# THE 'PARADOX OF THE MANIFESTOS' – SATISFIED USERS, CRITICAL METHODOLOGISTS

Ian Budge <budgi@essex.ac.uk>

Essex University

March 2013

The very extensive use of the Manifesto estimates by users other than the groups (MARPOR, CMP, MRG) which generated them, attests to their indispensability and coverage. The fact that they have supported so many satisfactory research conclusions (over 1200 Google citations to Mapping Policy Preferences I and II alone) also confirms their general validity. Validity in turn guarantees high reliability – you cannot consistently produce good substantive results with flawed measures. Direct checks on reliability (Klingemann et al, 2006, 89-104) produce a range of coefficients from .78 to .94. By way of comparison the widely used Party Identification variable has been estimated to have a reliability of .86 (Converse/Markus, 1979, 39). While these estimates mainly cover random error (noise) recent work has shown that RILE at any rate shows less systematic bias than survey-based measures of party positioning (electoral and expert) (Best et al., 2012) and computer-based estimates using these as input.

In spite of quite overwhelming evidence for their research validity many methodological articles dealing with the Manifesto estimates have been critical in tone. At a minimum they suggest improved Right-Left measures, (e.g. Gabel & Huber, 2000). Mostly they start off from the premise that there must be error in the dataset (unexceptional – all data has some error) – but then proceed as if this must be so great that it renders estimates quite untrustworthy. Ignoring the CMP documentation quoted above (McDonald & Mendès, 2001: Klingemann et al, 2006, 89-104) they variously find error in the excessive variation of the estimates (Benoit & Laver, 2007) or in an alleged centrist bias (Lowe et al, 2011: Benoit et al, 2012). No matter that the evidence for reliability and for non-bias quoted above contradicts these assertions – which are also inconsistent in themselves. Documentation and reliability statistics are ignored to the extent of claim-

ing that MARPOR and the CMP have never investigated or measured error (Benoit, Laver & Mikhaylov, 2009, 296).

Table 1 summarises the points made by one persistent group of critics, which clearly demonstrates inconsistencies between different critiques over time (e.g. 'excessive variation' in RILE estimates in 2007, and then 'crowding into the centre' in 2011 and 2012). Unfortunately each new critique proceeds without reference to the others or to MARPOR or CMP documentation, so they do not add up to a sustained and consistent overall argument.

What they do cumulatively however is to give the impression that the estimates are error-prone and must always be corrected and adjusted before being entered into any statistical analysis. Indeed this is a specific recommendation in at least one website (2011) that additional error adjustments 'can and … should be used in any, research that utilises the CMP data' www.kenbenoit.net/ .

The important consequence from the point of view of users is that concentrating on error in the Manifesto estimates to the exclusion of possible error in the other variables in the analysis, leads to inflating the former's influence – a Type 1 error. This can be shown in the following simulation which builds on an actual analysis reported in McDonald & Budge (2005, 220-223). The question is whether the left-right position of the median party in parliament (MPP) is an important consideration for understanding a central government's welfare policy regime. The welfare regime indicator is Esping-Andersen decommodification score for each of 17 advanced Western democracies—i.e., where decommodification refers to a composite summary indicator of social services rendered as a matter of right such that maintenance of one's living standard is possible without relying on the market (Esping-Andersen 1990, 22). Along with a hypothesised MPP effect, the McDonald-Budge model includes the percentage of the nation's population over age 65 and Arend Lijphart's measure of consensus democracy (1999). The question of most interest is whether the medium- to long-run political preferences of parliaments have an actual effect on the extensiveness of welfare regimes beyond (1) the demands placed on a system by an aging population and (2) the organization of politics in terms of consensual versus adversarial institutional arrangements.

Table 1: Replications of McDonald-Budge Analysis of Welfare State Organisation under Varying Conditions of Measurement ReliabilityA (N = 17 for all equations)
Dependent Variable = Decommodification Index

| Independent Variable | *Model 1* Original Equation | *Model 2* MPP $r_{xx}$ = .9 | *Model 3* MPP $r_{xx}$ = .9 CD $r_{xx}$ = .9 | *Model 4* MPP $r_{xx}$ = .8 |
|---|---|---|---|---|
| [redacted] | *b* ($s_b$) | *b* ($s_b$) | *b* ($s_b$) | *b* ($s_b$) |
| % of Pop > Age 65 | 1.940** (.471) | 1.908** (.453) | 1.867** (.449) | 1.862** (.427) |
| Median Party in Parliament (MPP) | -.259* (.105) | -.303* (.117) | -.285* (.119) | -.366* (.131) |
| Consensus Democracy (CD) | .020* (.010) | .018 (.010) | .021* (.012) | .015 (.010) |
| Intercept | .132 (6.03) | .422 (5.79) | .930 (5.72) | .831 (5.43) |
| $R^2$ | .828 | .842 | .849 | .862 |
| $S_{Y\|X}$ | 3.67 | 3.51 | 3.42 | 3.28 |

