Comparative Social Research

Shortcut URL to this
page: **http://www.sciencedirect.com/science/bookseries/01956310**

Volume 24, Pages 1-409 (2007)

Capitalisms Compared
Edited by: Lars Mjøset and Tommy H. Clausen
ISBN: 9780762313136

A Symposium on Methodology in Comparative Research

New Methods for Comparative Research?
Pages 373-389
Claude Rubinson and Charles C. Ragin

Rejoinder: Affirming Limits and Defending Alternatives to Multiple Regression
Pages 391-409
Michael Shalev

LIMITS AND ALTERNATIVES TO MULTIPLE REGRESSION IN COMPARATIVE RESEARCH

Michael Shalev

This paper criticizes the use of multiple regression (MR) in the fields of comparative social policy and political economy and proposes alternative methods of numerical analysis. The limitations of MR in its characteristic guise as a means of hypothesis-testing are well known. The emphasis here is on the specific difficulties of applying MR to the problem of explaining diverse outcomes across a limited range of country cases. Two principal conclusions will emerge. First, even though technical means are available to deal with many of the limitations of MR, these solutions are either unconvincing or else require such advanced technical skills that they offer questionable returns on scholarly investment. Second, dissatisfaction with MR does not necessarily mandate radical alternatives or abandonment of numerical methods altogether. ''Low-tech'' forms of analysis (tabular and graphical methods) and multivariate statistical techniques other than MR (such as factor analysis) constitute viable and useful alternatives.

The comparative study of welfare states is a good example of the characteristic methodological polarization that afflicts the social sciences. Historians and social policy analysts with an intrinsic interest in welfare states engage in descriptive and prescriptive studies, while at the other extreme are ''hard-nosed'' social scientists who regard the welfare state essentially as a

Capitalisms Compared
Comparative Social Research, Volume 24, 261–308

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24006-7

convenient source of data for testing abstract theoretical claims. The sociologists and political scientists who began studying social policy in the late 1970s were part of the quantitative revolution in comparative studies. Using simple correlation and regression analysis, they optimistically hoped to settle the competition between a handful of master explanations for variation in the size of welfare states (Amenta, 1993; Shalev, 1983). Over the last two decades there has been a compelling trend toward greater sophistication in quantitative work (for a pioneering compilation see Janoski & Hicks, 1994). Especially noteworthy is the growing recognition by comparativists of the limitations of simple cross-sectional uses of MR, and their attempts to overcome these limitations without sacrificing the power of regression. Indeed, refined data analysis is the hallmark of a new and statistically more literate generation of scholars (see particularly the series *Cambridge Studies in Comparative Politics* including works by Boix (1998), Garrett (1998), Iversen (1999), Franzese (2001) and Swank (2002)). At the center of these studies are complex analyses of pooled datasets that cover multiple countries at multiple moments in time.

Earlier works in comparative political economy tended to focus on explaining enduring cross-national differences (more rarely, they looked at differences between countries in historical dynamics). The standard tools of the trade were scatter-plots, correlations and primitive cross-sectional regressions (e.g. Tufte, 1978; Cameron, 1984). This was true even of methodologically advanced practitioners (e.g. Hibbs, 1978; cf. Shalev, 1979b). The turning point was a controversial cross-national regression study by Lange and Garrett (1985) which sought to show that the combination of strong unions and left governments was beneficial for economic growth following the first "oil shock". In a final response to their critics Garrett and Lange (1989) suggested that the debate could only be resolved by the use of a pooled cross-sectional time series design, which in addition to furnishing a much larger number of observations would enable researchers to directly study whether the effects of changes in government composition are conditioned by national institutional contexts. Two years later Alvarez, Garrett, and Lange (1991) published their seminal article "Government Partisanship, Labor Organization, and Macroeconomic Performance" which turned pooled regression into the design of choice for quantitative comparative political economists.

Alternative approaches include Ragin's (1987, 2000) innovative attempts to formalize the analytical approach of traditional comparative-historical scholarship, and Berg-Schlosser's demonstrations of alternative multivariate techniques (e.g. Berg-Schlosser & De Meur, 1997; Berg-Schlosser, 2002).

However, especially in the United States these methods have had little impact.[1] So far the only significant qualification to the dominance of MR in general and pooled models specifically in quantitative work on comparative political economy, has been the insistence of some practitioners on the necessity for constructive dialog between comparative history and multicountry regression analysis (see especially Hall, 2003). John Stephens and his collaborators have been the most committed exponents of this approach (Rueschemeyer, Huber-Stephens, & Stephens, 1992; Huber & Stephens, 2001), although case studies also play a subsidiary role in several notable applications of pooled regression (e.g. Boix, 1998; Iversen, 1999; Swank, 2002). Perhaps the most telling symptom of the hegemony of regression in quantitative comparative research is Gøsta Esping-Andersen's (1990) seminal work on welfare state regimes. It is striking that after offering a forceful critique of the core assumptions of conventional methodology, Esping-Andersen himself turned to MR in order to assess the empirical validity of his arguments.

The final section of this paper reanalyzes Esping-Andersen's data using techniques better suited to his theoretical and methodological premises. The preceding section offers an extended critique of pooled regression analysis. Prior to these two parts of the paper I first present an overview of the deficiencies of MR as a tool of macro-comparative research and then offer two detailed illustrations of how standard applications of MR in comparative research can generate misleading results that are inferior to those obtained using simpler methods.

STRENGTHS AND WEAKNESSES OF MULTIPLE REGRESSION

The difficulties that MR poses for comparativists were anticipated 40 years ago in Sidney Verba's essay "Some Dilemmas of Comparative Research", in which he called for a "disciplined configurative approach … based on general rules, but on complicated combinations of them" (Verba, 1967, p. 115). Charles Ragin's (1987) book *The Comparative Method* eloquently spelled out the mismatch between MR and causal explanation in comparative research. At the most basic level, like most other methods of multivariate statistical analysis MR works by rendering the cases invisible, treating them simply as the source of a set of empirical observations on dependent and independent variables. However, even when scholars embrace the analytical purpose of generalizing about relationships between variables, as opposed

to dwelling on specific differences between entities with proper names, the cases of interest in comparative political economy are limited in number and occupy a bounded universe.[2] They are thus both knowable and manageable. Consequently, retaining named cases in the analysis is an efficient way of conveying information and letting readers evaluate it.[3] Moreover, in practice most producers and consumers of comparative political economy are intrinsically interested in specific cases. Why not cater to this interest by keeping our cases visible?

Different views of causality are an equally celebrated source of the debate between case-oriented and variable-oriented researchers. Andrew Abbott (1998, p. 183) has cogently argued that "all too often general linear models have led to general linear reality, to a limited way of imagining the social process". Abbot notes the constricted theoretical scope of the notion of causality underlying linear models, which cannot recognize (or at least is unlikely to recognize) situations where the effect of any given causal variable is uneven, contradictory (dialectical), or part of a wider bundle of factors sharing an elective affinity. In the social world effects are typically contingent upon their setting, including two types of historical contingency: temporal context (period effects) and time paths (particular historical sequences or cumulations). The problem is not that MR does not have or could not invent technologies for dealing with such complexities. Non-linear functional forms, interaction effects and (in time series analysis) complex lag structures immediately come to mind. The point is that because such techniques are either difficult to employ or impose a steep statistical penalty due to the "small-n problem", they are rarely or insufficiently used.

Case-oriented analysis easily accommodates the nuances that concern Abbot and likeminded critics, because it assumes from the outset that the effect of any one cause depends on the broader constellation of forces in which it is embedded ("conjunctural causation" in Ragin's words). If MR models try to emulate this assumption they are likely to quickly exhaust available degrees of freedom. MR is even more challenged by another causal assumption that flourishes in case-oriented analysis, namely that there may be more than one constellation of causes capable of producing the phenomenon of interest. That is, some cases are explained by one causal configuration and others by a different configuration. Statisticians refer to the phenomenon of multiple pathways to a common outcome as causal heterogeneity. MR models cannot handle this simply by increasing the number of independent variables. The results will be ambiguous because they will be unable to distinguish between additive effects, conditional relationships and multiple causal pathways.

The difficulty may be illustrated by a well-known finding of comparative welfare state research. Two subtypes of European welfare states that developed under different political auspices – Social Democracy and Christian Democracy – are known to be high spenders (for landmark studies, see Korpi, 1983; Van Kersbergen, 1995). This presents no problem for the standard additive regression model provided that the two effects are equivalent and unrelated – if for instance a strong social-democratic party could be expected to have the same effect whether or not it governed in coalition with a Christian-Democratic party. However the Austrian experience suggests that this is unlikely since historically, the black half of the "red-black" coalition severely constrained its welfare state development (Esping-Andersen & Korpi, 1984). This suggests the need for an interactive (conditional) model.

A more radical challenge to the linear additive model is posed by Esping-Andersen's (1990), later claim that Christian-Democratic welfare states have both a policy logic and a political logic that are qualitatively different from those of Social Democracy. Although in terms of overall expenditure both social policy regimes are relatively costly, they represent two different causal syndromes that in respect to expenditure happen to result in similar outcomes. The standard regression model would treat the two political constellations as two independent variables and force them to compete to explain variance in the dependent variable. As a result the real effect of both would be diluted. And what of the hybrid Austrian case? In practice, except for the liberal English-speaking nations nearly all of the advanced political economies tend to be *either* Christian-Democratic or Social-Democratic. The peculiarities of Austrian social policy should thus be understood as the result of this cohabitation and its particular historical sequencing. They cannot be represented causally by summing the effects of the two political trends (additive model), or by trying to infer from the singular Austrian experience a law-like effect of their juxtaposition (interactive model).

To appreciate why MR is a problematic choice for comparativists, it is also helpful to consider why it may be a good choice for certain other kinds of social scientists. Economists are often interested in estimating the marginal effect of one economic variable on another, holding constant the impact of other presumed causes. If prices rise, what will be the likely effect on economic growth, net of other known influences like the rate of investment and the terms of trade? If people invest in a college degree, what will be the likely effect on their future income stream, net of other known influences like work experience? MR suits this project well. Estimating marginal effects under conditions of ceteris paribus is precisely what it aims to do.

In contrast, much of the curiosity of comparative political economists revolves around the presence or absence of certain conditions. Will economic growth be higher in the presence of corporatist trade unions (or a hegemonic social-democratic party, or an independent central bank)? It would be nice to know *how much* growth results from *how much* corporatism, but our theoretical interests are typically far more elementary and our predictions quite imprecise.

The evaluation of marginal effects in macro-comparative research is also dogged by the ambiguity of many of the variables of interest and the difficulty of measuring them precisely.[4] Concepts like corporatism are so contentious that even categorical measures exhibit worrying inconsistencies (Kenworthy, 2001; Shalev, 1990). Some theoretical approaches in comparative politics are almost immune to successful quantification. An example is state-centered theory (e.g., Weir & Skocpol, 1985). Although the problem may partly be theoretical slipperiness, only superficial aspects of the structure of states (such as constitutional provisions) have proven to be measurable (e.g. Huber, Ragin, & Stephens, 1993). The framing of political action and agendas by state capacities, policy legacies and the autonomous initiatives of state managers has not been given serious consideration except in non-formal historical research.[5] In contrast, naturally continuous variables like "left party cabinet representation" can be measured precisely. Unfortunately, however the use of such measures is rife with problems of both reliability and validity. Inter-country comparisons of long-term differences in left party power are plagued by the difficulty that, for example, a mean fraction of 50% of cabinet seats is consistent with either intermittent left government, stable left participation in cabinet coalitions, or a dominant left party which is unseated in midstream. Comparison over time is equally problematic, since the numbers alone cannot tell us whether the left's role in government has shifted between *qualitatively different conditions* like one-party dominance, wall-to-wall coalitions, junior partnership, pivot party facing a divided right, etc. MR could accommodate such complexity by replacing the continuous measure of left strength with a series of dummy variables, or perhaps by finding an appropriate non-linear functional form to capture discontinuities in the effect of left strength on the phenomenon of interest. But the first solution is "wasteful" of precious degrees of freedom and the second requires either good luck or an unlikely degree of theoretical sophistication.

In the behaviorist sub-fields of political science and related disciplines much of the appeal of MR derives from its comfortable fit with sample survey methodology. Because they enjoy a relatively high ratio of cases to

variables, survey researchers are able to use MR as a means of introducing statistical controls. Unlike economists they may not be motivated by an ontological view that is inherently marginalist. They use controls in the hope of dealing with causal forces that in the ideal experimental design would have been neutralized by random assignment of subjects to differential "treatments". This approach has been the subject of vigorous debate. In different ways David Freedman (1991) and Stanley Lieberson (1985) have made compelling arguments that proper statistical control would require much more sophisticated and complete causal theories than social researchers can hope to have.[6] Even assuming that comparative political economists had such theories, given the small number of cases included in their empirical research it is technically difficult for them to analyze the effect of more than a few independent variables at a time.

Staying with the survey researchers, we can identify a final reason why the appeal of MR outside of comparative research need not inspire its use within the field. To economize on resources, analysts of voter opinion or social mobility usually poll only a tiny fraction of their target population. As a result, a fair amount of the immense heterogeneity that characterizes a universe like "American voters" cannot possibly be captured in the typical sample of only one or two thousand. Nevertheless, even the most unlikely combinations of the independent variables probably do exist in the target population. From this viewpoint one of the advantages of MR is that using the observations in hand, its coefficients (marginal effects) project relationships across the whole spectrum of potential configurations of variables.

In cross-national quantitative research the situation is very different. We often analyze the entire universe of cases, and if not it is usually because of lack of data rather than sampling considerations. For the most part then, *if a particular configuration of attributes does not exist in a cross-national dataset, it does not exist at all*. To grasp the size of the problem, consider the following hypothetical example using only three independent variables and a crude level of measurement. Social security expenditure as a proportion of GDP is regressed on left party power, exposure to trade and proportion of the population over 65. All variables are measured on a 5-point scale. If we were to construct a multiway table with this dataset, it would have 625 ($5 \times 5 \times 5 \times 5$) cells. Since no study of the OECD area can have more than about 20 cases, this implies over 600 empty cells! MR in effect places imaginary countries in some of these empty cells when it seeks out the best linear fit that can be generated for the data at hand.[7] Because it estimates partial parameter effects as if all (linearly-fitting) configurations were possible, MR can easily yield problematic results.

The venerable social-democratic model of the welfare state illustrates this problem (Shalev, 1983). Andrew Martin's (1973) pioneering comparison of the US and Sweden inferred that social-democratic party dominance was the crucial difference responsible for Sweden's postwar commitment to the full-employment welfare state, compared with its glaring absence in the US. Numerous correlation and regression studies echoed this argument and went on to seemingly confirm its veracity across the whole spectrum of advanced capitalist democracies. Yet, this model could tell us little or nothing about the causes of policy variation between the US and other liberal political economies, or within the US over time. The coefficient for social-democratic rule generated by cross-sectional regressions yielded absurd inferences along the lines that with one additional decade of socialist rule, America (or a country like it) would probably boast an unemployment rate three points lower and child allowances 40% higher. This is an extreme example of the dangers of generalizing from empty cells when each of our cases is a complex historically bounded *gestalt*. Still, it cannot be denied that one of the tests of a useful causal model is that it will be capable of answering counterfactual questions – that is, of filling empty cells with hypothetical data. Indeed, it was precisely by asking how US policy would have developed under Swedish conditions that Martin and others were led to focus on the causal role of labor movement strength. However, some "cells" are so unlikely ever to be filled that they should not be part of either our computational space or our predictions (King & Zeng, 2002). The attributes of societies are not subject to infinite variation in unlimited combination with one another.

From an MR perspective, the problem of empty cells may not be intractable. If a variable capable of explaining differences between Sweden and the US offers no guidance to the contrast between Canada and the US, then our model must be either under-specified or mis-specified. If the problem was under-specification the appropriate response would be to add independent variables capable of accounting for the observed variation. But with these additional variables in the model, it might become too large to estimate on a small cross-sectional dataset. In response, we might be tempted to enlarge our dataset by combining cross-sectional observations for different years. This would have the added advantage of permitting the investigation of intra-country differences (i.e. within the US as well as between the US and other countries). As noted, this pooling strategy is the subject of a later section of the paper.

If mis-specification is the problem then the solution would be to find an explanation sufficiently general that it could accommodate a wider range of variation – between the US and Canada as well as vis-à-vis Sweden.

In contrast, comparativists steeped in the case-oriented tradition would be more likely to assume causal heterogeneity. Instead of looking for a new master explanation they would seek an additional one tailored to cases that are inconsistent with prevailing theory. Following this logic, in the comparative study of political economy and public policy it has become common to assume that distinctive causal trajectories apply to different "families of nations" (Castles, 1993). If MR is obviously not the best way of testing plural explanations, what is? This issue will be discussed later in the context of Esping-Andersen's claim that there are three distinctive welfare state regimes.

Before proceeding to the questions of whether pooling resolves the problem of "too many variables and not enough cases" and whether regression is capable of dealing with causal heterogeneity, the paper offers two specific examples of the everyday use of MR. These illustrations were chosen with an eye to countering two possible responses to the general critique of MR that has been offered so far. One of these would be to lower our expectations and utilize regression more as a means of partitioning empirically observed variance than of rigorously testing hypothesized causal relationships. Alternatively, it might be argued that the causal status of regression coefficients should indeed be treated tentatively, but that our confidence is strengthened if alternative types of numerical and non-numerical analysis yield convergent findings. Both approaches have their problems. The next section critiques an illustration of the use of MR as only a loose guide to the plausibility of alternative models. Using a different example, the section that follows shows that even convergence among different methodologies does not guarantee that the data will yield their fundamental secrets.

"CAUSAL ARGUMENTS" OR MERE "SUMMARIES"?

> With multidimensional data sets, regression may provide helpful summaries of the data. However, I do not think that regression can carry much of the burden in a causal argument. (Freedman, 1991, p. 292)

David Freedman is a statistician who believes in the power of numbers but has made it his mission to disabuse social scientists of their exaggerated belief in statistical inference as a tool of causal analysis (Freedman, 1985, 1987, 1991). The essence of the argument made by Freedman (see also Leamer, 1983) is that statistical hypothesis-testing requires that researchers have a well-developed theory and a hands-off relationship with the data

prior to the point at which testing is carried out. In practice social-science research is based on weak or incomplete theories and its empirical generalizations are almost always the outcome of numerous iterations. Accordingly, when forced to confront the fact that progress in social research rests on a "dialog of ideas and evidence" (Ragin, 1994b), one should concede that the most which can legitimately be done with MR is to use it to summarize multivariate datasets.

Given prevailing expectations regarding publishable research, few scholars have the courage to claim that their research objectives are purely descriptive (Abbott, 1998). Still, some comparative research has treated MR as less than a formal hypothesis-testing device and more like an economical method of sustaining broad empirical claims. An example of this low-expectations approach can be found in Rothstein's (1990) study of cross-national variation in union membership from a new institutionalism perspective. Although Rothstein's article was primarily based on comparative-historical analysis, it included a simple cross-country regression. The substantive background to the study was that under the so-called "Ghent system" unions bear responsibility for administering unemployment insurance, with the consequence that in periods of economic crisis or transformation their membership is unlikely to be eroded and may even increase. For theoretical reasons, Rothstein wished to demonstrate that the highest levels of unionization have been reached only in countries where this system is in place. His union density figures for 18 OECD countries in the mid-1980s reveal that Ghent is indeed present in all of the countries with the highest rates of union penetration, and only these countries. Hence, unless Ghent is but a spurious understudy for the real star of the causal show, it has been a necessary condition for rates of more than 70% unionization. Of course, this does not mean that the Ghent system is a *sufficient* condition for union success. Perhaps it merely amplifies the effects of other favorable conditions.

There are thus several possibilities that a simple table showing union membership alongside Ghent presence/absence cannot address: spurious association (alternative explanations), additional causes (complementary explanations), and interaction effects (conditional explanations). Following convention, Rothstein seeks to lay the first two of these issues to rest by executing a multiple regression that takes into account other probable influences on cross-country differences in unionization. These are left party participation in government, and potential union membership (the absolute number of employed and unemployed wage-earners).

Rothstein's model was re-estimated for this article using a modified version of his dataset.[8] Following the original, the coefficients are standardized *betas*.

$$\begin{aligned}\text{Percent Unionized} = {} & 0.47(\text{Ghent}) + 0.28(\text{Left Government}) \\ & - 0.34(\text{Log of Potential Membership})\end{aligned}$$

All coefficients are significant at conventional levels (although Left Government only marginally) and the adjusted R-squared is 0.73. The metric coefficient for the Ghent variable reveals that the net average difference in unionization between Ghent and non-Ghent systems is a striking 27-percentage points.

Notwithstanding these indications of success, it can be argued that Rothstein's use of MR is inappropriate and in part misleading. Rothstein is content, in his words, to show "that all three variables have an independent explanatory effect of about the same standardized size" (Rothstein, 1990, p. 41). However, a prerequisite for these "explanatory effects" to have causal meaning is that the model be theoretically plausible. Rothstein himself casts doubt on this, when he describes the argument for the significance of potential membership size as logically indefensible, and suggests that the left-government argument suffers from what econometricians call simultaneity bias. In addition, while the standardized coefficients indeed suggest that Ghent has at least as much empirical weight as rival explanations, because countries are invisible the results do not speak to Rothstein's core claim that it is Ghent, not left strength or small size, which differentiates between the most unionized countries and all the rest. True, this claim would have been negatively ruled out had the Ghent effect disappeared once the other variables were added to the equation. But the regression could not make a positive case for Rothstein's argument.

Beyond these specific limitations of MR in Rothstein's case, his model rests on a standard but questionable assumption. Rather than operating as a syndrome of elective affinities, the explanatory variables are assumed to exert causally distinct effects. Consequently, none of the effects is assumed to be conditional on the value of other variables – i.e. no interactions are anticipated.

