Methods of Social Movement Research. University of Minnesota Press, Minneapolis.

Lakatos, I., 1970. Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge, pp. 91–196.

Saris, W.E., Andrews, F.M., 1991. Evaluation of measurement instruments using a structural modelling approach. In: Biemer, Groves, P.P. (Eds.), Measurement Errors in Surveys. John Wiley, New York, pp. 575–598.

Saris, W.E., Satorra, A., Coenders, G., 2004. A new approach to evaluating the quality of measurement instruments: the split-ballot MTMM besign. Sociological Methodology 34, 311–347.

Steenbergen, M., 2005. Multi-method approaches to measuring party positions, unpublished.

Suen, H.K., 1990. Principles of Test Theories. Lawrence Erlbaum, Hillsdale: NJ.

Tarrow, S., 1995. Bridging the quantitative–qualitative divide in comparative politics. American Political Science Review 89, 471–474.

Wackerly, D.D., Mendenhall, W., Schaeffer, R.L., 1996. Mathematical Statistics with Applications. Duxbury Press, Belmont.

Zeller, R.A., Carmines, E.G., 1980. Measurement in the Social Sciences. Cambridge University Press, Cambridge.

Further reading

References to data sources on party positioning

2002 Chapel Hill Expert Survey on party positioning on European integration for 14 western and 10 central- and eastern European countries. Source: Available from: http://www.unc.edu/~hooghe.

Benoit, Laver, 2003. Party policy positions data set, in press. Party Policy in Modern Democracies. Routledge, London. Available from: http://www.tcd.ie/Political_Science/ppmd/ (Source: Kenneth Benoit and Michael Laver).

Castles, Mair, 1984. 1984 Left/right positioning data. "Left–rights Political Scales: Some Expert Judgements". European Journal of Political Science 12, 73–88 (Source: Frank Castles and Peter Mair).

Comparative Manifesto Dataset 1945–1998, 2001. Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998. Oxford University Press, Oxford, UK (Source: Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Eric Tannenbaum, Judith Bara) Available on CD-Rom included with book.

Huber, Inglehart, 1995. 1995 Left/right positioning data. "Expert Interpretations of Party Space and Party Locations in 42 Societies". Party Politics 1, 73–111 (Source: John Huber, and Ronald Inglehart).

Laver, Hunt, 1992. 1992 Party policy positions data. Policy and Party Competition. Routledge, New York, NY (Source: Michael Laver and W.B. Hunt).

Marks, Steenbergen, 2006. 1999 Expert survey on party positioning on European integration in 14 EU countries. "Evaluating Expert Surveys." European Journal of Political Research. Available from: http://www.unc.edu/~gwmarks/data.htm (Source: Marco Steenbergen and Gary Marks).

Ray, 1999. 1984–96 Expert survey on party positioning on European integration in 16 Western countries. 'Measuring Party Positions on European Integration: Results from an Expert Survey'. The European Journal of Political Research 36 (2), 283–306. Available from: http://www.lsu.edu/faculty/lray2/data/data.html (Source: Leonard Ray).

Rohrschneider, Whitefield, 2003. Expert survey on party cleavages in postcommunist societies.

Other data sources used

Eijk van der, C., Franklin, M., Schönbach, K., Schmitt, H., Semetko, H., et al., 2002. European Elections Study 1999: Design, Implementation and Results. Steinmetz Archives, Amsterdam. Available from: http://shakti.trincoll.edu/~mfrankli/EES.html (Computer file and codebook).

Hix, S., Noury, A., Roland, G., 2005. 'Power to the parties: cohesion and competition in the European parliament, 1979–2001'. British Journal of Political Science 35 (2), 209–234. Available from: http://personal.lse.ac.uk/hix/HixNouryRolandEPdata.HTM.

Katz, R., Norris, P., Thomassen, J., Wessels. B., 1999. 'The 1996 Political representation in Europe survey of members of 11 national parliaments and members of European parliaments'. Available from: http://www.gesis.org/ZUMA/.

Keman, H., Pennings, P., 2004. 'The development of Christian & social democracy across Europe: changing positions and political consequences'. In: Tsatsos, D.Th., Venizelos, E.V., Contiades, X.I. (Eds.), Political Parties in the 21st Century. Wissenschafts Verlag/Emile Bruylant, Berlin & Brussels.

Laver, M., Benoit, K., Garry, J., 2003. 'Extracting policy positions from political texts using words as data'. American Political Science Review 97, 311–330. Programming information available from: http://wordscores.com/.

Gary Marks [a,b,*]
[a] *University of North Carolina,*
Department of Political Science,
Chapel Hill, NC 27599-3265, USA
[b] *Vrije Universiteit, Amsterdam,*
The Netherlands
[*] Fax: +1 919 962 5375.
E-mail address: marks@unc.edu