The original McDonald-Budge results are reported in the left-most column of coefficients in Table 1. These results assume that none of the variables are measured with error. Assuming no measurement error, the MPP left-right variable, with a coefficient of -.26 (where high scores for the MPP preference indicate a parliament standing on the political right), indicates that a parliament at a centre right position, +10, has an expected decommodification score just over 5 points lower than a parliament at a centre-left position -10. [A five point difference on decommodification is the sort of distinction Esping-Anderson's scoring gives to the German versus Dutch welfare states.] The relative size of the post-retirement age population and the organisation of political institutions in consensual rather than adversarial forms also have reliably estimated effects, strongly so for the aged population but weakly for consensual institutional arrangements.

Column 2 of the table shows the estimated effects if the measurement of the MPP position is not totally but, instead, 90% reliable. All three variables retain their statistical significance at conventional levels ($p < .05$), but the magnitude of the MPP increases by 15% (from -.26 to -.30), with a slight increase in its standard error, while the effects of other variables decline slightly. Column 3 shows that if both the MPP and the consensus democracy variables have reliabilities of .9, nothing dramatic occurs in terms of the estimated effects compared to the original coefficients (column 1). Other possibilities associated with measurement error follow this same pattern: the lower the reliability of the one variable, the higher that variable's estimated effect and the lower the other variables' estimated effects. An MPP reliability of .8 increases the estimated effect to .37 (see column 4) and reduces the aged population effect slightly while reducing the consensus democracy effect to statistical insignificance. Another alternative, reducing the consensus democracy reliability to .9 while the MPP variable is perfectly reliable, increases the consensus democracy effect slightly and reduces the other two effects, also slightly.

Depending on the degree of reliability, the effect of MPP left-right position may thus have an effect as low as 5 decommodification units or as high as 7.5 units for parliaments on the centre-right versus on the centre-left. More specifically, depending on MPP's reliability, the estimated effect with total reliability is 5.2 units; with 90% accuracy it is 6.0 units; and with 80% accuracy it is 7.4 units. Imputing less reliability to the Manifesto estimate actually increases its inferred effects, rendering the assumption of total reliability in the estimates quite a conservative one.

In practice Manifesto data reliability has been estimated as between .8 and .9, with some other measures going up to 1.00 (Klingemann et al., 2006, 103) so the reliability range reported in the Table is entirely plausible. The trouble is often the absence of similar estimated reliabilities for the other variables (e.g. consensus democracy). Why should the Manifesto estimates be uniquely error prone? For comparability an assumption of total reliability for all variables may be the best we can make. These simulations also show that the estimated relationship with MPP is robust in the face of marginal error fluctuations even though its exact magnitude is sensitive to them. Such a result clearly undermines claims that additional error adjustments should be made.

Why should methodological assessments be generally critical of the Manifesto estimates while research experiences are generally positive? One reason is that methodologists rarely apply their conclusions to actual research – not even going so far as to compare the estimates their alternative approach produces compared to the originals. If they did so they would often find that they match very closely. For example Lowe et al's (2011) logit ratio scale has a correlation of $r = .94$ when applied to the data.

Benoit et al's (2009) adjusted SIMEX procedure correlates at .99! This does not only imply that the mean estimates are almost the same. It implies further that every adjusted estimate matches its original closely.

This raises the question of why bother with the adjustments? What drives critiques when the mass of evidence favours the originals? There seem to be two major forces involved:-

- The first is structural. Any methodological article which simply extolled the strength of the Manifesto estimates would very likely not be published. Journals demand originality. The quality of the estimates has been extensively investigated and documented. Any methodological review which simply repeated this would risk rejection as uninformative. There is a premium therefore in finding faults or improvements and elaborating on them. 'Manifesto estimates good' – not news. 'Manifesto estimates bad' – big news, given their extensive use and indispensability to so much research.
- The second force driving critiques is a basic incredulity that manual coding can produce better and even more reliable estimates than computerised procedures. Texts, with their potential for word counting, are a natural terrain for procedures such as Wordscores (Laver, Benoit, Garry, 2003) or Wordfish (Slapin & Proksch, 2008). Computerised procedures moreover have an undisputed ability to reproduce their results exactly – though in the following limited sense. Given the same input and the same texts, the programme will always produce the same results. Problems however arise both about reliability and validity when we look directly at the estimates the procedure makes. Some of the input the programme works on may be systematically biased e.g. the expert judgements of party positions usually input to Wordscores for example. The texts used to define the scorings may not be the only ones which could be used. If equally authoritative ones are substituted the estimates change – they are unreliable in this broader sense. Words are also not the natural unit of sense in our languages. Sentences and arguments are. Human codings of these may thus be more valid and – in the broader sense – reliable than computerised codings.