A straightforward way to address these issues is to summarize causes and effects in a way that identifies different combinations of conditions (causes) with the countries that "carry" them. This requires some forethought because Rothstein's model refers to three different causal variables and his dependent variable, unionization, is not easily collapsed (it is distributed

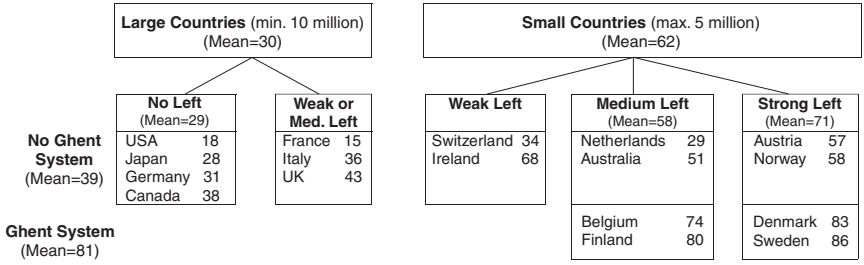

Chart 1. Reanalysis of Rothstein's Model of Union Membership.

fairly evenly across a broad spread). The proposed solution is a simple flow chart or "tree" showing exact values of unionization for different clusters of countries. These clusters were created simply by cross-tabulating the presence or absence of Ghent with categorical versions of Rothstein's two other causal variables.[9]

The results (Chart 1) offer interesting evidence of nested causal effects. This is immediately apparent from the systematic difference between extant and non-existent configurations. Substantial left party representation was only attained in small countries, and only countries with a substantial left had the Ghent system.[10] In the case of the affinity between Ghent and left strength, Rothstein himself pointed out that we cannot know which way the causal arrow points without branching into historical research. Indeed, this is true of all of the relationships among unionization, Ghent and left strength.[11] But we can say that c. 1985, it is the *combination* of smallness, "leftness" and Ghent that is associated with the highest rates of unionization. The results also hint at a more specific interaction. The Ghent effect may be stronger in countries with medium left strength than in the fully fledged social democracies.

This "unsophisticated" method of presenting the data reveals regularities that MR does not. In the process it more effectively vindicates Rothstein's thesis by making clear precisely what he wanted to demonstrate: that the Ghent effect is large and not spurious, and that it comes into play in countries where other conditions are broadly favorable to unions. But these results do something else important, which is to point the interested researcher to the most fertile questions for selective case comparisons that might help nail down how important Ghent really is.[12] In particular, it must be questioned whether the Ghent system alone can explain the very large differences in density between otherwise well-matched countries: Belgium vs. the Netherlands, and Sweden and Denmark vs. Norway.[13]

The visibility of the relationship between variables and cases in the simple diagrammatic presentation favored here may thus draw attention to anomalous cases, which reveal limitations in the theoretical model. Attending to outliers from a regression analysis is sometimes also a way of identifying anomalies, but not of the kind discussed here – namely countries that do not "make sense" *when viewed in relation to other similar cases*. Tabular or graphical presentation of the dataset with named observations permits this; inspection and diagnostic testing of regression residuals does not.

COMPLEMENTING REGRESSION WITH OTHER TYPES OF ANALYSIS

Peter Hall and Robert Franzese (1998) have contributed to a significant subfield of comparative political economy which challenges the preeminence of economists in studying central banks and their impact on economic performance (Iversen, Pontusson, & Soskice, 1999). Hall and Franzese argue that while independent banks are always anti-inflationary, under certain institutional conditions their impact on the labor market is far less salutary. Unless wage setting is centralized and coordinated the bargainers will fail to internalize bank "signals", and the result will be higher rather than lower unemployment.

In testing their argument Hall and Franzese proceed in three stages. First, they demonstrate its plausibility by referring to the paradigm case of West Germany. Second, they use data for 18 OECD countries over the entire postwar period, presented in a simplified tabular format. Finally, they use MR to test a more elaborate model at several levels of aggregation ranging from full-period means (pure cross-section) to pooled annual data. The results of each one of these analyses are consistent with their argument that the impact of central bank status on unemployment is conditional on the structure of wage bargaining.

In their initial quantitative analysis, Hall and Franzese collapse measures of central bank independence (hereafter CBI) and wage coordination and cross-tabulate them. The results clearly confirm the hypothesized interaction effect. However the authors recognize that this effect could be an artifact, the result of some confounding influence like countries' wealth, economic openness or government composition. In practice, the result survives the application of controls for these variables using MR. Conditional parameter estimates show that the interaction between independence and coordination is substantively as well as statistically significant. Moreover, diagnostic

testing indicates that these results do not depend on the presence of any particular case.

Hall and Franzese's study deserves close attention precisely because it offers such a thorough application of MR, which moreover very sensibly builds on prior qualitative research on the German case. Yet it will be shown that the study's tabular results are misleading. Missing from these results is an element which proved crucial in probing Rothstein's study, namely, identification of the cases (countries). Another issue is how best to group continuous data into categories in order to reveal multivariate relationships. It was relatively easy to categorize Rothstein's variables intuitively, but this is not the case for Hall and Franzese's data. Although formal methods are sometimes used for this purpose (e.g. Goodman's (1981) test of "collapsability"), most researchers rely on commonsense ways of determining cutoff points: substantive familiarity with the cases, aggregation into categories of similar size or tailoring the categories to breaks in the distribution of observations. Hall and Franzese provide no explicit rationale for their cutoff points. Taking advantage of the availability of their dataset,[14] Chart 2 permits direct examination of the distribution of cases along the two institutional dimensions. Visual inspection of each dimension offers no indications of categories that could be "naturally" amalgamated. Further, observing the two-dimensional patterning of the countries one is not struck by any

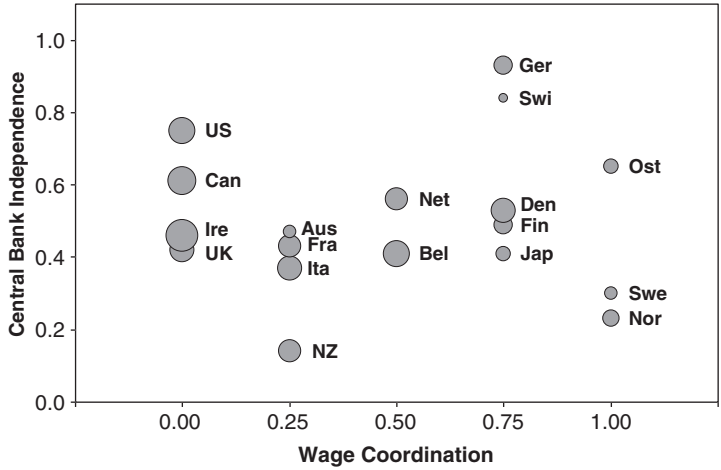

Chart 2. Institutional Configurations (X and Y axes) and Unemployment (Bubbles) (Based on Hall and Franzese).

Table 1. Institutional Effects on Unemployment (Derived from Hall and Franzese).

Wage Coordination	Central Bank Independence	UE 1955–1990	UE 1955–1973	UE 1984–1990
0.00	Lower (UK, Ire)	6.8	4.0	12.9
	Higher (US, Can)	6.2	4.9	7.6
0.25	Lower (NZ)	4.2	2.1	7.6
	Higher (Aus, Fra, Ita)	3.9	2.3	7.5
0.75	Lower (Den, Fin, Jap)	3.3	2.0	5.3
	Higher (Ger, Swi)	2.0	0.8	4.2
1.00	Lower (Nor, Swe)	2.0	1.8	2.6
	Higher (Ost)	2.2	1.8	3.5

Source: Hall and Franzese dataset (made available at the URL cited in note 15). Differences between the average unemployment rate for 1955–1990 reported here and in Table A.1 of Hall and Franzese (1998) are due to an error in the published table (Robert Franzese, personal correspondence, November 6, 2002). *Abbreviations:* Ire, Ireland; Can, Canada; NZ, New Zealand; Aus, Australia; Fra, France; Ita, Italy; Den, Denmark; Fin, Finland; Jap, Japan; Ger, Germany; Swi, Switzerland; Nor, Norway; Swe, Sweden; Ost, Austria.

obvious clustering. This suggests that Hall and Franzese may have erred in collapsing their institutional variables into dichotomies.

Is it possible without aggregation to discern the effects on unemployment, which were apparent in Hall and Franzese's aggregated figures (their Table 1)? The "bubbles" in our chart are proportionate in size to the mean unemployment rate for 1955–1990 in each country. Looking first for univariate effects, it is noticeable that as we move from left to right along the x-axis the jobless rate drops quite dramatically. No such clarity is evident when comparing unemployment rates at lower and higher levels of CBI (i.e. moving from the bottom to the top of the y-axis). Consequently, whereas unemployment is strongly correlated with wage centralization ($r = -0.74$) it is completely uncorrelated with CBI ($r = -0.07$).

The critical question though is whether "In nations where wage coordination is high, an increase in the independence of the central bank is associated with a very small increase in the rate of unemployment … . Where wage coordination is low, however, an increase in the independence of the central bank is associated with a substantial increase in the rate of unemployment" (Hall & Franzese, 1998, p. 518). Chart 2 provides no evidence for this proposition. In fact unemployment fails to rise with the extent of CBI at all levels of wage coordination. Apparently, the aggregation of Hall and Franzese's original data into categories inadvertently generated unfounded support for their hypothesis.

There is also an important substantive issue, which their analysis fails to reckon with. Studies that pool data from different points in time – whether by simple averages or complex panel analysis – implicitly assume stability in the causal relationships under consideration.[15] However, in the aftermath of the second oil shock, unemployment in most European economies rose dramatically while in North America it declined. Was this shift in international unemployment differentials, which persisted into the 1990s and beyond, accompanied by a change in the conditional impact of CBI? To find out, Table 1 compares unemployment in the postwar golden age (defined here as 1955–1973) with the period of global crisis from 1984–1990 (when the time series ends). Given that "our key institutional variables do not vary over time" (Hall & Franzese, 1998, p. 520), no attempt has been made to calculate sub-period measures of centralization and CBI. Further, to simplify the presentation Table 1 builds on the fact that *within each level* of wage coordination two groups of countries are discernable, one with higher CBI scores than the other.[16] The table permits us to evaluate whether relatively higher levels of CBI are associated with higher unemployment as coordination declines, in both the complete series and the two sub-periods.

The results confirm that the data for the postwar period as a whole do not fit expectations, but they show that in the period prior to 1974 there is some support for the predicted conditional relationship. This support would be stronger but for the fact that the two uncoordinated economies with low CBI, Ireland and the UK, experienced very different unemployment rates. The CBI "penalty" in this period thus turns heavily on the question of whether the role of the central bank can carry the main explanatory weight for the contrast between the UK, with well under 3% average unemployment; and the US and Canada with nearly 5%. I believe that a stronger explanation is provided by the absence of social democracy in North America compared to the paramount influence of the Labour Party on the terms of Britain's postwar settlement (Korpi, 1991). Turning to the later period of economic crisis, Table 1 shows that the results are at odds with Hall and Franzese's expectations. Among the least coordinated economies, North American unemployment was actually lower than in Britain or Ireland.

Perhaps one should not place too much weight on evidence concerning the gross effects of institutional context on economic performance. The authors of the study saw tabular analysis as only one building block in a longer evidentiary chain that included cross-country regressions controlling for key economic and political influences on unemployment (including the variable just referred to, government partisanship). Moreover with unusual thoroughness they ran these regressions not only on cross-sectional averages

for the entire postwar period, but also used pooled time series data in the form of either decade-long averages or annual observations. They report that the results of all of these tests were consistent with their leading hypothesis.

Nevertheless, there are reasons to take a cautious view of Hall and Franzese's multivariate analysis. With four control variables entered in aggregate cross-country regressions alongside the two institutional indicators and their interaction, the model is seriously overweight for application to only 18 cases. In theory, this limitation ought to be overcome once multiple observations for each country are combined at different time points. But for reasons that will be explicated in more detail in the next section of the paper, this is questionable. For instance, as we have just seen the postwar period 1955–1990 was far from homogeneous in its unemployment record. The models used by Hall and Franzese do control for over-time variability in the overall level of joblessness, but not for the equally plausible possibility that the determinants of unemployment altered over time.[17] In addition, whether tested in sparse cross-sectional format, decade-long panels or by pooling annual time series across countries, these regression models build on a great many empty cells. The vast majority of potential combinations of collective bargaining systems, CBI, union and left party strength and trading conditions have no empirical counterparts. As in most studies of this type, multiple time frames primarily add more cases to already-populated configurations.

The implications of limited diversity in the dataset utilized by Hall and Franzese are especially worrying for their most impressive evidence – decadal averages that simulate "what difference it makes". The authors' Table 4 presents expected levels of unemployment for 15 different institutional configurations, calculated by fixing control variables at their sample means. The results indicate that, as predicted, the effect of CBI is profoundly influenced by the degree of wage coordination. In completely uncoordinated systems unemployment is expected to be nearly *10 points higher* at maximum bank independence than at the minimum level of CBI. In completely coordinated systems there is a modest effect in the opposite direction. These results contrast very strongly with the uncontrolled effects that we have observed. However, it turns out that of the 15 cells in Hall and Franzese's table approximately two-thirds have no empirical counterparts. As it happens, the contrasts among the "extant" cells, while in the expected direction, are far more mild than those based on the hypothetical extremes of the institutional matrix.[18] Moreover the predicted levels of unemployment are seriously off the mark, higher than the real ones for decentralized systems and lower for the centralized ones.

There is a possible explanation for Hall and Franzese's inaccurate predictions of unemployment levels that also casts doubt on the veracity of their simulated effects of CBI (even for the realistic configurations). Both results may be traceable to the effect of elective affinities. As noted, Hall and Franzese adopted the typical procedure for such "what-if" exercises, allowing the explanatory variables of theoretical interest to vary while controlling for additional known influences by calculating their impact at mean levels. However as already noted in connection with Rothstein's study, different elements of the institutional context tend to cohere. For instance, coordination generally thrives in small, highly unionized economies with strong social-democratic parties but is stymied in liberal political economies with the opposite set of features. Consequently, by evaluating their control variables at the grand mean for all countries it is likely that Hall and Franzese inflated their predictions for the coordinated economies and understated them for the decentralized ones. The same bias may have exaggerated the deleterious effect of CBI in the decentralized context.

To sum up, Hall and Franzese present us with a study that is impressively well-rounded methodologically, integrating qualitative and quantitative research and moving stepwise from simple to sophisticated forms of numerical analysis. Despite this, their quantitative results are unconvincing. By failing to address temporality, limited diversity and elective affinities, their multivariate analyses almost certainly overstated the potency of the effects they sought to uncover. Their tabular analysis, based on questionable category groupings and abstracted from the cases under study, generated misleading results. In small-n comparative research even an analytical device as simple as a cross-tabulation needs to be applied with close attention to the data at hand. The pitfalls of the pooled regression models used by Hall and Franzese make it clear that more complex techniques offer no guarantee of yielding an empirically plausible account. While by now these pitfalls are well known they have not deterred comparative quantitative researchers from wholesale adoption of pooled MR as their technique of choice. The next section of the paper provides a fuller account of the problems this entails.

IS POOLING A PANACEA?

Some readers might view elements of the critique of the two articles discussed so far as just another illustration of a well-known problem: that because comparativists have "too many variables chasing too few cases", MR can only be applied either crudely (Rothstein) or else implausibly (Hall

and Franzese) in standard cross-sectional designs. My alternative approach might be criticized as a dishonorable retreat to rendering descriptive summaries of the data that are all too dependent on arbitrary decisions about how to group and present them. These critics would doubtless reject my argument that regression is fundamentally unsuited to macro-comparative analysis, and would prefer to focus their creative energies directly on solving the problem of insufficient cases (e.g. King, Keohane, & Verba, 1994, pp. 24, 30–31).

In this spirit, John Goldthorpe has argued that "*au fond* the small-N problem is not one of method at all but rather of data". Goldthorpe specifically recommends emulating the large number of researchers who "have 'pooled' data for the same set of nations for several different time-points. Observations – and degrees of freedom – are in this way increased …" (Goldthorpe, 1997, p. 8).[19] However, there are well-established reasons to believe that the most likely consequence of a turn to pooling is to muddy the causal waters still further. My critique proceeds in three stages. First, I explain why the rationale for using pooling as a means of adding statistical degrees of freedom is fundamentally flawed. Second, I demonstrate that creative attempts to overcome the difficulties of making causal inferences from pooled data are encouraging in principle but have been of limited practical benefit. Third, pooling encounters severe technical stumbling-blocks, and it is questionable whether growing methodological sophistication will reliably overcome these difficulties.

What does pooling entail?[20] Traditionally, quantitative macro-level research analyzed either "snapshots" of different countries at a single moment in time (cross-sectional data), or else period-to-period data for a single country (annual time series or sub-period averages). Pooled datasets merge these two views by "stacking" panels for multiple countries one on top of the other. Hence they embody both comparative variation between countries and dynamic variation over time. As a result analysts must contend with the technical complications characteristic of both cross-sectional and time series estimation, and practitioners face a bewildering range of technical problems and solutions. Even more basic is the well-grounded fear that pooling may be counter-productive "if thoughtful consideration is not given beforehand to the *meaning* of the aggregations in the pool" (Sayrs, 1989, p. 70).

Most comparative researchers who use pooled designs have been motivated by the traditional agenda of cross-sectional comparison, the desire to explain enduring differences between countries. These researchers implicitly regard each cross-sectional snapshot as just one more view of the same between-country variability. However, it has long been understood that the

effect of a given independent variable may be quite different in time series and cross-section "because the underlying causal structures differ" (Firebaugh, 1980, p. 333). For instance in their comparative and historical study of class conflict Korpi and Shalev (1980) observed that while temporal fluctuations in strikes followed an economic logic, with falling unemployment stimulating greater labor militancy, the cross-sectional variance followed a political logic, with lower unemployment operating as a disincentive to strong labor movements to employ the strike weapon. In this spirit, Hicks (1994, p. 171) promoted pooling precisely as a means of carrying out "systematic comparisons of cross-sectionally and longitudinally varying causal forces". But the reality is that most pooled designs utilize multiple cross-sections in order to fortify comparative generalizations, or multiple time series to fortify dynamic generalizations, on the implicit assumption that there is no difference in causality between the two dimensions.

A quite different, and more constructive approach to pooling, is to exploit the combination of comparative and over-time data in order to uncover and explain cross-national differences in over-time processes. Examples of this type of enquiry can also be found in studies of the political economy of class conflict (e.g. Hibbs, 1976; Shalev, 1979a). Time series regressions on strike activity in different countries yielded divergent results. Some scholars saw this simply as an antidote to exaggerated generalizations (Paldam & Pedersen, 1982). But others interpreted diverse parameter estimates as exemplifying the predictable effect of contextual forces on conflict dynamics (Snyder, 1975).

This has been the tack followed by the most thoughtful analysts of pooled datasets, Larry Griffin, Larry Isaac and their associates (Griffin, Barnhouse Walters, O'Connell, & Moor, 1986; Griffin, O'Connell, & McCammon, 1989). In what is still the best exposition of pooling for comparative political economists, Griffin et al. (1986) used annual data for 12 nations and 16 years to explore the effects of six economic and political variables on countries' expenditure on income maintenance. Their first finding was that the bulk of the variation in most of their independent variables was concentrated in *either* the time or cross-country dimension. This alone suggests that it would not have made sense to use a single model to explain both dimensions. And indeed, Griffin et al. found that "the average cross-national slopes and the average time series slopes … have very little in common" (p. 116). Even within the time and space dimensions, the contingency of causal relations could not be ignored. The results of annual cross-sections proved to be "extraordinarily unstable across years", even *contiguous* years (p. 111). While country-specific time series estimates were more stable, they

nevertheless seemed to "evoke markedly different processes" (p. 115). Despite these reasons not to treat pooled data simply as more data, it is rare for analysts to differentiate between over-time and cross-sectional effects or to take seriously the possibility of temporal or national specificity.[21] True, it is not uncommon for pooled models to include dichotomous variables intended to capture country or period effects. However, what these dummies actually measure are differences in the intercept or "baseline value" of the dependent variable in different countries or years. Interaction terms, far more costly in degrees of freedom, would be required to test country or period differences in *slopes*.[22]

For those mainly interested in explaining dynamic processes, on the other hand, pooling makes it possible to contemplate multiple explanations tailored to different contexts. The dynamics characteristic of a country or group of countries might be seen as both indicative of, and caused by, long-run (structural) differences. Griffin and his colleagues proposed a systematic methodology for this type of research. They suggested that time series parameters be estimated in regressions for individual countries. In a second round, these parameters would be treated as dependent variables to be explained cross-sectionally by broad-brush differences between countries (Griffin et al., 1986). While this technique may produce suggestive results (cf. Griffin et al., 1989), the credibility of the second-round results is, of course, dependent on the quality of the first round of time series estimates. Since these are typically based on short series, which may themselves be punctuated by causal heterogeneity, it is hard to be confident about these estimates.

Bruce Western (1996, 1998) has, however, offered an attractive approach to conceptualizing and estimating the type of multilevel design proposed by Griffin and his associates. Western (1996) sought to show that institutional factors like the presence or absence of corporatism could explain differences between countries in the dynamic effects of variables like government composition on fluctuations in unemployment.[23] He advocated a Bayesian approach to estimation that allows for possible contextual differences in causal dynamics, but differs in an important respect from Griffin's two-stage method. Western's technique permits estimates for individual countries to "borrow strength" from the whole sample. The implications of this are profound. It seemingly allows the analyst to take advantage of the more numerous observations and greater diversity afforded by pooled datasets, without having to assume identical causality in both time and space. Pooling would then be freed of most of the objections I have raised and, as Western explains, the issue of whether comparativists ought to generalize within or

beyond specific contexts would become a tractable empirical question rather than an epistemological conundrum.

Western's success in this regard is best assessed by considering the results of his own illustration, an analysis of unemployment using a pooled dataset for 18 OECD countries between 1964 and 1990 (Western, 1996). Impressively, he was able to demonstrate corporatism's implications in both the long and short run. Over the long run (cross-sectionally), corporatist countries were found to experience significantly lower rates of unemployment. From the dynamic (time series) perspective, the evidence supported the common claim that corporatism safeguards employment by improving the short-run tradeoff between wages and jobs. However, Western obtained puzzling findings for the dynamic effects of shifts in government composition. They appeared to show that in corporatist countries and other settings where collective bargaining is widespread, *increases in left party power cause unemployment to rise.* As always, the credibility of statistical conclusions needs to be checked against the cases. Chart 3 reproduces Western's estimates of the dynamic effects of changes in left cabinet representation. To highlight possible institutional consequences of the type Western was interested in, countries have been grouped using his indicators into three different settings –''unregulated'', ''regulated'' and ''corporatist''.[24]

At first sight, Chart 3 strongly confirms the finding that ''social democratic governments tend to raise unemployment where collective bargaining

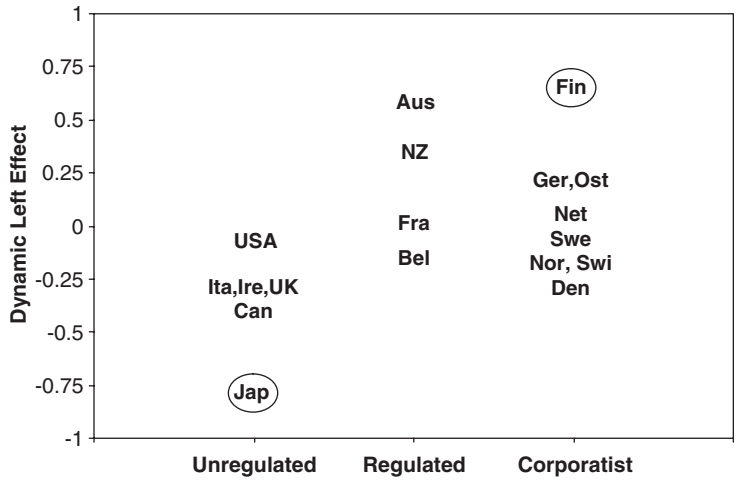

Chart 3. Western's Hierarchical Model of Unemployment.

coverage is extensive'' (Western, 1996, p. 25). However without two outliers – Japan and Finland – this tendency would be substantially weaker.[25] As it happens, the dynamic effects of leftwing governance in these two critical cases are highly problematic. During the period studied by Western, Finland experienced few significant shifts in the left's overall role in government. (What *did* vary was the relative role of the communist and socialist parties, a feature of government composition not measured in his study.) As for Japan, in the relevant period its left party representation was an unvarying zero.[26]

Western's hierarchical approach to utilizing pooled datasets holds out the possibility of harnessing their wealth of information while simultaneously respecting and even exploiting the difference between synchronic and diachronic causation. However, the key to reconciling these two objectives is ''borrowing strength''. In Western's words, ''Information from other countries will help provide an estimate for a coefficient in a particular country where, say, a given independent variable shows no variation'' (Western, 1998, p. 1240). This approach rests on a strong belief in the possibility of generalizing from ''populated cells'' to ''empty cells''. In the example at hand, the dynamic effects imputed to two cases generated extreme values that became the foundation on which a strong cross-national generalization was built. It is difficult to have confidence in such a generalization. This is a pity because Western's analytical strategy is very inviting to comparativists. Instead of merging repeated cross-sections simply in order to beef up the number of cases, he drew on the nested logic of multilevel modeling (Steenbergen & Jones, 2002). Moreover, he asked a question quintessential to the comparative method: do over-time relationships differ across countries and if so what stable differences between countries can predict those differences? Viewed this way, the pooled design offers an empirical way out of the controversy over whether causation is contextual (proper names are indispensable) or general (proper names surrender to variable names). In practice, however, since efficient estimation risks basing our ultimate conclusions on implausible counterfactual evidence, there may be no alternative to statistically unreliable country-by-country analyses.

Beyond issues concerning the analytical and practical justifications for the pooled design, as Stimson (1985, p. 945) pointed out at an early stage of the pooling revolution in political science, the technique suffers from ''a plethora of potential problems'' of a more technical kind. The validity of any regression estimate rests on assumptions about the statistical properties of the data, in particular the distribution of prediction errors. The characteristic problem for analysis of data collected at different time-points is serial correlation, which means that there is some kind of trend in the errors (e.g.

they tend to get bigger or smaller over time). For cross-sectional regressions comparing different units at a single moment in time, the typical challenge is "heteroskedasticity", meaning that the errors vary with the level of a predictor variable (e.g. corporatism may be a better predictor of unemployment in more corporatist than less corporatist countries). Further, cross-sectional errors may be "locally" interdependent. Examples commonly noted in comparative political economy are policy diffusion from one country to another through bilateral or multilateral coordination, or the economic impact of big countries on their smaller trading partners. From a technical point of view, pooled designs are the worst of both worlds. They expose regression estimates to the risks of trends in the error structure over time *and* systematic variation in the error term across units. To make matters worse these problems may appear in subtle combination, for instance heteroskedasticity could increase over time. In addition, if as we have suggested explanations may have differing applicability at different moments (or periods) and across different countries (or families of countries), then the errors will also be patterned by causal heterogeneity.

There are numerous ways to shield the accuracy and reliability of regression coefficients from these risks. However, many of them are atheoretical technical fixes that treat the deviant phenomena as "nuisance" rather than "substance" (Beck & Katz, 1996). In addition, the inferences generated by different remedies are often wildly dissimilar, while at the same time it is not entirely clear which remedy is the "right" one (Stimson, 1985). So far as causal heterogeneity is concerned, our earlier discussion has shown that conventional solutions to the problem are either wasteful of degrees of freedom or require heroic assumptions concerning the transferability of relationships from one context to another.

These issues are exhaustively treated in the pedagogical literature already referenced here (Beck & Katz, Griffin, Hicks, Stimson and others) as well as in standard econometrics texts. What bears emphasis is the questionable relationship between the costs and benefits of pooling, given that its technical complexities render it a risky and uncertain enterprise and at the same time one which imposes a steep and continuously rising learning curve. Most practitioners have responded to this dilemma by looking to "best practice" and following it faithfully – often with disastrous consequences. The breakthrough article by Alvarez, Garrett, and Lange (1991) referred to earlier utilized a Generalized Least Squares technique then regarded as state-of-the-art. However Beck et al. (1993) famously showed that because their dataset included more countries than time-points, this technique gravely inflated the significance of most parameter estimates. Subsequently,

Beck and Katz (1995) demonstrated that this problem invalidates the results of numerous well-known applications of the pooled design in comparative political economy and they introduced a new technique for estimating standard errors. Beck and Katz (1996) made the further suggestion that the dynamics generating serial correlation of time series errors should be modeled by including the lagged dependent variable as a predictor.

While Beck and Katz's proposals have subsequently become virtually canonical in modeling pooled data in political science, they have been sharply criticized by some other specialists. Achen believes that under typical conditions of high serial correlation and trended exogenous variables, ''the lagged [dependent] variable will falsely dominate the regression and suppress the legitimate effects of the other [independent] variables'' (Achen, 2000, p. 24). Specialists in international relations (where research designs are often much less constricted in degrees of freedom) have also engaged in heated debate concerning the use of pooled models.[27] An eminent econometrician has characterized Beck and Katz's prescriptions as ''not, strictly speaking, correct'', adding that ''the procedure of using OLS and reporting the 'panel corrected' standard errors is sweeping the problems under the rug'' (Maddala, 1998, pp. 60–61).

One of the few critical voices heard within comparative political economy is that of a European scholar, Bernhard Kittel. After reviewing many of its technical and practical deficiencies, Kittel (1999, p. 245) concluded that pooling adds statistical value to static cross-sectional regressions only ''under quite demanding conditions and to a very limited degree''. A more recent contribution by Kittel and Winner (2005) offers an exhaustive replication of a typical contemporary study, by Garrett and Mitchell (2001). On the basis of numerous alternative methods of testing and evaluation it is concluded that the results of this study are empirically unfounded. An even more sophisticated dissection of the same study by Plumper, Troeger, and Manow (2005) not only reveals additional technical deficiencies, but also challenges some of the main substantive conclusions drawn by Kittel and Winner.

The level of methodological expertise required to follow these kinds of debates over pooling has become prohibitive for many scholars. In rare but encouraging instances, analysts who are not professional methodologists have questioned technical orthodoxy because it generated results that simply did not make sense. Thus, Huber and Stephens (2001, Ch. 3) rejected the use of the lagged dependent variable as a predictor of social expenditure, arguing that it would have redefined their research question from assessing the long-run impact of differing political configurations to predicting short-run

fluctuations. Indeed, given the complexity of political dynamics and the poor likelihood of capturing them by crude measures like short-run changes in the proportion of the executive controlled by social or Christian-democratic parties, it is not surprising that in study after study political partisanship loses its explanatory efficacy once the design shifts from explaining levels to explaining dynamics. (See also Plumper et al., 2005; but compare Podesta, 2003.)

Because available techniques are constantly updated by statisticians and econometricians, quantitative political economists are tempted to devote much time and effort to refining their skills with pooled models. There are optimists who believe that such refinements can resolve the fundamental issues raised here, but in my judgment it is more likely that our theoretical understanding of causality will continue to far outstrip our measurement and estimation capabilities. Nevertheless, it should be noted that there has recently been a mushrooming of innovative statistical methods designed to address some of the problems discussed here.

Beck and Katz (2003) have suggested a variety of ways to systematically assess whether pooling multilevel data is justified, and Zorn (2001) has proposed a method of distinguishing between dynamic and cross-sectional effects. Braumoeller has developed new techniques for incorporating central goals of Ragin's approach into the regression framework – testing for the presence of necessary and sufficient conditions and modeling causal heterogeneity (Braumoeller & Goertz, 2000; Braumoeller, 2003). In a similar spirit, Girosi and King (2001) have devised a method of allowing explanations of over-time variation to vary across countries. But there is also bad news to report. Braumoeller's method of identifying multiple causal paths is only viable if the cases "represent all combinations of conditions" (Bear Braumoeller, personal correspondence July 23, 2005), while Girosi and King's technique seems to require a very large number of cases.

Finally, King, Tomz, and Wittenberg (2000) have proposed a simulation technique for increasing the amount of information on which statistical inferences are based, thereby enhancing their accuracy and certainty. King and his collaborators used this method to enthusiastically confirm a key finding of Geoffrey Garrett's influential book *Partisan Politics in the Global Economy*. Because this example poignantly illustrates the extent to which technique may outstrip data fundamentals, it deserves a closer look.[28]

Garrett's (1998) aim in using pooled regressions was to assess how the distribution of class power affects policy responses to globalization. These regression results were the basis for estimating expected levels of economic performance and public spending under different political configurations,

controlling for other relevant influences. Garrett's provocative findings (1998, Figs. 4.2, 5.2, 5.3, 5.4) appeared to demonstrate that in social-democratic and corporatist settings exposure to globalization pushes government spending upwards, while simultaneously enhancing these countries' superior record of unemployment and economic growth. King et al., (2000) argued that they were able to provide an even stronger foundation for these conclusions by generating 1,000 sets of simulated coefficients and expected values for the scenarios contrasted in Garrett's original study. Nevertheless, as shown by Garrett's own data (1998, Figs. 3.10, 3.12), at least until very late in the period of the investigation his key scenarios actually had no empirical counterparts.

Chart 4 provides a graphical view of the limited empirical variability of the institutional configurations tapped by Garrett.[29] The X and Y axes measure his two dimensions of exposure to globalization – trade openness and restrictions on capital mobility. The bubbles that represent each country are proportional in size to Garrett's index of "left-labor power". It is evident that the 14 countries included in the study fall into a limited number of groups that exhaust only part of the available property space. In the upper half of the chart we find a social-democratic cluster with high levels of capital restrictions. The countries with fewer restrictions fall into two main groups.

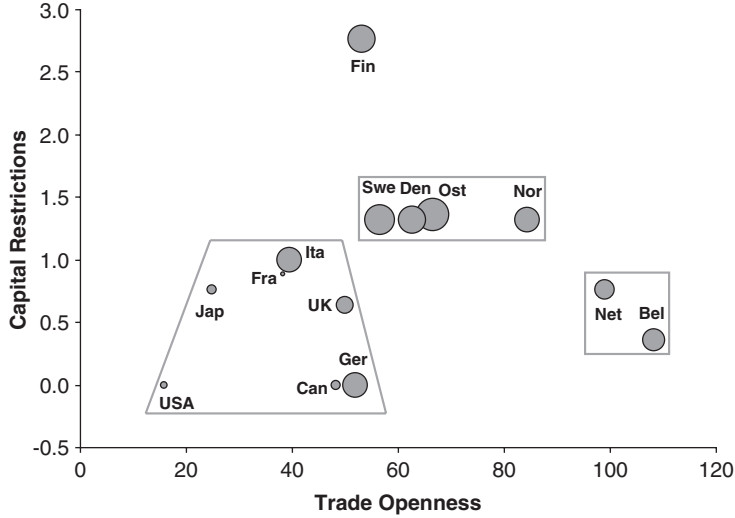

Chart 4. Garrett's Indicators of Globalization (X and Y Axes) and Left-Labor Power (Bubbles).

Belgium and the Netherlands are small states highly involved in trade (cf. Katzenstein, 1985). The remaining seven countries are all large and relatively autarchic with few capital controls, although they exhibit diverse levels of labor strength. As a result of this clustering of Garrett's key variables it is evident for instance that no countries have either very high left power and unrestricted mobility, or low power and high trade openness. Despite this, Garrett calculated estimates of how the outcomes of interest would respond to high levels of globalization under both high and low left-labor power.[30]

As King and Zeng (2002, p. 29) have argued in a different context, if "no evidence exists in our data with which to evaluate" a question, then "having time series–cross-sectional data with thousands of observations does not change this basic fact and will not make inferences like these any more secure". This reinforces my earlier contention that investments in hi-tech statistical analysis are of limited value in fields like comparative political economy, where both the number of cases and their variability are severely restricted. Indeed, as Beck & Katz have wisely cautioned, "complicated methods often move us away from looking at and thinking about the data" (Beck & Katz, 1996, p. 31).

TESTING THE "REGIME" APPROACH

If the typical practitioner of pooling is guilty of closing his or her eyes to causal complexity, in *The Three Worlds of Welfare Capitalism* Gøsta Esping-Andersen (1990) took complexity as his essential starting-point. Unusually, Esping-Andersen combined and made explicit the desiderata posited by diverse traditions of comparative research: (1) recognizing that there may be striking causal discontinuities across different contexts; (2) informing hypotheses about relationships between variables by drawing on knowledge of cases; and (3) using quantitative indicators to systematically test propositions across the entire universe of cases. As this paper has tried to explain, while obviously consistent with the third of these goals MR is markedly inhospitable to the first two.

In his quantitative analysis, Esping-Andersen adopted a two-stage approach reminiscent of Hall and Franzese – first descriptive analysis and then MR. He developed indices of "universalism", "decommodification" and "stratification" and used simple tables to show that his 18 OECD countries tend to fall into three distinct subgroups (Esping-Andersen, 1990, Tables 2.1, 3.3, 4.3). He then utilized MR to perform a causal analysis of cross-country variation in more than a dozen indicators, which were

regressed on political variables and in some cases control variables as well. However, Esping-Andersen's first technique (tabular analysis) was unnecessarily "soft", while the second (regression) is fundamentally in conflict with his analytical premises. There are better solutions, which exploit the rich data available on welfare states while respecting the theoretical assumption of causal complexity.

Esping-Andersen's tabular analysis relied heavily on his own judgment – both in the construction of indices and the identification of country clusters.[31] No *systematic* test was carried out of whether his ensemble of indicators of welfare state regimes actually do "hang together"; and if they do, whether countries indeed cluster in three distinct subgroups on underlying policy dimensions. It would have been a logical step to subject these claims to techniques like factor analysis, cluster analysis, correspondence analysis or multidimensional scaling that seek to reveal underlying proximities between different variables or cases.

Demonstration of the existence of three policy regimes was of course only a preliminary to Esping-Andersen's search for empirical support for his causal arguments. Central here was his view that different welfare state regimes embody different socio-political forces and state traditions. Using MR, Esping-Andersen did his best to demonstrate that his preferred (political) explanations garnered stronger empirical support than rival (e.g. demographic) explanatory variables. These empirical results are of questionable value, being based on regressions with 5 or 6 explanatory variables and only 18 cases. The key difficulty, however, is that asking whether political effects "matter" after "controlling for" other causes is a different and more banal question than what actually interested Esping-Andersen. As stated in his own critique of the quantitative, cross-sectional research tradition, "The dominant correlational approach is … marred by a frequent mismatch between theoretical intent and research practice" (Esping-Andersen, 1990, p. 106; see also Esping-Andersen, 1993).

The key causal argument of *The Three Worlds* is that *countries cluster on policy because they cluster on politics*. The regression approach, however, treats both policy and politics as continuous variables scattered across the whole spectrum of potential variation – not as a limited number of qualitatively different configurations with distinctive historical roots. In contrast to the causal thinking embodied in MR, Esping-Andersen would certainly *not* want to claim that, say, any discrete increment of Catholicism or absolutism ought to yield a discrete and uniform increment in the "corporativism" of pension programs. This is because *only* countries that are predominantly Catholic and/or have an absolutist past are expected to

exhibit the corporativist policy profile. By the same token, he would also not claim that the social policy of any given country may be understood precisely as the combined effect of Catholicism, absolutism and working class mobilization. (As in, "to make a loaf of bread combine 1 part yeast, 2 parts water and 10 parts flour …") On the contrary, a central purpose of his book was to demonstrate how the socialist, Catholic-Conservative and liberal political milieux have generated three different worlds of welfare. We may speculate that Esping-Andersen adopted MR out of deference to convention. He applied it as a blunt instrument for tapping gross differences between groups of countries, differences that arguably could have been more effectively conveyed by the use of tables and charts without the implication of constant linear effects across different contexts.[32]

How might Esping-Andersen have exploited his quantitative data without falling back on the conventional statistical paradigm, which is so out of keeping with the spirit of his analysis and his critique of earlier work? Three early investigations offered innovative suggestions. Ragin (1994a) carried out an elaborate study of pension policy using seven different explanatory variables, by means of his own technique of qualitative comparative analysis (QCA). In the same volume Kangas (1994) compared the performance of QCA with cluster analysis and traditional regression techniques for testing a simplified political model of the quality of sickness insurance. A third study, by Castles and Mitchell (1992), used descriptive data to build an alternative typology of four overall worlds of welfare capitalism. Methodologically, while Castles and Mitchell refrained from going beyond the presentation of simplified tabular data, both Ragin and Kangas utilized cluster analysis to assign countries to regimes. But these creative efforts ran into serious difficulties. Kangas had trouble finding the Liberal countries and Ragin was placed in the awkward position of having to assign one third of his countries to a "spare" category, which automatically excluded them from his analysis. In performing cluster analysis of countries both authors were forcing them to fit into a single regime, thereby predetermining an issue in need of empirical exploration.[33]

This issue has continued to bedevil subsequent research. A review by Arts and Gelissen (2002) concludes that Esping-Andersen's typology has received only partial support from the empirical literature. According to these authors the typology is challenged because a significant number of countries lie between regimes. In their view, the imperfect fit between country cases and Esping-Andersen's regimes indicates that more categories should be added to the typology. These conclusions reflect a common misunderstanding of the three worlds of welfare capitalism as referring literally to three discrete and mutually exclusive groupings of countries. However Esping-Andersen's

core analytical concept was not ''worlds'' but ''regimes'', that is to say *ideal-typical policy profiles*. As ideal-types they can be expected to resonate with the experience of some nations, but not to accurately describe all of them. On the contrary, hybrid cases are to be expected and the typology should help characterize and understand them more clearly. Finally, as already noted Esping-Andersen sees welfare regimes as reflecting three different political contexts. Hence the empirical usefulness of the regime typology should also be judged by whether countries' placement with respect to regimes is paralleled by their political characteristics.

To summarize: (1) It is policy profiles and not necessarily countries that ought to follow a tripartite division; (2) The proximity or distance of a country's policy profile from the three ideal-types should be matched by its political configuration; and (3) Policy regimes and their political underpinnings should together inform our understanding of individual countries. It follows that rather than seeking to assign countries to regimes, researchers should aspire to uncover underlying dimensions or profiles from cross-country correlations among policy indicators. Put differently, reducing a battery of *variables* to a few underlying dimensions is preferable to grouping *cases* into a few clusters. In light of this distinction it is not surprising that in Arts and Gelissen's review of empirical tests of Esping-Andersen's typology, the former methodology generated more supportive results than the latter.[34]

Practically speaking, researchers interesting in uncovering policy regimes can choose from a variety of techniques, including factor analysis (Shalev, 1996) and its cousin, Principal Components Analysis (de Beer et al., 2001; Hicks & Kenworthy, 2003).[35] One of the attractive features of these methods of reducing data into a smaller number of dimensions is that they are not at all fazed by a multiplicity of variables. On the contrary, while the existence of a wealth of explanatory variables is the acknowledged bane of cross-national research, multiple indicators are actually desirable if the purpose is to more parsimoniously characterize the dependent variable.