This does not however prevent true computer believers trying to pick holes in manual procedures whenever they can – despite the limited take up and scope of computerised text processing up to now.

Table 2: Negative Critiques of the Manifesto Estimates 1990-2011

| *Date* | *Publication* | *Nature of Criticism* | *Follow Up/Responses* |
|---|---|---|---|
| 1992 | Laver and Hunt *Policy and Party Competition* | Expert Placements on Policy Dimensions (not Left-Right) suggested as more nuanced alternative to Manifesto Estimates | Used as panel data with expert judgements reported by Castles and Mair (1984); Huber and Inglehart (1995) and Benoit and Laver (2006). Hampered by absence of Left-Right placements available from the other expert surveys |
| 2001 | Laver and Garry 'Estimating Policy Positions from Party Manifestos' | Alternative, partly computerised, coding of party pro and con positions based on key words distinguishing parties: criticises saliency assumptions underlying (many of) MRG-CMP coding categories | Abandoned in favour of scoring system suggested by Kleinjenhuis and Pennings (2001), developed as Wordscores - and based on saliency assumptions! |
| 2003 | Laver, Benoit and Garry 'Estimating Policy Positions ... Using Words as Data' | Computerised count of words in texts 'Wordscores' can be used to score them with absolute reliability unlike MRG-CMP manual codings | Budge and Pennings (2007) point out that Wordscores position estimates are unreliable as they fluctuate depending on what text is used to score the others. Limited use of Wordscores owing to this difficulty |
| 2006 | Benoit and Laver *Party Policy in Modern Democracies* | Error statistics available for experts' judgements (within countries). Left-Right should be conceived as a contentless dimension involving different issues at different points in time and in different countries | Not widely used except to make up over-time panels with earlier expert surveys. Suffers from centrist bias endemic to survey-based estimates, which eliminates cross-national variation |
| 2007 | Special Edition of *Electoral Studies* ed. Marks party positioning | Benoit and Laver (2007b) 'Response' to Budge and Pennings criticizes absence of error and uncertainty estimates in Manifesto data, Benoit & Laver 2007a attribute error to excessive variation in the Manifesto estimates leading to systematic exaggeration of party policy change to centre or extremes | Absence of error and uncertainty measures continues to be main criticism of Manifesto estimates in spite of reliability estimates and confidence intervals published in Klingemann et al, 2006, 86-104. Criticism of excessive variation somewhat contradicts later 'centrist' criticism |

| *Date* | *Publication* | *Nature of Criticism* | *Follow Up/Responses* |
|---|---|---|---|
| 2008 | Mikhaylov, Laver, BenoitCoder Reliability & Misclassification in CMP codings | CMP inter-coder reliability test simulated by coders working for Mikhaylov, Laver and Benoit with bad results. Concludes that Manifesto data are unreliable as a whole and that Left-Right scale is systematically biased towards centrist placements (2009) or rightist placements (2010) | Klingemann et al, 2006, 106- 7 points out that the test is part of coder training not production coding which is carried out by different coding simulation procedures. The simulation therefore is irrelevant to the quality of estimates. However results have continued to provide a basis of criticism |
| 2009 | Benoit, Laver, Mikhaylov 'Uncertainty in Text Statements of Policy Positions' | Published manifestos are randomly sampled from a population of alternative policy statements so constituent (quasi-) sentences can be randomly dropped and replicated to see how estimates shift. Longer documents are more stable than shorter. Results can be used to calculate confidence intervals for every manifesto-based policy estimate which should be adjusted before being used in any (regression or other) analysis | Klingemann et al, 2006, xvi and *passim* stresses that manifestos are:- a) a *population* of *authoritative* policy statements by party; b) produced by intensive scrutiny of every (quasi) sentence in the text. These cannot be repeated or dropped without changing the true meaning. Reliability coefficients and confidence intervals for final point estimates can be calculated on this basis (MPPII, 90-104). No adjustments to estimates needed before multivariate analyses, which have inbuilt tests of error and uncertainty. |
| 2011 | Mikhaylov, Laver, Benoit Coder Reliability and Misclassification in CMP codings | CMP inter-coder reliability test simulated by coders working for Mikhaylov, Laver and Benoit with bad results. Concludes that Manifesto data are unreliable as a whole and that Left-Right scale is systematically biased towards centrist placements. | Klingemann et al, 2006, 106-7 points out that the test is part of coder training not production coding which is carried out by different coding simulation procedures. Simulated test is therefore irrelevant to final estimates. However results have continued to provide a basis of criticism |
| 2011, 2012 | Lowe et al, 2011: Benoit et al; 2012 | RILE suffers from a centrist bias which necessitates wholesale substitution either with logit ratio scale or specific policy sub-scale constructed in the same way | Logit procedure produces estimates which correlate with RILE estimates ($r$ = .94). Where they differ this is due to substitution of .5 for zero in logit procedure |