What underlying dimensions would we expect to find if Esping-Andersen's typology is correct? I believe that analytically his triplet of regimes rests on two dimensions of policy. One of them is a dichotomy that is unabashedly similar to Titmuss' (1974) classic distinction between ''residual'' and ''institutional'' welfare state principles, often illustrated by contrasting the United States with Sweden. A second dimension, dubbed ''corporativism'' by Esping-Andersen, captures the fragmented, hierarchical and status-preserving measures pioneered by Catholic-Conservative welfare states, measures that were anathema to *both* socialist and bourgeois forces. It follows that if

Esping-Andersen is right about there being three ideal-typical worlds, we should be able to parsimoniously characterize the policies of actual welfare states in terms of these two dimensions.[36]

Esping-Andersen's original *The Three Worlds* volume identified several different loci of welfare state variation: social rights, social spending, the public/private division, and employment policy. The present reanalysis is based on 13 of Esping-Andersen's policy indicators[37] and uses factor analysis to test whether the distribution of specific indicators follows the hypothesized two dimensions.[38] Factors are economical linear combinations of variables. They are generated in such a way that there is strong correlation between the variables with the highest "loadings" on a given factor, but minimal correlation between different factors (ideally they are completely uncorrelated or "orthogonal").[39]

The results of an unrotated principal component factor analysis are reported in Chart 5. The first two factors together account for the majority (nearly 60%) of the variance, good news for Esping-Andersen's model. The first factor, which runs between the East and West of the chart, evidently captures the residual/institutional dimension. It exhibits high positive loadings on public employment, active labor market expenditure, benefit equality and social security spending; and strong negative loadings on poor relief and indicators of the scope of private health and pension provision. The

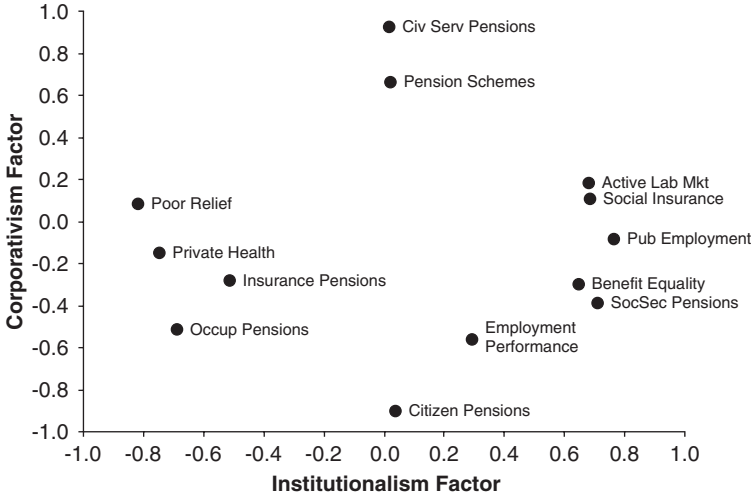

Chart 5. Two Factor Solution for Esping-Andersen Data.

second (North-South) factor signifies the corporativist dimension of policy. It has high positive loadings on the number of pension schemes and the prominence of civil service pensions, and a high negative loading on the role of ''citizen pensions'' (social security). The factors are not completely orthogonal, but the areas of overlap are intelligible. For instance, the results confirm that both the corporativist and institutional policy clusters are alienated from occupational pensions. They also imply that in the 1980s, when Esping-Andersen's data were collected, employment performance (low unemployment and high job creation) was stronger in the institutional regime than in the residual or corporativist regimes.

We now evaluate Esping-Andersen's political explanation for the origins of the three policy regimes. Chart 6 arrays the 18 nations in his study in accordance with their scores on our two factors. The evident linkage between policies and their political context generates an illuminating cross-national mapping. In particular, the findings support the clear distinction in Esping-Andersen's (1990) book between the following three families of nations:

- *Socialist:* The Scandinavian social democracies, characterized by levels of working class mobilization almost without peer in other Western nations.
- *Catholic-Conservative:* Continental European nations – Italy, France, Belgium, Austria and Ireland – which share an absolutist past, relatively late-blooming democracy and a largely Catholic population.

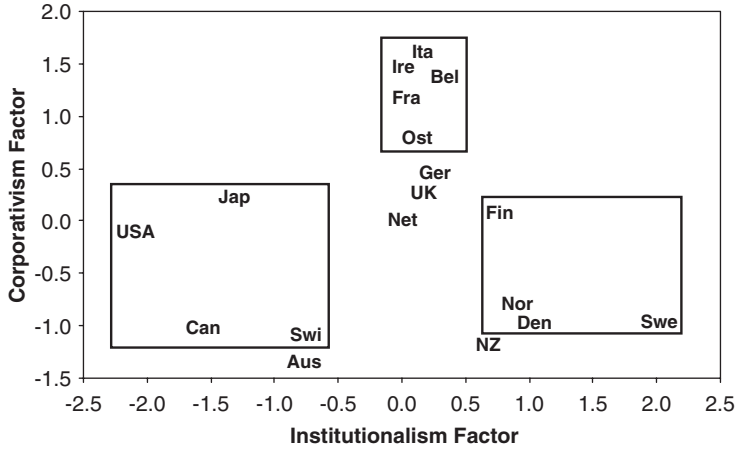

Chart 6. Social Policy Factor Scores.

- *Liberal:* The USA, Canada, Switzerland and Japan – in which working class mobilization is very weak and, in North America, the conservative heritage is absent.

The remaining five countries in Esping-Andersen's study are more difficult to classify. They have experienced moderate levels of working class mobilization but their state traditions are either close to the conservative group (Germany and the Netherlands), or were exposed in formative periods to liberal influences (the UK) or to the peculiar conditions of Antipodean settler societies (Australia and New Zealand).[40]

The fit between the three political clusters and countries' placement on the two policy factors is substantial. The liberal states and Australia have the most negative institutionalism scores, while the Scandinavian states along with New Zealand have the highest positive scores. Most of the remaining countries are conservative states, and as expected they score indifferently on institutionalism but above average on the corporativism factor. Two mixed cases (Britain and the Netherlands) score close to zero on both factors, confirming their ambiguous status rather than making us wish they would go away.[41]

Our analysis largely supports Esping-Andersen's vision of three different policy constellations powered by three different constellations of political power. The key point is that this empirical support was garnered without the mismatch between ontology and methodology that is exemplified by the use of MR in *The Three Worlds*. Esping-Andersen's analytical reliance on ideal-types in the context of an ambitious program of comparative and historical research recalls the classic sociological tradition, one which continues to inspire many comparativists. His goal of subjecting the theory of welfare state regimes to systematic empirical test was also admirable, but MR was ill-suited to this task. I have tried to show that methodological alternatives are available which do not require sacrificing either quantification or the ambition of supporting causal claims through empirical generalization.

CONCLUSION

Despite considerable methodological debate and innovation among comparativists in recent years, MR remains by far the predominant mode of numerical data analysis and most of its critics see qualitative analysis (whether formal or not) as the only real alternative. This paper seeks to promote a third way. I recognize that Charles Ragin's innovations, QCA

and more recently "fuzzy-set" analysis (Ragin, 1987, 2000), point to another strategic alternative. Ragin's techniques constitute a synthesis of the qualitative and quantitative traditions aimed at explicitly testing the kind of "causal pathways" arguments typical of classical comparative-historical research in the genre of Weber, Moore, Rokkan and Skocpol. The desire to systematically evaluate the evidence for such arguments is not new (Somers, 1971). But Ragin (1987) is the first to have offered formal procedures for parsimoniously identifying the regularities that underlie a series of case configurations.

Ragin's methods are not "qualitative" in the sense of relying on the interpretive skills of analysts wading knee-deep in thick description. If anything, as Griffin and Ragin (1994, p. 10) have insisted, QCA is more like MR: both apply rules that are independent of the researcher, and both treat cases as "discrete, multiple instances of more general phenomena". While controversial,[42] in principle Ragin's methods have great advantages because of their fidelity to principles of case-oriented analysis. One feature, which is especially valuable in the context of small-n macro-comparisons, but lacking in MR, is visibility of and dialog with the cases. However, the advantages of Ragin's techniques are not exclusive to his methods. My reanalysis of diverse MR-based studies in this paper poses alternatives to both QCA and MR. In closing, I incorporate these suggestions into a summary statement of the major options (other than Ragin's methods) open to quantitative researchers who are troubled by the limitations of MR.

1. *Refinement.* This is the optimistic approach best represented in the present survey by Bruce Western's variant of pooled regression. However, the discovery of a serious limitation of Western's method heightens our pessimism concerning the payoffs from technical refinement. Western was unable to resolve the problem of simultaneously combining and separating cross-country and over-time effects. This is only one issue in MR analysis for which political scientists have sought inspiration from their technically more advanced counterparts in economics and statistics. In this connection it is sobering that G.S. Maddala, one of the most respected figures in the econometric world, considers its achievements both modest and contested. Moreover, he believes that leading political methodologists have mistakenly or misguidedly emulated shallow econometric fads (Maddala, 1998). Sadly, Maddala's criticisms and cautions appear to have fallen on deaf ears.[43] More encouraging is the emerging trend, noted earlier, of efforts to find original econometric solutions to some of the lacunae of MR highlighted in this paper. However it is too

early to predict the fate of these new methods. They are as likely to spark new rounds of technical debate or simply be ignored as to triumph over researchers' customary methodological conservatism.

2. *Triangulation.* This means combining MR with other types of analysis – quantitative, qualitative or both. Hall and Franzese adopted this approach to strengthen their empirical case by citing the convergent findings produced by different ways of researching the same topic. Alternatively, the complementarity of different approaches may rest on the distinctive contributions made by each one of them. This is the strategy underpinning Esping-Andersen's work on welfare states, and several ambitious comparative and historical studies by John Stephens, Evelyn Huber & their collaborators (Rueschemeyer et al., 1992; Huber & Stephens, 2001; see also Huber, Ragin, & Stephens, 1991; Rueschemeyer & Stephens, 1997). They have proposed that comparative research be based on dialog between broad-spectrum quantitative comparisons and historically oriented country studies (see also Esping-Andersen, 1993). The results of MR should be confronted by both theory and knowledge of cases, and if causal anomalies arise they should be put to the test of historical process-tracing across multiple countries.

   This approach is attractive but also very demanding; it is virtually impossible without long-term collaborative research. In practice, when triangulation does occur it is usually more modest than in the hands of Stephens and his collaborators. Occasionally, researchers employ multiple statistical techniques to analyze the same data or problem, looking for convergent results (e.g. the use of both MR and QCA by Kangas, 1994; Ebbinghaus & Visser, 1999). In addition, some book-length studies have utilized both case-studies and pooled regressions, using the qualitative materials either to illustrate their argument (e.g. Boix, 1998) or as a genuine complement to statistical findings (e.g. Swank, 2002).[44] This kind of hybrid analysis is a welcome development, but the insularity of different methodological traditions and the difficulty of publishing multimethod articles in journal format both limit its likely spread.

3. *Substitution*. The present paper has promoted the use of alternative methods of quantitative analysis as another strategy for dealing with the problems of MR. The second and third sections presented tables or tree diagrams in which countries are clearly identified.[45] It was shown that these simple techniques overcome some of the most unattractive limitations of MR while incorporating key elements of the case-oriented approach. They are able to plainly convey complex analytical ideas like elective affinities and causal hierarchies. They also draw attention to cases

deserving of additional, more focused comparative scrutiny, which is a blind spot of most other methods. I have suggested as well that, provided they fit researchers' theoretical assumptions, there is no reason why inductive multivariate statistical methods should not be exploited by comparativists. The utility of factor analysis in clarifying the evidence for Esping-Andersen's approach to welfare state diversity was the illustration offered here,[46] but many other methods of exposing latent variables are available. Such methods hold the delicious promise of turning the traditional handicap of more indicators than cases from a burden into an asset. Of course, generating better measures of the phenomena of interest cannot resolve the difficulties of testing causal explanations in cross-national research. It has been argued here that data analysis aimed at theory testing and theory building should strive to reveal how the cases are located in relation to each other as well as to cause and effect variables.

ACKNOWLEDGMENTS

Participants at several workshops and conferences where this paper was presented were kind enough to offer comments and advice. In addition I wish to acknowledge valuable input from Neal Beck, Frank Castles, David Freedman, Peter Hall, Robert Franzese, Orit Kedar, Bernhard Kittel, Walter Korpi, Noah Lewin-Epstein, Hadas Mandel, Jonathon Moses, Herbert Obinger, Meir Shabat, Aage Sorensen, David Soskice, John Stephens, Uwe Wagschal and Bruce Western.

NOTES

1. In contrast, interest in formal methods tailored to small-n research is relatively strong in Europe, with an extensive website devoted to the topic (http://www.compasss.org).

2. It is, of course, debatable just how bounded the research universe is or should be. Conventionally, comparative policy studies focus on the approximately 18 rich, capitalist countries with longstanding democratic polities and non-trivial populations. Such conventions may be theoretically arbitrary and should always be open to challenge. Many studies have incorporated Greece, Spain and Portugal after democratization (and more practically, after their inclusion in OECD databases). Other candidates for inclusion in studies of what have until now been known as "the Western nations" might be found in the former Soviet bloc states, Latin America and East Asia. There are good arguments both for and against expanding the universe of comparative studies. For instance, compare Geddes (1990) and Boyer (1997).

3. Even the well-known injunction of Przeworski and Teune (1970) that comparativists should strive to turn the proper names of countries into the abstract names of variables did not entirely contradict this view. It should be remembered that Przeworski and Teune were railing against the dominance of comparative politics by "area studies" specialists and urging their colleagues to avoid particularizing arguments that could easily strait-jacket both theory and comparison. Many contemporary advocates of case-oriented analysis (including Ragin) would have no quarrel with this assessment.

4. The criticism here is not the standard one that quantification over-simplifies complex reality. There is always a trade-off between accuracy and parsimony in social research, whether analysis uses quantitative measures or narrative representations. The point is that the use of MR encourages what may well be a mistaken belief that our measures are precise and continuous.

5. An exception is Amenta and Poulson's (1996) use of MR and QCA in a comparative study of the American states. This exception proves the rule, however, since the measurement of such concepts as "administrative strength" was possible only because this research compared sub-national units of a uniform national entity.

6. More recently, Lieberson and Lynn (2002) have offered a more fundamental critique of the quasi-experimental epistemology prevalent in sociology and similar disciplines.

7. Abbot has offered an elegant formulation of this problem. Variable-oriented approaches "seek to understand the social process by developing linear transformations from a high-dimensional space (of 'main effects' and occasionally of interactions between them) into a single dimension (the dependent variable) … Now this strategy … is useful only if the data space is more or less uniformly filled" (Abbott, 1997, p. 86).

8. I excluded picayune Iceland with only 80,000 potential union members. I also replaced Rothstein's left party representation indicator borrowed from Wilensky (1981) and based on the entire 1919–1979 period which includes disruptions and discontinuities during the interwar years. Since the unionization data reveal that cross-national differentials stabilized after about 1965, I treat the first two postwar decades as the politically formative period. Figures for average left cabinet strength in this period were taken from the dataset assembled by Korpi and Shalev (1980). It turns out that these modifications strengthen the effect of the Ghent variable.

9. Potential membership was dichotomized after exploratory charts revealed that it had an evident threshold effect on unionization. With the exceptions of only Switzerland and the Netherlands, all small countries (no more than 5 million potential members) had more than 50% density, while all the large countries (10 million and up) scored less than 50%. Within these two categories no relationship was discernible between the two variables.

Left strength was grouped into four categories that reflect breaks in its distribution. "None" were cases with zero or trivial (up to Japan's 4%) left party representation in cabinet; "weak" 7–15%; "medium" 22–29% plus an intermediate case (the UK) with 36%; "strong" 45% or more.

10. On the other hand, left strength discriminates only weakly between the unionization rates of small countries, and not at all between the large ones (except perhaps for the British case).

11. It should be pointed out however that although only careful comparative historical research can speak to this type of causal question, as a result of theoretical, evidentiary and interpretive differences there is no guarantee that a consensual account will emerge. On the contrary, a sizable literature relevant to the role of the Ghent system has failed to arrive at clear-cut conclusions. In addition to Rothstein's article, see Hancke (1993), Scruggs (2002), Oskarsson (2003) and Swenson (2002).

12. The significance of these kinds of anomalies for scientific progress has been strongly argued by Rogowski (1995).

13. Visser (1992) has suggested that most of the vast difference between Belgian and Dutch unionization can be attributed to the fact that Dutch unions have no presence in the workplace. The origins of Norway's laggard status are less clear, but they might be traceable to the Norwegian union movement's lesser effectiveness in some of the sectors that grew from the 1960s, when Norway's density plateaued while Sweden's entered a long period of growth. Data collected by D'Agostino (1992) reveal substantial gaps in union density favoring Sweden in the following categories: women, private sector trade and services, and white-collar workers.

14. See http://www-personal.umich.edu/~franzese/h&f_data.TXT

15. The assumption of causal stability over time can be relaxed, but as in Hall and Franzese's study it typically is not. Although Hall and Franzese tested for effects of different data periodicities (annual, decadal or full-period), they did not examine the consistency of their model across sub-periods.

16. Except at the intermediate level of coordination (0.5), where there is only a small difference in CBI between Belgium and the Netherlands. Since the Hall–Franzese model in any case makes no specific prediction for this configuration I do not include it in Table 1.

17. Hall and Franzese included dummy variables for each decade or year in their pooled regressions, but they were not interacted with any of the causal variables.

18. Hall and Franzese's simulation estimated 9.7 percentage points more unemployment at the highest than the lowest levels of CBI in decentralized systems, whereas the simulated gap between the actually existing poles of CBI is only 2.4 points.

19. Goldthorpe recommends even more strongly that researchers widen the "geographical and sociocultural range" of their research. In this matter, however, it cannot be said (as it can of pooling) that the recommended solution is a popular one. As Goldthorpe concedes, data quality and availability are limited outside of the bloc – the OECD countries – which interests his intended audience (and mine). Moreover it is widely understood that what might be called the "specification costs" of going beyond the OECD (additional casual factors and alternative causal paths) usually outweigh the potential benefits. Even in a theoretically developed field (the economics of growth) where it was possible to gather comparable data for a stunning 119 countries, Levine and Renelt (1992) found themselves hopelessly unable to use cross-national regressions to adjudicate between rival theories.

20. In political science, where pooling has been most popular, foundational treatments are Stimson (1985), Sayrs (1989) and Hicks (1994).

21. In Kittel and Winner's (2005, p. 8) pithy summary, "practically all published contributions to comparative political economy using panel data assume poolability by fiat".

22. A compromise that is more sensitive to context but less exhaustive of degrees of freedom, is to permit both intercept and slope parameters to vary across *groups* of nations or years. For a rare example see O'Connell (1994).

23. Western's 1998 article is the published version of a paper dated December 1996 which was circulated electronically (Western, 1996). In the final version a partly different empirical example was substituted for the one in the preprint version (economic growth became the dependent variable instead of unemployment). I refer here to the findings reported in the 1996 version since they highlight a problem, which I believe to be endemic to the technique that Western proposed.

24. ''Unregulated'' labor markets are those in which no more than half of the workforce was covered by collective bargaining. Classification of the other countries was based on Western's dichotomous measure of corporatism. I adopted Western's classification of Switzerland as corporatist even though it had less than 50% collective bargaining coverage.

25. For example, if the time-series coefficients for left cabinet strength are regressed cross-nationally on collective bargaining coverage, the resulting coefficient is 1.00 ($t = 3.4$) for all countries but only 0.59 ($t = 1.5$, non-significant) without Japan and Finland.

26. Western (1996, p. 26) indeed noted that the left government variable for Japan was constant and counseled against ''substantive interpretation'' of the Japanese result. However the statistical generalization yielded by the cross-sectional level of his hierarchical model was clearly based in part on the Japanese case.

27. The debate took place in a special issue of *International Organization*. For a judicious summary, see the contribution by King (2001).

28. For additional wider-ranging critiques of Garrett's study, see Hay (2000) and Moses (2001).

29. Chart 4 is based on averages for the full period of Garrett's investigation (1966–1990) which I calculated using the dataset on his Yale University website (http://pantheon.yale.edu/~gmg8) in August 2000.

30. In a private communication dated March 7, 2001, Garrett concurred that with one temporary and partial exception no country in his dataset with a strong left exhibited weak capital controls, but he argued that out-of-sample experience in the 1990s subsequently vindicated his predictions.

31. Recent research has sought to replicate and/or update Esping-Andersen's decommodification scores. Lyle Scruggs is highly critical of Esping-Andersen's methodology (see his ''Comparative Welfare State Entitlements'' website at http://sp.uconn.edu/~scruggs/wp.htm and Scruggs and Allen (2006)), while Bambra (2004) reports similar results to Esping-Andersen using updated sources.

32. In his more recent work Esping-Andersen (1999) adopted a different variant of MR, multinomial logistic regression. In keeping with the spirit of the regime approach, this technique has the advantage of permitting explanatory weights to vary across different categories of the dependent variable. But in the context of cross-national research of this type, the category-specific coefficients must be estimated on ludicrously small numbers of cases.

33. Both of the standard approaches to clustering – hierarchical and *k*-means – allocate cases to mutually exclusive clusters, although they provide information on how well each case fits its group.

34. For an exception published after Arts and Gelissen's survey see Powell and Barrientos (2004).

35. In addition to the techniques mentioned, other methods of revealing underlying "dimensions" are MDS (multidimensional scaling) and CA (correspondence analysis). These methods are appropriate to ordinal or even nominal data and do not assume linear relationships among variables. Another flexible option, utilized by de Beer, Vrooman, and Wildeboer Schut (2001), is the non-linear version of Principal Components Analysis known in SPSS as PRINCALS. Since the results generated by factor analysis in my original study (Shalev, 1996) are replicated using other methods, they remain the basis for the findings reported here.

36. Hicks and Kenworthy (2003) also advocate a dimensional approach to verifying Esping-Andersen's typology. However, these authors seem to interpret their finding that welfare state indicators reduce to two dimensions as evidence against the existence of three regimes. In contrast, I argue that if Esping-Andersen is correct then policies (again – *not* countries) should follow two underlying continua which provide the coordinates of the three regimes.

37. In view of objections raised by Castles and Mitchell (1992) concerning his coding of Australia and New Zealand, I did not include two of Esping-Andersen's key indicators – "decommodification" and "universalism" (Esping-Andersen, 1990, Tables 2.2, 3.1). The 13 indicators summarized in Chart 4 were obtained as follows; references are to Esping-Andersen (1990): social insurance spending (Table 5.1, source data from the author); number of pension schemes ("Corporatism" in Table 3.1), Civil Servants' pensions ("Etatism" in Table 3.1), benefit equality (Table 3.1); "poor relief" (Table 3.1); the public–private division in health (Table 3.1) and pensions (Table 4.3); "full-employment performance" (Table 5.9, data from the author). Active manpower program expenditures relative to GDP (c. 1975) and public employment as a percentage of total employment (in 1980) are mentioned in Esping-Andersen (1990) and analyzed in Esping-Andersen (1985), but the source data were obtained directly from the author.

38. The findings presented below were originally reported in the introduction to Shalev (1996).

39. Thus the researcher hopes that each item will load high on only one of the factors. The procedure known as factor "rotation" is designed to encourage this to happen, but I opted here for the more pristine test of an unrotated analysis.

40. On the complexity and importance of state traditions as a causal variable in comparative research, see Crouch (1993).

41. The contradictions of the British welfare state are well known, and if anything they are exemplified by the contrasting experiments launched by Thatcher and Blair. On the mixed Dutch case, see Wildeboer Schut, Vrooman, and de Beer (2001).

42. QCA has been vociferously criticized, particularly for its dichotomous measurement of variables and abandonment of probabilistic generalizations in favor of deterministic ones (see especially Lieberson, 1994, 1991; Goldthorpe, 1997). Ragin's "fuzzy logic" technique at least partially answers these criticisms.

43. In quest of evidence for political methodologists' inattention to critiques of pooling, I used the Social Sciences Citation Index to search for articles that cited Maddala (1998). As of July 1, 2005, there were only five citations, two of them authored by political methodologists. In contrast, another article by Maddala

(on unit roots and cointegration) published the same year has been cited more than 100 times.

44. In an intriguing recent contribution, Gordon and Smith (2004) offer a method for introducing qualitative findings into causal statistical models (which however has already given rise to debate; see *Political Analysis*, Vol. 13, No. 3).

45. For an independent application of these techniques, see Marks and Wilson (2000, pp. 445, 450).

46. See also Leertouwer (2002), who used factor analysis to uncover the latent dimensions of corporatism and central bank independence by analyzing a wide range of empirical indicators proposed by previous researchers.

REFERENCES

Abbott, A. (1997). On the concept of turning point. *Comparative Social Research*, *16*, 85–105.

Abbott, A. (1998). The causal devolution. *Sociological Methods and Research*, *27*(2), 148–181.

Achen, C. H. (2000). Why lagged dependent variables can suppress the explanatory power of other independent variables. Paper presented at the annual meeting of the political methodology section of the American Political Science Association, UCLA, July 20–22, http://web.polmeth.ufl.edu/papers/00/achen00.zip

Alvarez, R. M., Garrett, G., & Lange, P. (1991). Government partisanship, labor organization, and macroeconomic performance. *American Political Science Review*, *85*(2), 539–556.

Amenta, E. (1993). The state of the art in welfare state research on social spending efforts in capitalist democracies since 1960. *American Journal of Sociology*, *99*(3), 750–763.

Amenta, E., & Poulsen, J. D. (1996). Social politics in context: The institutional politics theory and social spending at the end of the New Deal. *Social Forces*, *75*(1), 33–59.

Arts, W., & Gelissen, J. (2002). Three worlds of welfare capitalism or more? A state-of-the-art report. *Journal of European Social Policy*, *12*(2), 137–158.

Bambra, C. (2004). Weathering the storm: Convergence, divergence and the robustness of the 'worlds of welfare'. *The Social Policy Journal*, *3*(3), 3–23.

Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, *89*(3), 634–647.

Beck, N., & Katz, J. N. (1996). Nuisance vs. substance: Specifying and estimating time-series cross-section models. *Political Analysis*, *6*, 1–36.

Beck, N., & Katz, J. N. (2003). Random coefficient models for time-series–cross-section data. Paper presented at the 2nd ECPR conference, Marburg (Germany), 18–21 September, http://www.essex.ac.uk/ecpr/events/generalconference/marburg/papers/6/2/Katz.pdf

Beck, N., Katz, J. N., Alvarez, R. M., Garrett, G., & Lange, P. (1993). Government partisanship, labor organization, and macroeconomic performance – a corrigendum. *American Political Science Review*, *87*(4), 945–954.

Berg-Schlosser, D. (2002). Macro-quantitative vs. macro-qualitative methods in the social sciences – testing empirical theories of democracy. Paper presented at the XV world congress of the International Sociological Association, Brisbane, July 7–3.

Berg-Schlosser, D., & De Meur, G. (1997). Reduction of complexity for a small-N analysis: A stepwise multi-methodological approach. *Comparative Social Research*, *16*, 133–162.

Boix, C. (1998). *Political parties, growth and equality: Conservative and social democratic economic strategies in the world economy*. Cambridge: Cambridge University Press.

Boyer, R. (1997). Comparing Germany and Japan: Methodological issues, main findings from the regulation approach, and an agenda for further research. Paper presented at the conference on Germany and Japan: The Future of Nationally Embedded Capitalism in a Global Economy, University of Washington, Seattle, April 10–12.

Braumoeller, B. F. (2003). Causal complexity and the study of politics. *Political Analysis*, *11*(3), 209–233.

Braumoeller, B. F., & Goertz, G. (2000). The methodology of necessary conditions. *American Journal of Political Science*, *44*(4), 844–858.

Cameron, D. R. (1984). Social democracy, corporatism, labour quiescence and the representation of economic interest in advanced capitalist society. In: J. H. Goldthorpe (Ed.), *Order and conflict in contemporary capitalism* (pp. 143–178). Oxford: Clarendon Press.

Castles, F. G. (1993). *Families of nations: Patterns of public policy in western democracies*. Brookfield, VT: Dartmouth.

Castles, F. G., & Mitchell, D. (1992). Identifying welfare state regimes: The links between politics, instruments, and outcomes. *Governance*, *5*(1), 1–26.

Crouch, C. (1993). *Industrial relations and European state traditions*. Oxford: Clarendon Press.

D'Agostino, H. (1992). *Why do workers join unions? A comparison of Sweden and OECD countries*. Stockholm: SOFI (Swedish Institute for Social Research).

de Beer, P., Vrooman, C., & Wildeboer Schut, J. M. (2001). *Measuring welfare state performance: Three or two worlds of welfare capitalism?* Luxembourg Income Study Working Paper no. 276, May.

Ebbinghaus, B., & Visser, J. (1999). When institutions matter: Union growth and decline in Western Europe, 1950–1995. *European Sociological Review*, *15*(2), 135–158.

Esping-Andersen, G. (1985). Power and distributional regimes. *Politics and Society*, *14*(2), 223–256.

Esping-Andersen, G. (1990). *The three worlds of welfare capitalism*. Cambridge: Polity Press.

Esping-Andersen, G. (1993). The comparative macro-sociology of welfare states. In: L. Moreno (Ed.), *Social exchange and welfare development* (pp. 123–136). Madrid: CSIS.

Esping-Andersen, G. (1999). *Social foundations of postindustrial economies*. Oxford: Oxford University Press.

Esping-Andersen, G., & Korpi, W. (1984). Social policy as class politics in post-war capitalism: Scandinavia, Austria, and Germany. In: J. H. Goldthorpe (Ed.), *Order and conflict in contemporary capitalism* (pp. 179–208). Oxford: Oxford University Press.

Firebaugh, G. (1980). Cross-national versus historical regression models: Conditions of equivalence in comparative research. *Comparative Social Research*, *3*, 333–344.

Franzese, R. J. (2001). *Macroeconomic policies of developed democracies*. New York: Cambridge University Press.

Freedman, D. (1985). Statistics and the scientific method. In: W. M. Mason & S. E. Fienberg (Eds), *Cohort analysis in social research: Beyond the identification problem* (pp. 345–390). New York: Verlag.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics*, *12*, 101–223.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology*, *21*, 291–313.

Garrett, G. (1998). *Partisan politics in the global economy*. Cambridge: Cambridge University Press.

Garrett, G., & Lange, P. (1989). Government partisanship and economic performance – when and how does who governs matter. *Journal of Politics*, *51*(3), 676–693.

Garrett, G., & Mitchell, D. (2001). Globalization, government spending and taxation in the OECD. *European Journal of Political Research*, *39*(2), 145–177.

Geddes, B. (1990). How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political Analysis*, *2*, 131–150.

Girosi, F., & King, G. (2001). *Time series cross-sectional analyses with different explanatory variables in each cross-section. The Global Burden of Disease 2000 In Aging Populations.* Research Paper no. 10, Harvard University, July, http://www.hsph.harvard.edu/burdenofdisease/publications/papers/Time Series Cross-Sectional Analyses.pdf

Goldthorpe, J. H. (1997). Current issues in comparative macrosociology: A debate on methodological issues. *Comparative Social Research*, *16*, 1–26.

Goodman, L. A. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology*, *87*(3), 612–650.

Gordon, S. C., & Smith, A. (2004). Quantitative leverage through qualitative knowledge: Augmenting the statistical analysis of complex causes. *Political Analysis*, *12*(3), 233–255.

Griffin, L., Barnhouse Walters, P., O'Connell, P., & Moor, E. (1986). Methodological innovations in the analysis of welfare-state development: Pooling cross sections and time series. In: N. Furniss (Ed.), *The futures for the welfare state* (pp. 101–138). Bloomington: Indiana University Press.

Griffin, L. J., O'Connell, P. J., & McCammon, H. J. (1989). National variation in the context of struggle: Post-war class conflict and market distribution in the capitalist democracies. *Canadian Review of Sociology and Anthropology*, *26*(1), 37–68.

Griffin, L., & Ragin, C. C. (1994). Some observations on formal methods of qualitative analysis. *Sociological Methods & Research*, *23*(1), 4–21.

Hall, P. A. (2003). Aligning ontology and methodology in comparative research. In: J. Mahoney & D. Rueschmeyer (Eds), *Comparative historical research in the social sciences,* (pp. 373–404). New York: Cambridge University Press.

Hall, P. A., & Franzese, R. J., Jr. (1998). Mixed signals: Central bank independence, coordinated wage bargaining, and European monetary union. *International Organization*, *52*(3), 505–535.

Hancke, B. (1993). Trade union membership in Europe, 1960–1990 – rediscovering local unions. *British Journal of Industrial Relations*, *31*(4), 593–613.

Hay, C. (2000). Globalization, social democracy and the persistence of partisan politics: A commentary on Garrett. *Review of International Political Economy*, *7*(1), 138–152.

Hibbs, D. A., Jr. (1976). Industrial conflict in advanced industrial societies. *American Political Science Review*, *70*(4), 1033–1058.

Hibbs, D. A., Jr. (1978). On the political economy of long-run trends in strike activity. *British Journal of Political Science*, *8*(2), 153–175.

Hicks, A. (1994). Introduction to pooling. In: T. Janoski & A. M. Hicks (Eds), *The comparative political economy of the welfare state* (pp. 169–188). Cambridge: Cambridge University Press.

Hicks, A., & Kenworthy, L. (2003). Varieties of welfare capitalism. *Socio-Economic Review*, *1*(1), 27–61.

Huber, E., Ragin, C., & Stephens, J. D. (1991). Quantitative studies of variation among welfare states: Towards a resolution of the controversy. Paper presented at the ISA conference on comparative studies of welfare state development, Helsinki.

Huber, E., Ragin, C., & Stephens, J. D. (1993). Social democracy, Christian democracy, constitutional structure, and the welfare state. *American Journal of Sociology*, *99*(3), 711–749.

Huber, E., & Stephens, J. D. (2001). *Development and crisis of the welfare state: Parties and policies in global markets*. Chicago: The University of Chicago Press.

Iversen, T. (1999). *Contested economic institutions: The politics of macroeconomics and wage bargaining in advanced democracies*. New York: Cambridge University Press.

Iversen, T., Pontusson, J., & Soskice, D. W. (1999). *Unions, employers, and central banks: Macroeconomic coordination and institutional change in social market economies*. New York: Cambridge University Press.

Janoski, T., & Hicks, A. M. (1994). *The comparative political economy of the welfare state*. Cambridge: Cambridge University Press.

Kangas, O. (1994). The politics of social security: On regressions, qualitative comparisons and cluster analysis. In: T. Janoski & A. M. Hicks (Eds), *The comparative political economy of the welfare state* (pp. 346–364). Cambridge: Cambridge University Press.

Katzenstein, P. (1985). *Small states in world markets: Industrial policy in Europe*. Ithaca, NY: Cornell University Press.

Kenworthy, L. (2001). Wage-setting measures – a survey and assessment. *World Politics*, *54*(1), 57–98.

King, G. (2001). Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization*, *55*(2), 497–507 U7.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, *44*(2), 347–361.

King, G., & Zeng, L. (2002). *When can history be our guide? The pitfalls of counterfactual inference*. Unpublished paper, Harvard University Department of Government, May 17, http://gking.harvard.edu/files/counterf.pdf

Kittel, B. (1999). Sense and sensitivity in pooled analysis of political data. *European Journal of Political Research*, *35*(2), 225–253.

Kittel, B., & Winner, H. (2005). How reliable is pooled analysis in political economy? The globalization-welfare state nexus revisited. *European Journal of Political Research*, *44*(1), 1–25.

Korpi, W. (1983). *The democratic class struggle*. London: Routledge and Kegan Paul.

Korpi, W. (1991). Political and economic explanations for unemployment: A cross-national and long-term analysis. *British Journal of Political Science*, *21*(3), 315–348.

Korpi, W., & Shalev, M. (1980). Strikes, power and politics in the western nations, 1900–1976. *Political Power and Social Theory*, *1*, 301–334.

Lange, P., & Garrett, G. (1985). The politics of growth: Strategic interaction and economic performance in the advanced industrial democracies, 1974–1980. *Journal of Politics*, *47*, 792–827.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, *73*(1), 31–44.

Leertouwer, E. (2002). *Measurement issues in political economy*. Groningen: University of Groningen SOM, http://www.ub.rug.nl/eldoc/dis/eco/e.c.leertouwer

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review*, *82*(4), 942–963.

Lieberson, S. (1985). *Making it count*. Berkeley, CA: University of California Press.

Lieberson, S. (1991). Small N's and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces*, *70*(2), 307–320.

Lieberson, S. (1994). More on the uneasy case for using Mill-type methods in small-N comparative studies. *Social Forces*, *72*(4), 1225–1237.

Lieberson, S., & Lynn, F. B. (2002). Barking up the wrong branch: Scientific alternatives to the current model of sociological science. *Annual Review of Sociology*, *28*, 1–19.

Maddala, G. S. (1998). Recent developments in dynamic econometric modelling: A personal viewpoint. *Political Analysis*, *7*, 59–87.

Marks, G., & Wilson, C. J. (2000). The past in the present: A cleavage theory of party response to European integration. *British Journal of Political Science*, *30*(3), 433–459.

Martin, A. (1973). *The politics of economic policy in the United States: A tentative view from a comparative perspective*. Beverly Hills: Sage (Sage Professional Papers in Comparative Politics, No. 01-040).

Moses, J. W. (2001). A methodological critique of statistical studies of globalization. Paper presented to the Max Planck Institute for the study of societies, Cologne, 7 June.

O'Connell, P. (1994). National variation in the fortunes of labor: A pooled and cross-sectional analysis of the impact of economic crisis in the advanced capitalist nations. In: T. Janoski & A. M. Hicks (Eds), *The comparative political economy of the welfare state* (pp. 218–242). Cambridge: Cambridge University Press.

Oskarsson, S. (2003). Institutional explanations of union strength: An assessment. *Politics & Society*, *31*(4), 609–635.

Paldam, M., & Pedersen, P. J. (1982). The macroeconomic strike model: A study of 17 countries, 1948–1975. *Industrial and Labor Relations Review*, *35*(4), 504–521.

Plumper, T., Troeger, V. E., & Manow, P. (2005). Panel data analysis in comparative politics: Linking method to theory. *European Journal of Political Research*, *44*(2), 327–354.

Podesta, F. (2003). Econometric solutions vs. substantive results: A crucial tradeoff in the time-series–cross-section analysis. Paper presented at the 2nd ECPR conference, Marburg (Germany), 18–21 September, http://econpapers.repec.org/paper/eseiserwp/2003-34.htm

Powell, M., & Barrientos, A. (2004). Welfare regimes and the welfare mix. *European Journal of Political Research*, *43*(1), 83–105.

Przeworksi, A., & Tuene, H. (1970). *The logic of comparative social inquiry*. New York: Wiley.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley, CA: University of California Press.

Ragin, C. C. (1994a). A qualitative comparative analysis of pension systems. In: T. Janoski & A. M. Hicks (Eds), *The comparative political economy of the welfare state* (pp. 320–345). Cambridge: Cambridge University Press.

Ragin, C. C. (1994b). *Constructing social research: The unity and diversity of method*. Thousand Oaks, CA: Pine Forge Press.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Rogowski, R. (1995). The role of theory and anomaly in social-scientific inference. *American Political Science Review*, *89*(2), 467–470.

Rothstein, B. (1990). Labour market institutions and working class strength. In: S. Steinmo, K. Thelen & F. Longstreth (Eds), *Structuring politics: Historical institutionalism in comparative analysis* (pp. 33–56). Cambridge: Cambridge University Press.

Rueschemeyer, D., Huber-Stephens, E., & Stephens, J. D. (1992). *Capitalist development and democracy*. Cambridge: Polity Press.

Rueschemeyer, D., & Stephens, J. D. (1997). Comparing historical sequences – a powerful tool for causal analysis. In: L. Mjøset & F. Engelstadt (Eds), *Methodological issues in comparative social science* (pp. 33–56). Greenwich, CT: JAI Press.

Sayrs, L. (1989). *Pooled time series analysis*. London: Sage (Quantitative Applications in the Social Sciences No. 70).

Scruggs, L. (2002). The Ghent system and union membership in Europe, 1970–1996. *Political Research Quarterly*, *55*(2), 275–297.

Scruggs, L., & Allen, J. (2006). Welfare-state decommodification in 18 OECD countries: A replication and revision. *Journal of European Social Policy*, *16*(1), 55–72.

Shalev, M. (1979a). *The strike: Theoretical and empirical studies of industrial conflict across time and nations*. Ph.D. Thesis, Industrial Relations Research Institute, University of Wisconsin, Madison, June.

Shalev, M. (1979b). Strikers and the state: A comment. *British Journal of Political Science*, *8*(4), 479–492.

Shalev, M. (1983). The social democratic model and beyond: Two 'generations' of comparative research on the welfare state. *Comparative Social Research*, *6*, 315–351.

Shalev, M. (1990). Class conflict, corporatism and comparison: The Japanese enigma. In: S. N. Eisenstadt & E. Ben-Ari (Eds), *Japanese models of conflict resolution* (pp. 60–93). London: Kegan Paul International.

Shalev, M. (Ed.) (1996). *The privatization of social policy? Occupational welfare and the welfare state in America, Scandinavia and Japan*. London: Macmillan.

Snyder, D. (1975). Institutional setting and industrial conflict: Comparative analyses of France, Italy and the United States. *American Sociological Review*, *40*(3), 259–278.

Somers, R. H. (1971). Applications of an expanded survey research model to comparative institutional studies. In: I. Vallier (Ed.), *Comparative methods in sociology: Essays on trends and applications*. Berkeley, CA: University of California Press.

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

Stimson, J. A. (1985). Regression in space and time: A statistical essay. *American Journal of Political Science*, *29*, 915–947.

Swank, D. (2002). *Global capital, political institutions, and policy change in developed welfare states*. Camgridge: Cambridge University Press.

Swenson, P. (2002). *Capitalists against markets: The making of labor markets and welfare states in the United States and Sweden*. Oxford: Oxford University Press.

Titmuss, R. M. (1974). *Social policy*. London: George Allen and Unwin (edited posthumously by Brian Abel-Smith and Kay Titmuss).

Tufte, E. R. (1978). *Political control of the economy*. Princeton: Princeton University Press.

Van Kersbergen, K. (1995). *Social capitalism: A study of Christian democracy and the welfare state*. London: Routledge.

Verba, S. (1967). Some dilemmas in comparative research. *World Politics*, *20*, 111–127.

Visser, J. (1992). The strength of union movements in advanced capitalist democracies: Social and organizational variations. In: M. Regini (Ed.), *Labour movements towards the year 2000*. London: Sage.

Weir, M., & Skocpol, T. (1985). State structures and the possibilities for 'Keynesian' responses to the Great Depression in Sweden, Britain, and the United States. In: P. B. Evans, D. Rueschemeyer & T. Skocpol (Eds), *Bringing the state back in*. Cambridge: Cambridge University Press.

Western, B. (1996). *Causal heterogeneity in comparative research: A Bayesian hierarchical modeling approach*. Unpublished paper, Department of Sociology, Princeton University, December.

Western, B. (1998). Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach. *American Journal of Political Science*, *42*(4), 1233–1259.

Wildeboer Schut, J. M., Vrooman, J. C., & deBeer, P. (2001). *On worlds of welfare: Institutions and their effects in eleven welfare states*. The Hague: Social and Cultural Planning Office.

Wilensky, H. L. (1981). Leftism, Catholicism, and democratic corporatism: The role of political parties in recent welfare state development. In: P. Flora & A. J. Heidenheimer (Eds), *The development of welfare states in Europe and America* (pp. 345–382). New Brunswick, NJ: Transaction.

Zorn, C. (2001). Estimating between- and within-cluster covariate effects, with an application to models of international disputes. *International Interactions*, *27*(4), 433–445.

WHAT'S MULTIPLE REGRESSION GOT TO DO WITH IT?

Lyle Scruggs

Unfortunately, many people like to do their statistical work as they say their prayers – merely substitute in a formula found in a highly respected book. (Hotelling et al., 1948 [cited in Kennedy, 2002*])*

I want to begin by thanking Michael Shalev and the editors for providing a forum for discussing the role of quantitative techniques in comparative social science. My particular interest in this debate comes from two angles. First, though I am not trained as a methodologist, I regularly teach statistics to graduate students. This gives me a certain affinity with the frustrations expressed in his paper concerning the use and abuse of regression analysis. It is hard trying to explain to graduate students that while statistical software is a useful hammer, everything is not, in fact, a nail. It is even more frustrating that they can, with some justification, interject to my proscriptions: "but don't a lot of published papers in our field do that." Second, I have written several papers dealing directly with the examples discussed in his paper, most of them employing multiple regression (MR) techniques.