*Notes for Table:*

Kenneth Benoit and Michael Laver (2006). Party Policy in Modern Democracies, London, Routledge

Benoit, Kenneth, et al. (2012) "How to scale coded text units without bias: A response to Gemenis." Electoral Studies 30: 1-4.

Michael Laver and W. Ben, Hunt (1992). Policy and Party Competition, New York and London, Routledge

Michael Laver and John Garry (2000) Estimating Policy Positions from Political Texts, American Journal of Political Science, 44: 619-34

Michael Laver, Kenneth Benoit and John Garry (2003). Extracting Policy Positions from Political Texts using Words as Data, American Political Science Review 97: 311-31

W. Lowe, Kenneth Benoit, Slava Mikhaylov, Michael Laver (2011). Scaling Policy Preferences from Coded Politics. Legislative Studies Quarterly, 36, 123-155

Kenneth Benoit, Michael Laver and Slava Mikhaylov (2009). Treating Words as Data with Error – Uncertainty in Text Statements of Policy Positions' American Journal of Political Science 53: 495-513

Slava Mikhaylov, Michael Laver and Kenneth Benoit (2008, 2009, 2010). 'Coder Reliability and Misclassification in CMP Codings. Paper for 77th Midwest Political Science Association Annual National Conference and following papers on websites as 'Coder Reliability and Misclassification in this Human Coding of Party Manifestos' final version published in Political Analysis (2011), 20.1, 78-91.

# References


Benoit, Kenneth, and Michael Laver (2007). "Estimating party policy positions: Comparing expert surveys and hand-coded content analysis." Electoral Studies: 90-107.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov (2009). "Treating words as data with error: Uncertainty in text statements of policy positions." American Journal of Political Science: 495-513.

Benoit, Kenneth, et al (2012). "How to scale coded text units without bias: A response to Gemenis." Electoral Studies 30: 1-4.

Best, Robin E., Ian Budge, Michael McDonald (2012). "Representation as a Median Mandate: Taking Cross-national Differences Seriously". European Journal of Political Research: 1–23.

Converse, Philip E., and Gregory B. Markus (1979). "Plus ca change...: The new CPS election study panel." The American Political Science Review: 32-49.

Esping-Andersen, Gøsta (1990). The three worlds of welfare capitalism. Vol. 6. Cambridge: Polity press.

Gabel, Matthew J., and John D. Huber (2000). "Putting parties in their place: Inferring party left-right ideological positions from party manifestos data." American Journal of Political Science: 94-103.

Klingemann, Hans-Dieter/Andrea Volkens/Judith Bara/Ian Budge/Michael Macdonald (Eds.) (2006). Mapping policy preferences II: Estimates for parties, electors and governments in Central and Eastern Europe, European Union and OECD 1990-2003. Oxford: Oxford University Press (2006).

Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting policy positions from political texts using words as data." American Political Science Review : 311-331.

Lijphart, Arend (1999). Patterns of Democracy: Government Forms & Performance in Thirty-six Countries. New Haven: Yale University Press.

Lowe, Will, et al. (2011). "Scaling policy preferences from coded political texts." Legislative studies quarterly: 123-155.

McDonald, Michael D., and Ian Budge (2005). Elections, parties, democracy: conferring the median mandate. Oxford: Oxford University Press.

McDonald, Michael D., and Silvia M. Mendes (2001). "The policy space of party manifestos." In: Michael Laver (ed.). Estimating the Policy Position of Political Actors 20: 90.

Slapin, Jonathan B., and Sven Oliver Proksch (2008). "A scaling model for estimating time–series party positions from texts." American Journal of Political Science: 705-722.