Let me start by laying out what I basically agree with in the paper. First, regarding multivariate regression:

1. The standard estimators in MR, contain many assumptions that are often not verified by researchers and are often conveniently ignored in the actual research process. This is perhaps most true in the sense that statistical estimates are often (incorrectly) reported on a "sample" that is

Capitalisms Compared
Comparative Social Research, Volume 24, 309–323

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24007-9

actually a population, with no intent of generalizing beyond that population. It is also true that many linear regression applications are based on data that do not come close to spanning the tested model's parameter space, or, in other cases, are inadequately specified to justify pooling. The operationalization and estimation of MR models should be the last (or at least a late) stage in the development and testing of hypotheses. The earlier stages involve developing a theoretical model, mapping the theoretical model to observables, developing a statistical model of the process generating the observables.

2. There has been a tendency for (easily implemented) techniques to run ahead of and amok in empirical analysis. This is due partly to methodological developments, but is due mostly to cheap computational power and multipurpose software packages. These factors make it very easy to access estimation procedures that, unlike OLS, are arcane knowledge to most practitioners.
3. Because of the first two points, statistical analysis is used atrociously in a lot, if not most, comparative social science. How this state of affairs can come to be in a discipline with peer review is an interesting social science question in itself.

On several points of "causation" I am also in agreement with the paper:

1. Correlation cannot determine causation, nor can it eliminate the possibility that an unspecified alternative explanation explains a particular outcome. That correlation seems to hold either status among practitioners is no doubt a failing of training in statistical methodology, if not in social sciences more broadly. MR can only really confirm that empirical results are consistent with a hypothesis.
2. The effects of particular "forces" are generally contingent on the presence/absence/level of others.
3. Related to 2, there is seldom a single, isolated cause for a phenomenon of interest.

With these things said, I am concerned that this article tends to throw the baby out with the proverbial bathwater, leaving us with nothing in the tub. Regarding causation, my points of agreement are completely consistent with taking a diametrically opposite position to Shalev's with respect to the appropriate methodology for confirming causation.

First of all, almost all of the substantive problems with applying MR that are discussed in the paper are also addressed in basic econometrics texts.[1] Moreover, I think the paper just has it flat wrong about basic aspects of

what regression analysis can do. Most of the mistakes that the paper correctly identifies about how MR is too often conducted are seldom solved, and sometimes made worse, by appealing "case analysis" or "other qualitative techniques" (see Seawright, 2004).

Second, leaving aside critiques of pooling for the moment, a number of the articles singled out for criticism in Shalev's paper: (a) have been criticized on the empirics, (b) make as much of a case for *textbook* MR than any alternative approach, particularly one whose details are not clearly specified or are anyway part of the basic MR toolbox. Finally, one has to ask what is the alternative approach to MR for *evaluating theories?* As I hope to make clear below, the paper's most extended discussion of "alternatives to multiple regression" does not provide anything approaching a basis for establishing a causal relationship between the theoretical variables.

MULTIPLE REGRESSION AND CAUSAL EXPLANATION

The first thing most people learn in statistics is that correlation is not causation, and that inferring causation from statistical results requires that there is a theoretical model (a good reason to think there is a causal effect), not just a statistical one. Usually, this implies a theory with some "mechanisms" that may also be subject to investigation. Except in some quite limited senses of the term, almost no one thinks that any MR results justify a causal claim (Goldthorpe, 2001).

Nonetheless, I can attest from my teaching experience and reviewing manuscripts for scholarly journals that it is common for users of statistics to forget all of this. Why this is so is an interesting question. I have some guesses – e.g., researchers operate in a community that may not know enough about statistics to speak out about inappropriate use; they succumb to the temptation to ignore poor statistical methods when they produce results that seem to support their pet "causes." But these are only guesses.

What I found unclear in the paper is a definition of a cause, and the criteria for stating and establishing one. How attaching names to cases, for example, does anything to resolve the issue of establishing causation is a mystery to me. In later sections of the paper, it seems that this is a means by which one can introduce explanations in an *ad* (*post*?) *hoc* manner, with no recognition that this can easily result in a *unique* configuration of causes for each case. I do not think that is Shalev's intent. I know of no theories that are stated in terms of particular observations or cases.

Shalev rightly criticizes theoretical approaches that start with a dependent variable, add "independent variables" until most variation in the sample is explained, and then claim to have a model of causes. But one would be hard-pressed to find a modern econometrics text that does not reject such an approach.

At various places, Shalev raises the prospect that causal relationships can vary across units and across time in an effort to critique MR approaches. He seems to ignore the fact that unit homogeneity is necessary for *any* verifiable causal explanation in science. One can always claim that an explanation might not *always* hold in *all* places, just as one cannot refute the claim that a cause only "seems" to apply in times and places other than the observed case.

WHAT *CAN* REGRESSION DO?

Consider the following two quotes that are drawn from an early section of the paper. I select them, because I think that they are widely repeated claims against MR in contrast to a case-study method.

"[Case oriented research] assumes from the outset that the effect of any one cause depends on the broader constellation of forces in which it is embedded" (p. 5)

"MR is even more challenged by another causal assumption that flourishes in case-oriented analysis, namely that there may be more than one constellation of causes capable of producing the phenomenon of interest." (p. 5)

These objections are metaphysical ones, in the sense that they really undermine any attempt at explanation or verification in the sciences. The first statement amounts to saying that a case-oriented research *assumes* that any cause cannot be separated from a broader constellation of causes, and implicitly asserts that variable-oriented research assumes that it can be. I find the first assumption inscrutable as a basis for *comparative social science*. If causal forces cannot be isolated from one another and identified across units of comparison, how does one move beyond explaining all differences among cases as due to irreducible differences in the cases themselves.

This causal perspective would seem to imply, for example, that differences in welfare spending are ultimately explained by different "national characters" (understood broadly to include culture, history, and institutions), not by leftist governments, strong unions, or the level of economic development, or some combination of *just* those three factors. If each cause is considered to be embedded in other "forces," we (even the historians among us) should

be required to specify what we think those forces are and how they affect "causes" we are interested in explaining. And these explanations should be subject to some criteria of rejection, which implies a domain beyond a single event.

The second statement amounts to a claim that from the infinite *set* of factors that comprise a "constellation of forces" needed to explain an event, more than one such *set* of conditions may cause the event. This makes *any* causal explanation largely irrefutable. Why does the United States have no socialist party? If my explanation is "because it was a former British colony," identifying some British colonies with socialist parties is not sufficient to refute the causal claim definitively, because those other former colonies are not the United States. Indeed, if we did find a condition (X) that was, empirically, unique to those countries without strong socialist parties, one could still not refute the causal claim that, for the United States, condition X was only operative because the United States was among other things, a former British colony. (The counterfactual would be that condition X would not have precluded the development of a socialist party in the United States, if it had been, say, a French colony.) If nature behaved this way, MR would certainly be humbled, but no less thoroughly than any alternative approach to evaluating causal *regularities*.

CAN REGRESSION DEAL WITH CONJUNCTURAL CAUSATION AND CAUSAL HETEROGENEITY?

The previous section suggested that *any* approach to explanation must specify what is supposed to matter and how it matters. Here, I want to object to a narrower claim that MR cannot really accommodate conjunctural causation and causal heterogeneity. MR *does* require that whatever causal possibilities we posit to exist in theory must be specified and operationalized in an empirical model *beforehand*. But doing that is perfectly compatible with the reality of conjunctural causality and causal heterogeneity.

Conjunctural causation can essentially be accounted for by some type of "interaction term" in a regression model. This would test whether the effect of two things together is greater (or less) than the sum of the parts. In a simple case, one can simply take the interaction as the intersection of two variables. If, for example, having A or B alone is jointly bad for you, but having A and B (together) is good for you, this can be incorporated into an MR model. Interaction terms, particularly dummy-variable interaction

terms, which allow for the effect of a variable to be different to two contexts, i.e., government spending produces inflation in non-corporatist systems, but not in corporatist ones, are standard fare in regression texts. (A related, but more complicated, causal structure amenable to MR is hierarchical models, which Shalev praises later in his paper.)

The fact that a particular regression model fails to include (or consider) interaction possibilities is a theoretical or a model specification problem, not a technical one. While it is convenient to blame this lack of creativity on making students take statistics courses – Shalev cites Abbott's claim that using linear models causes us to think that causal effects are linear (p. 5) – the widespread confusion about conjunctural causation is really an argument why students desperately need more *good* statistical training, not less statistical training.

Shalev suggests that the problem with an interaction specification in MR is that it takes up degrees of freedom. This is a pretty widespread claim about the advantages of case-oriented approach. But how a case approach, which, if anything, leads to a *reduction* in the number of cases analyzed, can more adequately discern the validity of an explanation with one more "moving part" is hard to understand.

It is more often in "substantive" (i.e., case) approaches that one finds much vaguer specifications about the relationships between variables. I can only draw one straight-line curve between points; but I can draw a lot of non-straight curves. So what does it mean to move from a claim that a relationship is "linear" to a claim that it is "non-linear"?[2]

To illustrate how MR deals with conjunctural causation, consider the following example of ten observations (Table 1).

A standard regression model $Y = \mathbf{b0} + \mathbf{b1}A + \mathbf{b2}B$ yields

$$y = -0.9 + 0.5A + 0.5B$$

$$\text{se} \quad (0.57) \ (0.57)$$

and the overall model explains no variance Y (R^2 is around 0).

Given a theoretical reason (or just a hunch) of conjunctural causation between A and B, we posit a different regression model

Table 1. Data Example 1.

A =	0	0	0	1	1	1	1	1	0	0
B =	1	1	1	0	0	0	1	1	0	0
Y =	−1	−1	−1	−1	−1	−1	1	1	0	0

$Y = \mathbf{b0} + \mathbf{b1}A + \mathbf{b2}B + \mathbf{b3}C$. C is a new variable (A*B) estimating that model with the same sample of data, $Y = 0 + (-1)A + (-1)B + (3)C$, and predicts the data perfectly. Individually, A and B have a negative effect, but jointly, their total effect is positive.

Note that if our hunch arose simply from eyeballing the data (which you can do in this case) and not for some a priori theoretical reason, then one has simply summarized the data. The question of whether that model is good cannot come from mechanically fitting the data.

Causal Heterogeneity

In contrast to the common assertion that MR cannot handle causal heterogeneity, the possibility that different combinations of variable values can produce the same outcome is precisely what MR allows for. Indeed, when I was a graduate student, I learned that one reason for using MR as opposed to simpler, bi-variate regression analysis was that variation in most of the variables that social scientists are interested in is unlikely to have a single cause. Though some of my students do have trouble seeing it at first, a regression estimate produces a predicted value for each case, and it generates predicted values for all possible combination of variables in the model, even if some combinations are not represented by specific cases.[3]

To see this, Table 2 presents another simple set of seven observations.

A and B are both associated with Y. Regressing A and B on Y in the form $Y = \mathbf{b1}A + \mathbf{b2}B$ produces a result $(0.2 + 0.6A + 0.6B)$ that seems odd at first, because it implies that A and B do not perfectly predict Y. (If $A = 1$ and $B = 0$, Y(predicted) = 0.8.) However, knowing that Y only takes a 0 or 1 value, the regular ordinary least squares (OLS) regression model is flawed. You need a logit estimator, which is a relatively minor variation on the OLS technique, and is least introduced in most basic econometrics texts. Estimating these data with a logit model produces a result that perfectly *classifies* all of the cases![4] As for the claim that MR does not distinguish between additive, conditional, or multiple pathways as the causal forces, they are easily obtained from the predicted values of the actual cases.

Table 2. Data Example 2.

A =	1	1	0	0	1	0	0
B =	0	0	1	1	1	0	0
Y =	1	1	1	1	1	0	0

One thing that is sometimes overlooked is that MR approaches (OLS or variants like logit, ordered logit, etc.) can estimate parameters when variables are measured dichotomously (0 or 1) up to continuous measurement. MR approaches are also generally robust to reductions in the number of categories of measurement. Major alternative approaches, like Qualitative Comparative Analysis are only intuitive when the data are dichotomous for all of the variables. Too much may be made of estimation on a "continuum" when the measured concepts are not really so refined. That may be a temptation that MR permits, but it is not a cause of poor measurement.

Spanning Large Parameter Spaces

Shalev is certainly correct when he critiques how many MR studies "span" many empty cells and convey an impression of linear effects that is not really justified. For the relationship displayed in Fig. 1, OLS reports a "statistically significant" regression line, and would predict $Y = 12$ given $X = 13$. That prediction is based on the *assumption* that the relationship is linear, and the data obviously fails to support that assumption. ("More supportive data" for a linear effect would be that the observations $(X,Y) = (9,4)$ and $(15,21)$ were actually, say, $(9,8)$ and $(15,15)$.)

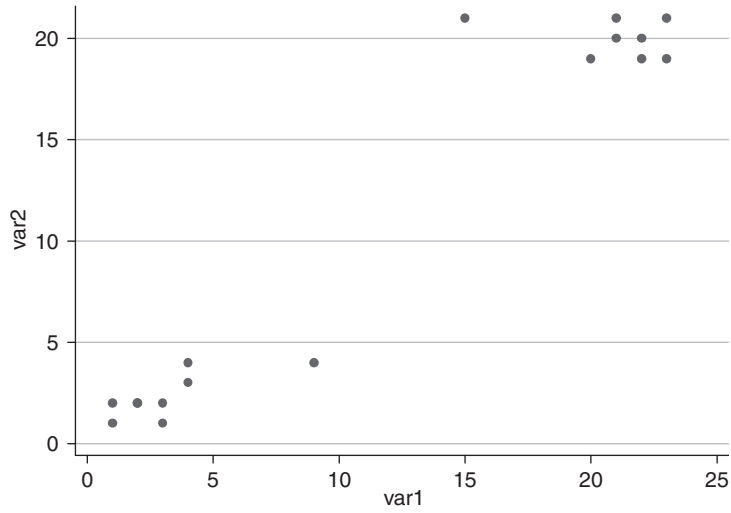

Fig. 1. Spanning Parameter Spaces.

But is this a problem that is particularly likely to plague MR as a technique? One step in developing and evaluating regression models is to examine assumptions about functional form, error distributions, and other so-called residual diagnostics before accepting MR results as genuine. Most basic econometrics texts have sections on residual diagnostics and functional form, and walk through all the basic assumptions of linear models. Econometricians like Leamer and Kennedy, and many econometrics texts, provide fundamentals on how to test the robustness of regression estimates, often in ways that reveal problems that Shalev's paper identifies. It is thus hard to characterize most of these "spanning" problems as unique to MR, let alone a justification for not using MR.

The example in Fig. 1 is a convenient illustration, because the data do follow a binary pattern. There is very obvious break. When the data actually varies more continuously over the range of values, imposing a binary classification on data can make results very sensitive to where one assigns the cut-point. Binary classifications are only simple if the cases are really discreet without many cases "somewhere in-between."

With respect to the capability of regression analysis to identify "problems" in the data, consider the contribution of Lange and Garrett, which Shalev mentions in several places in his paper. Lange and Garrett's initial findings (based on a simple model with an interaction term) were immediately contested Robert Jackman (1987) based on an assessment of the predicted values and errors. Hicks (1988) and Scruggs (2001) also present refined analyses.[5]

What is the solution to avoid spanning large parameter spaces? Why are those middle cells empty? Shalev's suggestion seems to be that cells are empty because there is, in fact, not independence in regressors. Such clustering can show up in MR analyses as correlated independent variables (collinearity). This problem makes it hard for MR to isolate with confidence the effect of any one independent variable, while still *allowing inferences to be made about "joint effects" of several variables*. In other words, if three factors coexist with the dependent variable in most of our sample of cases, MR analysis will show that the *set* of factors is associated with the dependent variable, and that primacy of these factors cannot be disentangled empirically.

More problematic is when there are situations like in Fig. 1. Simple regression can produce misleading results, but proper econometric analysis alerts us to these problems in two ways. First, scatterplots like Fig. 1 will raise a red flag. Second, the residuals from a regression analysis of these data are not normally distributed. Both checks would tell us to be wary of the MR estimates of a simply linear relationship.

Finally, Shalev does not make a real "positive" argument for a case-centered approach being any better in diagnosing or dealing with this kind of problem. Given clustering like Fig. 1, what is the causal explanation? Informed MR diagnoses the problem. But the approach advocated in the paper, like much case-study work, assumes determinism and perfect measurement of concepts. It infers that any residual is thus due to model misspecification. What if the residual is due to a measurement problem? Or due to sampling variation? Or to simple indeterminacy? It would seem that the paper's approach would always result in overfitting the data.

Population

Shalev points out that the data that comparative welfare state researchers use is problematic as a basis of evaluating their empirical models. MR *estimates* are not useful if you know the population values. Assuming determinacy, this might seem to imply that there should be no residuals, and that overfitting is not really possible. But social scientists generally want to be able to use their explanations to *predict.* Some notion of prediction (or counterfactual condition) is implicit in most definitions of causality.[6] This means that the "population" of 18 OECD countries is really a "sample" of outcomes, which we use to create explanatory models that will inform future policy choices or which are consistent with a well-developed theory. While prediction is not necessary for an explanation to be correct, there are many conceivable explanations of a given phenomenon. This makes any explanation's "predictive" power a good basis for parsing among competing explanations.

Shalev's discussion about this problem is unclear to me. On page 11, he cites Freedman and Leamer to the effect that hypothesis testing requires well-developed theory and data that has not been used to create the model in the first place. He then cites Ragin's claim that data and theory are in a constant dialog, and infers that Freedman and Leamer plus Ragin implies that we can, in fact, only count on MR as a way to summarize data, not to test hypotheses.

I think this totally misconstrues Ragin and Leamer and Freedman. MR can never simply be used to "summarize" relationships. The "product" of a model is valid to the extent that it can explain data that is independently derived. Leamer and Freedman (and, once again, many basic econometrics texts) do not promote "purist" notions of separating theory and data. What they suggest (also see de Marchi, 2006; Granger, 1999; Kennedy, 2002) is

that researchers who want to use particular data to assist in constructing an explanatory model should not use the fit of the model to those same data as a test of the model. Instead, researchers should test the model on new data. In practice, this calls for a strategy of (a) dividing your dataset into "model building" and "model testing" subsets, or (b) following an approach that looks for other observable implications of a model and testing those other "observable implications" of the theory (King, Keohane, & Verba, 1995).

COMMENTS ON SPECIFIC EXAMPLES

In the rest of this commentary, I briefly discuss specific examples mentioned at length in Shalev's paper. Where he invokes the importance of paying attention to specific proper names in resolving issues of causation, I submit that this is not a fruitful line of attack. First, a careful analysis of residuals from a conventional MR procedure is adequate to uncover many of the problematic results he points out. Second, there is no attempt in the paper to evaluate competing explanations that are both consistent with the data at hand and validated by new data (such as data for the same countries later in time or for a different set of countries). In principle, I would have no objection to the idea that knowing which specific cases are "outliers" can help to generate new and meaningful explanations. However, if one accepts that there is not complete determinism, or that there is measurement error between a concept and what is observable, outliers can represent the "white noise" inherent in any variation that we want to explain.

The Ghent System, MR and Causation

I am unsure why Shalev suggests that the Ghent unionization relationship is undertheorized. The papers cited in this section (and others) have evaluated a variety of causal mechanisms, and, while all make strong claims for an independent effect of a Ghent system on the rate of unionization, none claims that the Ghent system explains all variation in union density. The essence of his claim seems to be that the only way to prove a Ghent causal effect is to demonstrate a case in which all other causes are "turned off."

This seems to be a wild goose chase. If one actually could demonstrate it for the three explanations offered in Shalev's paper, one could bring other explanations into the fray. Shalev mentions the importance of a union's access to the workplace, for example, to account for differences in union

density in Belgium (a quasi-Ghent country) and the Netherlands (a non-Ghent country). He fails to point out, however, that the workplace access condition fails to explain the difference in union density in Norway and Sweden. (Of course, if you invoked "causal heterogeneity," this inconsistency is not a problem: each case could be allowed to have its own unique cause of its particular high or low density. But, comparative analysis would then seem irrelevant.

If Canada or Germany were to adopt a Ghent system, would they experience an increase in unionization? Shalev's argument suggests not, because he maintains that density is caused by a *combination* of three factors – small size, left governments, and Ghent – and Canada and Germany lack all three. The prediction from the MR model predict that adding Ghent institutions alone would raise density in Canada and Germany 20–30 points. I would wager that density would go up a lot if Canada switched to a Ghent system, but maintained its large size and less "leftist" governments.

Hall and Franzese

In the section on the Ghent system, Shalev collapsed continuous variables in a dichotomy to suggest a conjunctural type of causation. Here, he criticizes Hall and Franzese for doing just that, and suggests that, if they had used a more fine-grained measurement in their key explanatory variables, their results disappear. But perhaps the most important issues discussed in this section are pooling and the appeal to specific names in comparative analysis.

The first quibble that I have with this section is that there is no need to appeal to specific cases to demonstrate the results Shalev shows in his Chart 2. If one removes the names of the countries, Hall and Franzese's results still fall apart.

With respect to pooling, Shalev points out a common (and often untenable) assumption in such analyses that causal processes are the same in different time periods. This objection, of course, is not necessarily a temporal one – causal relationships may vary across time, across groups of countries, across configurations of space time, and across combinations of any variables. It is, in fact, impossible to count the number of alternative scenarios of variable causality. But this goes back to my initial discussion of causation. We might as well throw up our hands.

I think the more valid objections to pooling are based on the fact that pooling is quite often done by researchers interested simply in explaining cross-national variations, and that it is done with little realization that

adding observations does not necessarily add information that is (as assumed) independent of cross-sectional variation.

Shalev identifies several specific countries as "driving" Hall and Franzese's results. Knowing these cases, he then adds a new variable into the explanation. This approach seems like an exercise in "ad hoc'ism". Is the new explanation actually measured for all other cases and the model's results re-evaluated? Is this not just the type of "add variables until the R^2 approaches 1" approach to MR that is almost universally condemned? Finally, contrary to the admonitions of MR econometricians like Freedman and Leamer, the paper does not attempt to evaluate the implications of these added variables in a "more complete" model of political outcomes using new data (or observable implications of the model). Like Ptolemy's epicycles, this approach adds something that accounts for an empirical discrepancy, but that does not make it the most appealing causal explanation.

Esping-Andersen

Finally, I turn to Shalev's factor analysis of Esping-Andersen's data regarding welfare state regimes. Having recently spent several years trying to independently replicate the data that form the basis of Esping-Andersen's path-breaking book, I am not convinced that this example is very informative.[7] Leaving aside the issue of the validity of the reported measurements, what does the clustering Shalev provides actually represent? By excluding the decommodification index from his analysis, Shalev excluded perhaps half of the empirical basis of the "three worlds" typology.[8] Almost half of the indicators of "worlds of welfare" (6 of 13) that are included in the factor analysis deal with the structure of the old age pension system, obviously an important element of the welfare state, but hardly one that leaps out from the "worlds of welfare" narrative.

Second, it is taken as axiomatic that the underlying measures (and the way that they coalesce) are valid measures of some underlying concepts that differentiate the three regimes, yet the concept validity for most of these indicators is not really discussed in the original source or here. (Admittedly, this is more of an argument about the appropriateness of the example than anything Shalev does in the paper.)

I will close with one example of what I mean here. It is motivated by specific knowledge of programs and of cases, yet it reveals why it is so often *measurement*, not just theory or methodology, that matters. Is it plausible that civil service pension spending is not correlated with (i.e., loads on the

same factor as) public employment? It would seem that countries with lots of public employees must have big civil service pension bills. Chart 5 in Shalev's paper clearly shows this is not true for Esping-Andersen's data. Are civil servants just especially lavishly treated under "statist" and not lavishly treated in "social democratic" regimes? No. The terms of civil servant pensions are quite generous in every social democratic regime that I am aware of.

There are two reasons for low civil servant pension spending *as measured here*, both quite unrelated to the "Three Regimes" story. First, the expansion of the public sector is comparatively recent in social democracies, so the full fiscal impact of the state pension commitment (but not the commitment itself) is not manifest in civil servant spending in 1980, when the Three Worlds data is collected. Second, the civil servant pension system is a separate occupational pillar in statist countries, e.g., Germany and France, while civil servant pension spending is merely a (generous) top-up to the universal benefits offered in social democracies. In other words, the differences in civil servants pension spending in Germany and Sweden have little to do with what civil servants expect to get in the two countries and mostly to do with mundane accounting.

NOTES

1. I admit that I have not undertaken an exhaustive survey of these texts. Most of what I say here refers to several texts in economics that I have used in the past and which are quite popular. (They have been published in numerous editions.) They are Gujarati (2003), *Basic Econometrics 4*th *edition*, Kennedy (2003), *A Guide to Econometrics 5*th *edition*, and *Undergraduate Econometrics 2*nd *edition* by Hill, Judge, and Griffiths (2001).

2. To illustrate the problem, with $n = 2$, variation in X perfectly explains the variation in Y (i.e., $R^2 = 1.00$). With $n = 3$, variation in X perfectly explains the variation in Y if my model is $Y = b1X + b2X^2$; with $n = 4$, $R^2 = 1$ with $Y = b1X + b2X^2 + b3X^3$; and so on. All except the first are "non-linear" relationships between X and Y. Thus, for all X–Y relationships, there is *some* non-linear specification of X that perfectly explains the co-variation between X and Y.

3. Of course, some of these hypothetical values – outside of the range of observed X's – are often absurd. But most econometrics texts tell you so and suggest problems with predictions for cases outside of the range of observed Xs.

4. It is also possible to use OLS to test this relationship (again ignoring the identity of cases) if one hypothesizes that A or B may explain Y, and that A and B could occur jointly. Simply specify $Y = \mathbf{b1}A + \mathbf{b2}B + \mathbf{b3}A*B$, with the expectation that $b1 = b2 = 1$ and $b3 = -1$. This implies that when A or B = 1, Y is hypothesized to be $1(1) + 1(0) + (-1)(1*0) = 1(0) + 1(1) + (-1)(0*1) = 1$, and when A and B are both 1, Y also is expected to be 1: $1(1) + 1(1) + (-)1(1*1) = 1$.

5. Furthermore, the pooled model in Alvarez et al. was shown to be pretty limited once flaws in the estimating methodology were corrected. Only one of the three significant findings was exonerated.

6. It is sometimes said that one might also want some inherent variance to "leave room" for human agency. The problem for this in social science is that the effect of human agency is something that we very much want to explain.

7. See Scruggs and Allan (2006a, 2006b). Simply put, in replicating the decommodification and stratification indices we found some major inconsistencies in the scoring provided in *Three Worlds*, all of which seem to "interpret" the data in a manner supporting the theoretical structure. For example, scoring "rubrics" appeared to be ignored whenever they would produce results seemed not to "fit" the "three regime" framework. As for close knowledge of the cases, this seemed quite skewed toward familiarity of the Nordic countries, and usually to the detriment of, for example, non-European cases. On this score, due to suspicions about the underlying data (not an error on Shalev's part), the results in this section of his paper may not particularly reliable.

8. How the antipodes are treated in the decommodification index could have been easily corrected to make it consistent with Castles and Mitchell's objections.

REFERENCES

de Marchi, S. (2006). *Computational and mathematical modeling in the social sciences*. New York: Cambridge University Press.

Granger, C. (1999). *Empirical modeling in economics: Specification and evaluation*. New York: Cambridge University Press.

Goldthorpe, J. (2001). Causation, statistics, and sociology. *European Sociological Review*, *17*, 1–20.

Gujarati, D. (2003). *Basic econometrics* (4th ed.). New York: McGraw Hill.

Hicks, A. (1988). Social democratic corporatism and economic growth. *Journal of Politics*, *50*, 677–704.

Jackman, R. (1987). The politics of economic growth in advanced democracies, 1974–1980. Leftist strength or North Sea oil? *Journal of Politics*, *49*, 202–212.

Hill, R. C., Judge, G. G., & Griffiths, W. E. (2001). *Undergraduate econometrics* (2nd ed.). New York: Wiley.

Kennedy, P. (2002). Sinning in the basement: What are the rules? The ten commandments of applied econometrics. *Journal of Economic Surveys*, *16*, 569–590.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge, USA: MIT.

King, G., Keohane, R., & Verba, S. (1995). *Designing social inquiry*. Princeton, NJ: Princeton University Press.

Scruggs, L. (2001). The politics of growth revisited. *Journal of Politics*, *63*, 120–140.

Scruggs, L., & Allan, J. (2006a). Welfare-state decommodification in 18 OECD countries: A replication and revision. *Journal of European Social Policy*, *16*, 55–72.

Scruggs, L., & Allan, J. (2006b). Welfare state decommodification and poverty in advanced industrial democracies. *Comparative Political Studies*, *30*, 880–904.

Seawright, J. (2004). Qualitative comparative analysis vis-à-vis regression. Paper presented at the 2004 Annual Meeting of the American Political Science Association.

METHODS IN COMPARATIVE POLITICAL ECONOMY

Jonas Pontusson

Michael Shalev's provocative essay deserves the attention of quantitatively oriented students of comparative political economy. I agree with many of the points that Shalev makes, but I disagree with the basic thrust of his discussion. I wish that Shalev had written a paper that identified the pitfalls of uncritically applying standard regression techniques in comparative political economy and then explored various ways that these pitfalls have been or might be avoided. The paper that Shalev actually wrote contains elements of the paper that I wish he had written, but these elements seem out of place, for the bottom line of Shalev's paper is that multiple regression should not be employed by comparative political economists (and, presumably, anyone else engaged in comparison of "macro social units"). The implication of Shalev's discussion seems to be that the regression-based comparative political literature of the last 10–15 years has yielded very little, if anything, by way of new empirical and analytical insights. That strike me as far too harsh and sweeping a claim. Backing off that claim, as I imagine that Shalev would want to do, implies some recognition that multiple regression may be useful, at least in certain forms and for certain purposes.

For Shalev, there are two basic reasons why multiple regression (henceforth MR) is an inappropriate methodology for our field of inquiry. The first reason is a pragmatic one: we simply do not have a sufficient number of observations at our disposal to do MR properly. The second reason has to

Capitalisms Compared
Comparative Social Research, Volume 24, 325–333

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24008-0

do with causal complexity: estimating marginal effects based on a linear and additive conception of causality, MR does not provide an appropriate means to test the kinds of theories that animate comparative political economy. Shalev's second reason for rejecting MR is logically prior to the first reason in the sense that the "small-N problem" is moot if MR is fundamentally inappropriate from a theoretical point of view.

My commentary will focus on the fit between theory and methodology, but let me begin with a preliminary point about the "small-N problem." Much of Shalev's discussion seems to presume that comparative political economists are more or less exclusively concerned with between-country differences in institutions, policies and outcomes. If this is correct, we do indeed have a serious "small-N problem" and pooling cross-section and time-series observations does not constitute a defensible solution to the problem. But is Shalev right in his (often implicit) characterization of what "comparative political economy" is all about? As I will indicate in the course of the following discussion, I think that comparative political economists, regardless of whether they primarily employ quantitative or qualitative methods, should be and have increasingly become interested in explaining change over time and other forms of "within-country variation" as well as "between-country variation."

Turning to the fit between theory and methodology, is it really the case that regression analysis assumes a linear-additive conception of causality? Some of Shalev's general formulations read as if the quantitative comparative political economy literature consisted entirely of empirical models designed to identify the linear effects of each independent variable in serial fashion, but this is surely not the dominant style of analysis in the literature of the last 10–15 years. In fact, testing conditional causal arguments is a key feature of several of the works that Shalev scrutinizes and criticizes. While Hall and Franzese (1998) seek to demonstrate that the effects of central bank independence are conditioned by wage-bargaining coordination, Garrett's (1998) central argument posits that partisan responses to globalization are conditioned by labor encompassment, and Western's (1998) analysis is designed to show that short-term macro-economic effects of changes in government partisanship depend on the institutional constellation regulating labor markets. Shalev may be right that the execution of these analyses is flawed, but the flaws that he identifies do not seem to derive from assuming linear-additive causality.

I agree with Shalev that the kind of variable-specific interaction models pioneered by Garrett and Lange (Alvarez, Geoffrey, & Lange, 1991) do not fully capture the idea of "causal syndromes" and that this idea is indeed

central to much theorizing in comparative political economy, notably the typological approaches of Esping-Andersen (1990) and Hall and Soskice (2001). The typological tradition does not conceive clusters of countries simply as "bands" in the distribution of discrete variables, with relationships between variables being constant across clusters. Rather, this literature suggests that causal effects, not just the values that causal variables take on, should vary across clusters in systematic ways. As Rueda and Pontusson (2000) illustrate, such propositions can be tested within the framework of MR; indeed, I fail to see how else they might be tested (rather than simply being assumed). Pooling cross-section and time-series observations for 16 OECD countries, Rueda and I seek to ascertain whether the determinants of wage inequality are distinctly different in "liberal market economies" and "social market economies" by interacting dummy variables for each of these political-economy types with all the independent variables included in their model. While certain variables (e.g., union density) turn out to have essentially the same effects in both clusters, other variables (e.g., government partisanship) operate differently depending on the broader institutional configuration.

Even more so than variable-specific interaction models, the "syndrome-probing" interaction models estimated by Western (1998) as well as Rueda and Pontusson (2000) presuppose a relatively large number of observations. Shalev's objections to this approach seem to hinge entirely on his objections to pooling cross-section and time-series observations. Shalev points out that pooling entails a number of technical estimation issues, pertaining, in the first instance, to the potential for serial correlation and heteroskedasticity. I have neither the space nor the competence to sort out these issues, on which there is no clear consensus among methodologists.[1] Let me simply say that it clearly behooves practitioners of pooled MR to compare (and report) the results of models with different technical specifications. This has indeed become increasingly common practice in the quantitative comparative political economy literature.

Shalev's main objection to pooling is a substantive one: that pooling ignores the different causal structures underlying cross-sectional and time-series variation. There is a curious dualism at work in Shalev's argumentation on this score. Surely, we cannot be satisfied with observing that one set of variables are associated with cross-national variation on some outcome while an entirely different set of variables are associated with change in the same outcome over time. There are undoubtedly cyclical processes involved in time-series data that do not "add up" to level differences between countries (as Shalev notes, for instance, temporal fluctuations in

strike activity are associated with economic cycles). But this logic does not work the other way: if we believe that government partisanship is a cause of cross-national differences in levels of social spending, then we must also believe that government partisanship affects changes in social spending in ways that do add up. In other words, time-series variation is just as relevant to the proposition that government partisanship matters as cross-sectional variation.

I hasten to add that I think that Shalev is absolutely right in emphasizing that causal dynamics are likely to vary over time – in other words, that we should not assume, as many practitioners of pooled MR do, that the causal effects of any given X variable are constant over the time period covered by our data. Like the question of whether clusters of countries partake in different causal syndromes, the question of causal heterogeneity over time can be fruitfully explored within an MR framework. One obvious way do so is to estimate separate regression coefficients for different periods within our dataset. My favorite version of this approach is moving windows analysis, in which the same regression model is re-estimated for consecutive fixed time periods (dropping the earliest year and adding a more recent year to each new window). The beauty of this simple technique is that it allows us to track changes in causal effects over time and thus avoid the problem of arbitrary periodization.[2]

Invoking the authority of statistician David Freedman, Shalev urges practitioners of MR to concede that "the most which can legitimately be done with MR is […] to summarize multivariate datasets" (p. 11). Though Shalev does not elaborate much, the basic argument to which he alludes here is familiar and widely accepted, at least in theory, by methodologists and MR practitioners. Simply put, the regression coefficients that we obtain by estimating some regression model tell us about the statistical association between two variables, but not really about the causal relationship between them. MR is essentially a more complicated form of correlational analysis. Our estimate of the association between Y and X_1 takes into account the associations between Y and other Xs included in the model (reducing the probability that the association between Y and X_1 is spurious), but does not shed light on the causal mechanisms behind the association between Y and X_1 (or even the direction of causality). Hence we should not think of regression coefficients as doing the explaining; they themselves must be explained. As methodologists constantly remind us, regression results make little sense without a well-specified causal theory. Beyond this, I would also argue that regression analysis should be complemented by in-depth case studies that probe causal mechanisms by analyzing the sequencing of

changes in the variables of interest as well as political processes and the motivations of political actors.

Are the alternative (''low-tech'') methods of quantitative analysis proposed by Shalev any more immune to the objection that ''correlation does not equal causation'' than standard regression methods? I fail to see a compelling argument why this should be so in Shalev's paper. Rather, Shalev's argument seems to be that these methods invite and can be more easily integrated with case studies. By retaining ''named cases'' (i.e., the proper names of countries), they allow us to identify anomalies as well as paired comparisons that deserve further exploration. This advantage must be weighed against an obvious disadvantage illustrated by Shalev's own examples: Shalev's techniques only work well with two or three explanatory variables and, with three explanatory variables, the interaction effects among these variables become quite opaque.

It is undoubtedly true that ''complicated methods often move us away from looking at and thinking about the data'' (Beck & Katz, 1996, p. 31, cited by Shalev, p. 33). The need to explore and present patterns in the data rather than simply reporting regression coefficients has become a common theme among methodologists in recent years and political-economy practitioners of MR will hopefully follow suit. I firmly believe in using tree diagrams, scatterplots and cross-tabulations of data to complement and illustrate the results of regression analyses. Like Shalev, I also believe that such techniques are useful for the purpose of building (or improving on) our theoretical models. Most obviously, identifying anomalous cases might (should) lead us to add new variables to the original model. Case visibility is essential to this step in the theory-building process, but the ultimate goal remains to ''substitute variable names for proper names'' (Przeworski & Teune, 1970).

For certain expository purposes, I would be quite content to stay away from regression analysis altogether (cf. Pontusson, 2005), but I do believe that regression analysis has been and remains a necessary component of advancing the comparative political economy project. If our goal is to ascertain how government partisanship or some other variable of interest affects some outcome such as unemployment rates or growth rates, social spending, unionization, poverty rates or other measures of income distribution – to mention the most obvious concerns of the literature that Shalev reviews – we must surely control for the effects of economic and demographic structures. Even if regression results constitute ''mere summaries'' of patterns in the data, these are arguably better summaries than those yielded by Shalev's alternative techniques in the sense that they provide

more accurate estimates of the "statistical associations" to be explained. The need to control for the effects of several (many) variables that are causally relevant would seem to constitute a compelling reason for sacrificing visible, properly named, cases at some point in the process of data analysis.

Shalev may be right that comparative political economists initially turned to pooling without any substantive interest in over-time variation, viewing pooling simply as a means to analyze cross-section variation with a larger dataset, but this is less obviously true for more recent work. Arguably, the practice of pooling has itself generated an increased interest in modeling dynamics of change among quantitatively oriented comparative political economists. Certainly, real-world developments have brought the challenge of explaining changes in institutions and policies to the fore.

Shalev asks, "Is pooling a panacea?" His negative answer may be a useful antidote to the current pooling fad, but the question does not strike me as particularly interesting. After all, we all know that there are no methodological panaceas! Leaving the well-known limitations of case studies aside, let me mention here that I worry about the qualitative tradition of comparative political economy becoming impoverished. It seems to me that as case-oriented scholars have become increasingly interested in making broad comparative arguments, their "cases" have often become increasingly superficial in the sense that they are treated as single observations on some outcome variable and on a series of potential causal variables. More or less explicitly relying on Millian logic, many comparative case studies seem to reject certain explanatory arguments and embrace others simply by matching causal variables with outcomes across country cases. The conception of causality underlying such work is often strikingly similar to the linear-additive conception that informs simple-minded regression models, but with a highly deterministic twist (cf. Lieberson, 1992). As suggested above, I prefer case studies that explore causal processes and historical sequencing, treating each case as multiple observations of the variables of interest (see Swenson, 2002 and Thelen, 2004, for prominent recent examples). In my view, these are the kinds of case studies that are needed to complement regression analysis and to generate new analytical insights.[3]

It should also be noted that the quantitative political economy literature has recently begun to engage with individual-level data on economic insecurity, income, skills, social policy preferences and political behavior. The most developed strand of this new literature seeks to explore micro-foundational arguments of the earlier "macro" literature on globalization, deindustrialization and the welfare state (e.g., Iversen & Soskice, 2001). Another promising avenue in this vein is to exploit household-level data

from the Luxembourg Income Study to explore the determinants of the distribution of market income and the redistributive effects of the welfare state. While Shalev considers regression analysis to be entirely appropriate for the purpose of analyzing individual-level data, the position he adopts seems to preclude the next step: to incorporate macro-level variables into such analyses by way of hierarchical modeling (e.g., Anderson & Pontusson, 2005).

Nested or hierarchical modeling upholds the promise not only of integrating individual-level and country-level data, but also of addressing the question of why causal effects (or "statistical associations") vary across time or across clusters of countries. In the simplest version of this approach, sufficient to illustrate the logic involved, the coefficients generated by country-specific time-series models are treated as the "dependent variable" in a second cross-sectional model (e.g., Griffin, O'Connell, & McCammon, 1989). To be sure, the second-stage results still beg the question of causality, but this type of analysis does shed some light on "the structure of causality."

Finally, I would like to briefly comment on Shalev's argument that MR is inappropriate for comparative political economy because MR deals with marginal effects and presupposes theoretical precision. Is not MR actually an ideal method for testing imprecise theories? Most current theories in comparative political economy are probabilistic and might best be operationalized as "more of X will be associated with more of Y." With pooling, MR readily allows us to explore threshold effects or other non-linear direct effects. In my view, the paucity of empirical models that depart from the assumption of linear direct effects derives from the limitations of our current theories rather than methodological limitations.[4]

To sum up, I am in complete agreement with Shalev about the desirability of exploring causal syndromes that (may) vary across clusters of countries as well as variable-specific interaction effects. I agree that practitioners of pooled MR should be concerned with over-time heterogeneity in causal effects, and that we should pay more attention to historical dynamics in our theorizing. We should also be concerned about "limited diversity" and more technical issues pertaining to serial correlation and heteroskedasticity. We should not pool unless we have significant variation in the time-series observations on variables that we care about. And we should strive to retain visible cases for the purpose of theory-building as well as the exposition of theory and empirical results.

Shalev's paper is highly instructive, but fails to make the case that MR is an inappropriate methodology in comparative political economy. As suggested above, MR is a useful framework in which to explore some of the

methodological issues that Shalev raises. Rather than treating them as substitutes for MR, it might be more fruitful to conceive the "low-tech" quantitative methods advocated by Shalev as a bridge between MR and theoretically informed, process-oriented case studies.

ACKNOWLEDGMENT

For comments on a previous draft, I wish to thank Mary O'Sullivan, Michael Shalev and Bruce Western.

NOTES

1. Space also prevents me from engaging Shalev's discussion of "limited diversity" or, in other words, the problem of "empty cells." One quick and simple comment: I agree with Shalev that simulations of substantive effects should be based on reasonable ("within-sample") observations of the independent variables of interest.

2. Kwon and Pontusson (2005) use moving-windows analysis to estimate time-varying effects of government partisanship on social spending growth over the period 1962–2000. Our results indicate that Left governments were no more spending-prone than Right governments in the 1960s, but became significantly more spending-prone in the course of the 1970s and 1980s and that partisan effects declined sharply in the course of the 1990s. The results for the 1970s and 1980s come as good news to those who believe that government partisanship matters to the size of the welfare state over the long run. As Shalev notes (p. 30), previous literature that is not sensitive to temporal heterogeneity suggests that "political partisanship loses its explanatory efficacy once the design shifts from explaining levels to explaining dynamics."

3. Ragin's (1987) Qualitative Comparative Analysis provides a means to capture conditional causality based on case studies, but his approach is not sensitive to historical dynamics.

4. Illustrating this point nicely, Olson's (1982) encompassment thesis has inspired several models that estimate non-linear effects of labor-market institutions (e.g., Garrett & Way, 1999).

REFERENCES

Alvarez, M., Geoffrey, G., & Lange, P. (1991). Government partisanship, labor organization and macroeconomic performance. *American Political Science Review*, *85*(2), 539–556.

Anderson, C., & Pontusson, J. (2005). Workers, worries and welfare states. *European Journal of Political Research* (forthcoming).

Beck, N., & Katz, J. (1996). Nuisance vs. substance. *Political Analysis*, *6*, 1–36.

Esping-Andersen, G. (1990). *The three worlds of welfare capitalism*. Princeton: Princeton University Press.

Garrett, G. (1998). *Partisan politics in the global economy*. Cambridge: Cambridge University Press.

Garrett, G., & Way, C. (1999). Public-sector unions, corporatism and wage determination. In: T. Iverson, J. Pontusson & D. Soskice (Eds), *Unions, employers and central banks* (pp. 267–291). Cambridge: Cambridge University Press.

Griffin, L., O'Connell, P., & McCammon, H. (1989). National variation in the context of struggle. *Canadian Review of Sociology and Anthropology*, *26*(1), 37–68.

Hall, P., & Franzese, R. (1998). Mixed signals. *International Organization*, *52*(3), 505–535.

Hall, P., & Soskice, D. (Eds) (2001). *Varieties of capitalism*. Oxford: Oxford University Press.

Iversen, T., & Soskice, D. (2001). An asset theory of social policy preferences. *American Political Science Review*, *95*(4), 875–893.

Kwon, H. Y., & Pontusson, J. (2005). *The rise and fall of government partisanship*. Unpublished paper, Department of Politics, Princeton University.

Lieberson, S. (1992). Small N's and big conclusions. In: C. Ragin (Ed.), *What is a case?* (pp. 105–118). New York: Cambridge University Press.

Olson, M. (1982). *The rise and decline of nations*. New Haven: Yale University Press.

Pontusson, J. (2005). *Inequality and prosperity*. Ithaca: Cornell University Press.

Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry*. New York: Wiley.

Ragin, C. (1987). *The comparative method*. Berkeley: University of California Press.

Rueda, D., & Pontusson, J. (2000). Wage inequality and varieties of capitalism. *World Politics*, *52*(3), 350–383.

Swenson, P. (2002). *Capitalists against markets*. Oxford: Oxford University Press.

Thelen, K. (2004). *How institutions evolve*. Cambridge: Cambridge University Press.

Western, B. (1998). Causal heterogeneity in comparative research. *American Journal of Political Science*, *42*(4), 1233–1259.

MULTIPLE REGRESSION IN SMALL-N COMPARISONS

Gosta Esping-Andersen

INTRODUCTION

Michael Shalev has turned his attention, once again, to the bad methodological habits that social scientists – like myself – often adopt. As always, he presents us with thoughtful, rigorous, and penetrating criticism, but also with a generous dose of constructive prescription. His target is the widespread use of regression techniques in cross-national comparative research. The gist of the argument is that multiple regression (MR) is a far too blunt instrument if our aim is to arrive at a robust identification of crucial causal mechanisms. MR, as he puts it (p. 42), renders the cases invisible and, hence, precludes researchers from having any dialogue with them. The case becomes a set of scores; the causal mechanisms are reduced to correlation coefficients. As a result, analytical power is sacrificed rather than gained. Shalev advocates simpler 'low-tech' approaches such as tabular representations, tree diagrams, or clustering techniques either as substitutes for, or as companions to, regression analysis.

It is almost impossible not to agree with Shalev. As one of the three main targets of his paper, my *Three Worlds of Welfare Capitalism* analyses are, I am happy to see, not completely torn to shreds. The distinctiveness of welfare regimes more or less remains when subjected to alternative treatments, such as Shalev's factor analytical approach. I am more than ready to

Capitalisms Compared
Comparative Social Research, Volume 24, 335–342

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24009-2

concede that my use of MR to explain welfare regime differences was rather inappropriate for the purpose at hand. I am now older and perhaps also wiser, and would certainly have done it all differently today. But would I now follow Shalev's prescriptions? Yes and no. I am in full agreement that triangulation, i.e. combining MR with qualitative inspection, should be a favoured approach in small-N comparisons. I am less persuaded with his call for *substituting* regressions with more qualitative, lower-tech alternatives. Neither am I convinced that substituting MR with factor analysis (as Shalev does in his reanalysis of my data) will yield more analytical insight compared to scrutinizing residual plots from MR.

MR estimation on small-N country samples implies that we easily violate basic key assumptions, such as monotonic linear effects, statistical independence, the absence of selection bias, and conditional independence. Small-N regressions are therefore not very useful – and easily counterproductive – if used primarily to identify the strength of the statistical relationship. But all this does not mean that we should abandon MR.

Below I shall argue two major points. Firstly if MR is utilized as a diagnostic tool, explicitly aimed at detecting such violations, it provides, in my view, unrivalled potential for identification. Secondly, the 'low-tech' alternatives that Shalev espouses are not superior with regard to distinguishing wrong from correct causal mechanisms, in particular under conditions of selection and endogeneity.[1] My view is that we should use MR not to identify causal mechanisms via the βs, but rather as a 'Popperian' devise. The strength of a statistical association will not tell us much about the real causal mechanisms at work, but the diagnositics that we can obtain from MR residual plots are a minefield of information, truly powerful instruments for fine-tuning and possibly correcting our hypotheses, and subsequently for selecting appropriate alternative instruments. If our true aim is identification (following Manski, 1995), we should not throw MR out with the bathwater.

IDENTIFICATION WITH MULTIPLE REGRESSION IN SMALL-N COMPARISONS

We very often face important macro-level questions that cannot be answered. We usually have few cases but many rival explanations. We easily confound nation charactersitics with the dimensions we measure and, hence, it is basically unclear what explains what. The explicit aim of the 'politics

matter' literature is to demonstrate that left power (x) matters for welfare state development (y), conditional on a vector (z) of other plausible factors (such as economic growth). The standard approach is to sample the 20-odd OECD democracies and then regress the $y = f(x|z)$.

As Fearon (1991) insists, the smaller the sample size, the greater is the need to make counterfactuals explicit. In our sample we will observe Italy's welfare state size, and that Italy was ruled by Christian Democrats throughout the post-war era. The obvious counterfactual is that its welfare state would have been 'bigger' or 'better' had it been ruled by social democrats. In other words, our sample needs to include another Italy, a country that matches Italy on all relevant z values but differs on x. No such country is likely to be found in the 20-odd OECD sample and we are therefore left with no cell-match. When we add to this that small-N studies make it impossible to condition on all relevant z variables, true causal identification is for all practical purposes stifled.

Furthermore, as Shalev also argues, the choice of OLS regressions implies that we assume monotonic linear effects. Sweden is always the top-scorer on left power *cum* welfare statism. If MR gives us an $x\beta$ estimate of 0.3, we are then led to believe that a 5-point increase in left power (Denmark closes the gap with Sweden) will result in a 1.5-point increase in the Danish welfare state size. For Germany, the equivalent effect would be a 4.2-point increase. This is pretty much a non-sensical estimation and is, besides, not what the researcher should aim to identify. I readily admit my guilt in falling into this regression-trap on more than one occasion.

What we truly aim to uncover are the precise causal mechanisms that link x to y. Deep historical scrutiny of all the countries will, no doubt, help the researcher identify how, exactly, left power in Denmark influenced welfare state growth. Doing this rigorously for 20-odd countries would amount to a lifelong project. The reason that I support Shalev's call for triangulation is that MR can be employed very productively in the pursuit of more precise identification – in particular when aided by explicit and systematic use of counterfactuals.

The really valuable information in MR lies in the residual plots, not in the βs and R^2. Small N's are frustrating, but they do have the advantage that scrutiny of the residuals is easy: you can quickly put a name on each point. In this sense, Shalev errs when he claims that MR impedes dialogue with the cases. There are three key issues related to identification where good diagnostic use of MR can become a major asset: dependence among the observations, selection, and endogeneity.

INDEPENDENCE

If our observations are not fully independent of one another, we will violate a basic MR assumption. This will show up as heteroskedasticity in the residual behaviour. Dependency is theoretically interesting because it suggests that some cases follow a similar logic by way of, for example, mutually influencing each other. In time series analysis the logic is similar. If last year's values influence next year's we have some evidence in favour of a path-dependency hypothesis. Frank Castles (1993) took dependency to its logical conclusion in his *families of nations* argument. Many MR practitioners simply do not bother about heteroskedasticity. My *Three Worlds* study was premised on the idea of clustered regime logics and I should, therefore, have actively tested for independency.

There are two options if the assumption is violated. One can correct for it via country-group dummies (say a Scandinavia dummy). If Castles is right and there are four families, the dummy solution results in paralysis because it will exhaust just about all degrees of freedom. More to the point, regressing with 'family dummies' will not get us much closer to identifying real causal mechanisms. The second and much more alluring option is to launch an in-depth study (as Castles did) of *why* or *how* diffusion came about – one example of why triangulation via MR can be scientifically fertile. Had I seized upon this opportunity I would probably have paid far more attention to whether regimes emerge from similar behaviour on x or from a policy diffusion process that is unrelated to x. My thesis would clearly have encountered problems were the latter true.

When we regress welfare state size on left power we will inevitably identify a cluster that 'overshoots'. Putting names to the dots tells you immediately that they are countries with a strong Christian democratic tradition. Van Kersbergen (1995) noted this and subsequently conducted an in-depth examination of similarities and differences in the evolution of Christian Democratic welfare states. Here is another telling illustration of how dependency diagnosis plus qualitative analysis can yield good sociological research.

New countries are spawned from old ones almost on a yearly basis, and there are undoubtedly many MR practitioners that see this as a welcome addition of N's. United Nations membership has leapt by 50 percent in the past decade with the birth of new nations. One should, of course, not assume that the new Slovakia and Slovenia are statistically independent from the old Czechoslovakia and Yugoslavia. We might also ask ourselves whether intensified EU integration has diminished the degree of independence that once existed between, say, Finland and France. Some clues may be found in the MR residual plot.

SELECTION AND ENDOGENEITY

Our main challenge is to distinguish true from spurious relationships. Our inferences will be seriously biased if *y and x* are both the outcome of some, possibly unobservable, heterogeneity or if x is not truly exogenously determined. Selection bias and endogeneity are essentially two facets of a similar problem, namely that if they are present we will make incorrect conclusions regarding the causal mechanisms we care about. Using welfare state research again as illustration, it is very possible that strong social democracy *and* large welfare states are jointly determined by some unidentified factor that, perhaps, lies deeply buried in history. Take Sweden: the seemingly obvious connection between left power and welfare state growth may, in reality, be incorrectly identified. It is theoretically equally possible that both attributes of modern Sweden have their roots in any number of historical peculiarities, be it patterns of landholding in pre-industrial ages, the nature of absolutism, industrial structure, or the transition to democracy. We must, likewise, assume that the welfare state – once in place – will have had substantial influence on the social democrats' electoral fortunes, both in the short and long run. If so, the x for Sweden is influenced by y and the assumption of conditional independence is violated.

Selection and endogeneity are often difficult to detect and manage. The simpler methodologies that Shalev advocates are, as far as I can see, not better equipped to handle either, at least when compared to MR.

Selection bias may be related to observables or unobservables (Heckman, 1988). The former occurs when the expected covariance $E(u_j u|z) \neq 0$, but it disappears once we control for the observed variables Z, so that $E(u_j u|z) = 0$. The latter is present when $E(u_j u) \neq 0$ and $E(u_j u|z) \neq 0$. In this situation, controlling for the factors observed by the investigator does not remove the covariance between the errors in the outcome and the selection equations. Now note that the regression coefficient $\sigma_{ju} = \mathrm{cov}(u_j u)/\mathrm{var}(u_j)$. If selection is on unobservables, controlling for some variable x in the outcome equation may reduce the error variance u_j without equally reducing the covariance $u_j u$. Hence, the coefficient on the omitted variable will be larger and the bias will be exacerbated.

Accordingly, the expected values of the observed cases will be biased because they co-vary with the variable that determines which cases are observed. This bias can be corrected by conventional controlling procedures. But if bias stems from unobservables, such controls will only worsen the bias. As Heckman (opp.cit: 7) argues, the dilemma is that different methods of correcting for selection bias are robust if there is *no* bias to begin with; if there *is*, there is no guarantee that the methods are robust.

The problem is similar whether we study large or small N's (Fearon, 1991). De Toqueville provided a nice exemplification with his observation that revolutions do not seem to change anything. The reason might be that they occur only in countries where it is difficult to change society in the first place. Accordingly, even studies based on $N = 1$ may suffer from selection bias: the French revolution may have been caused by the same conditions that made social change so difficult. It is possible that a revolution in a country where social relations are easier to change would have provoked change. But then a revolution would not have been necessary.

This suggests that more qualitative case-specific methods that prioritize dialogue between researcher and the case hold no special advantage over MR as far as selection bias is concerned. In essence, the only genuine method to correct for selection bias is to construct counterfactuals, to fill in the unobserved values in the distribution of y for all x's. Comparative analysis of the case-study variety cannot benefit from statistical distributions to generate the counterfactuals. In this respect there is accordingly something to be said for methods, like MR, because they provide such distributions and because they permit us to estimate covariance coefficients.

The problems related to endogeneity are virtually identical and require, therefore, less elaboration. There are, however, a few small points to add. Endogeneity is present when our x's are conditionally dependent on y. Using welfare state research again to exemplify, this can be because social policies directly influence the parliamentary fortunes of social democracy (the Swedes love their welfare state and vote Left to ensure its continuity). It can also be because Sweden's welfare state *and* Sweden's unique variety of social democracy are part and parcel of 'everything that is Sweden'. In the latter case, the true x and y for Sweden is not left power, nor welfare statism, but a full list of all that is uniquely Swedish.

If this is so, the fixed-effects panel estimation approach will go wrong since it assumes that x will have an identical impact on y regardless of which country. But if the left power effect on welfare statism is 'Sweden' or, perhaps, 'Germany' specific we should assume non-identical effects. Similar to the identification of selection on observables, we might therefore introduce controls for everything that is Sweden or Germany specific. Small-N studies with strong endogeneity have little capacity to extend the number of potentially necessary controls. The solution is therefore, once again, to concentrate on the theoretical elaboration by means of counterfactuals.

One promising avenue is to redefine the dependent variable so that it is less likely to incorporate all that is Swedish, and/or so that it is less likely to directly pattern voters' party preferences. Indeed, the welfare state literature

has to a degree moved in this direction by replacing aggregate measures (such as social expenditure) with narrower indicators that measure specific properties of welfare states. However, the underlying problem may still remain if such properties are, once again, the mirror image of 'all that is Swedish' rather than verifiably related to specific values of $(x|z)$.

If MR is applied to a sufficiently large number of N's and used for diagnostic purposes, it can be a powerful and efficient method for detecting endogeneity – certainly superior to the kinds of low-tech alternatives discussed by Shalev. We do have good testing procedures to detect non-identical x-effects in fixed-effects regressions or, alternatively, we can use an IV approach within two-stage least squares estimation. These options are typically precluded in small-N studies and we are, therefore, back again to the importance of counterfactuals as our only realistic alternative.

In brief, my response to Michael Shalev's argumentation is that we should favour whichever method delivers superior information about the underlying statistical distributions. In some cases, MR may be the relevant choice; in others, possibly not. We should, above all, be careful not to throw the baby out with the bath water. MR has very powerful and easy-to-use diagnostic tools that can be mobilized for what statistical analysis really should pursue, namely to search for the true causal mechanisms. If, instead, we continue the past tradition of employing MR to show that our β and R^2 are bigger than others', then I agree whole-heartedly with Shalev. His lower-tech alternatives are less likely to produce violations of basic estimation assumptions than is the uncritical MR-based search for superior R-squares.

NOTE

1. Many of the points to be covered in this paper were previously examined in Esping-Andersen and Przeworski (2000). For illustrative purposes, I will draw primarily on examples from my own work on comparative welfare states.

REFERENCES

Castles, F. (Ed.) (1993). *Families of nations*. Dartmouth: Aldershot.

Esping-Andersen, G., & Przeworski, A. (2000). Quantitative cross-national research methods. In: *International encyclopedia of the social and behavioral sciences*. Section 2.3. Holland: Elsevier.

Fearon, J. (1991). Counterfactuals and hypothesis testing in political science. *World Politics*, *43*, 169–195.

Heckman, J. (1988). The microeconomic evaluation of social programs and economic institutions. In: *Chung-Hua series of lectures by invited eminent economists no. 14*. Teipei: The Institute of Economics, Academia Sinica.

Manski, C. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Van Kersbergen, K. (1995). *Social capitalism*. London: Routledge.

TOWARD IMPROVED USE OF REGRESSION IN MACRO-COMPARATIVE ANALYSIS

Lane Kenworthy

I agree with much of what Michael Shalev (2007) says in his paper, both about the limits of multiple regression and about how to improve quantitative analysis in macro-comparative research. With respect to the latter, Shalev suggests three avenues for advance: (1) improve regression through technical refinement; (2) combine regression with case studies (triangulation); (3) turn to alternative methods of quantitative analysis such as multivariate tables and graphs or factor analysis (substitution). I want to suggest some additional ways in which the use of regression in macro-comparative analysis could be improved. None involves technical refinement. Instead, most have to do with relatively basic aspects of quantitative analysis that seem, in my view, to be commonly ignored or overlooked.

LOOK AT THE DATA

Shalev's third suggested path for progress consists of using tables, graphs, and tree diagrams to examine causal hierarchy and complexity and to identify cases meriting more in-depth scrutiny. This should be viewed not as (or at least not solely as) a substitute for regression but rather as a critical component of regression analysis. All of us were (I hope) taught in our first

Capitalisms Compared
Comparative Social Research, Volume 24, 343–350

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24010-9

regression course that it is not enough to simply get the data, estimate some regression equations, and then draw conclusions. It also is necessary to get a feel for the data, in large part by examining descriptive statistics and looking at bivariate and/or multivariate patterns. Too many macro-comparativists, I suspect, either do not do this at all or do not do it sufficiently carefully.

In some instances what one finds by looking carefully at the data enhances or enriches what regression analysis tells us. Sometimes it calls into question the utility of using regression. Sometimes it suggests ways of altering the regression, for example by adding interaction effects, considering alternative functional forms of relationship, excluding certain cases, and so on. The tree diagram Shalev shows in discussing Bo Rothstein's (1990) analysis of determinants of unionization and the graph he uses in discussing Peter Hall and Robert Franzese's (1998) analysis of the impact of central bank independence and wage-setting coordination on unemployment are useful examples of what one can learn by spending a great deal of time looking at and thinking about the data. (That is not to suggest that either Rothstein or Hall and Franzese necessarily failed to do this. Sometimes we miss things, no matter how hard we look.)

It almost always is best to look at the data in graphical form. There are circumstances in which we can spot interesting patterns in tables. But it is much easier to do so when data are displayed graphically (Cleveland, 1993, 1994; Tufte, 2001; Wilkinson, 2001; Gelman, Pasarica, & Dodhia, 2002). Happily, these days the investment required to learn how to create both simple and relatively complex graphs is minor.

SHOW THE DATA

Quantitative analysis involves data reduction. But in my view, most recent quantitative macro-comparative work goes too far in this direction. The typical analysis includes 18 or so countries. In this type of research, unlike in analyses of thousands of individuals, the cases both are of substantial interest in and of themselves and can matter for our interpretation of regression findings.

The typical paper includes a few tables showing regression results and perhaps an appendix listing means and correlations among the variables. This is helpful information. But much more could be made available to readers. In particular, it is possible without taking up too much space to let readers see most of the raw data. In a cross-sectional macro-comparative paper the author can actually list all of the data used in the analyses – that

is, the values for each country on each variable – in a table. For analyses that utilize longitudinal data, graphs displaying the time series for each country for key variables can be included. Bivariate or multivariate scatterplots can help readers (and authors) to see patterns in the data that warrant scrutiny.

GREATER TRANSPARENCY IN PRESENTING REGRESSION FINDINGS

In the typical textbook explication of multiple regression, the author shows the results of a bivariate regression (one independent variable), then introduces the notion of spuriousness and the concept of "controlling" with non-experimental data, and then proceeds to add one or more additional independent variables. The coefficient for the original independent variable changes, and the reader thereby learns about partial associations and omitted variable bias.

This analytical strategy is not only useful for pedagogical purposes. It is an appropriate way to proceed in "independent-variable-centered" analyses. In such analyses the research question concerns the effect of one (or sometimes two or three) independent variable on the outcome. The question is "what is the impact of X_1 on Y?" Sometimes, by contrast, the research is "dependent-variable-centered": the research question is "what causes Y?" In a dependent-variable-centered analysis it may be more appropriate to begin with a large number of (theory guided) independent variables and then gradually reduce the number according to criteria such as statistical significance or contribution to adjusted R^2.

Most analyses in macro-comparative research are independent-variable-centered. The question is something like "What is the effect of left government on social policy generosity?" or "What is the impact of wage-setting arrangements on unemployment?" Yet most analysts proceed by including as many controls as possible in their initial regression. Sometimes this is the only regression presented; in other instances some of the variables are then dropped and a second (and perhaps third and fourth) regression is shown.

A common circumstance is that we have fairly strong reason to suspect there will be an association between the hypothesized causal factor and the outcome, and the expected association is there at the bivariate level, but then it disappears in a multivariate analysis. Also common is that we have a not-terribly-compelling theory suggesting a link but no bivariate association, yet in the regression with 10 or so control variables the association appears.

Sometimes these multivariate findings are correct. But we should be suspicious. Those who have done enough multivariate regression analysis know well that it is sometimes (not always) possible to get the expected and/or hoped for finding to emerge if enough model specifications are tried.

As researchers and as consumers of others' research, we should want to know exactly how such a finding has emerged. That requires going step-by-step through the regressions, from bivariate patterns to the results of adding each of the various controls. Which particular control or set of controls makes the association change? Is that particular specification more theoretically compelling than others? How robust is the association to alternative specifications (not to mention measurement choices, groups of countries, and time periods)? Walk the reader through the analyses and findings. Allay suspicion by making it as transparent as possible what is going on in the data. Of course, space constraints typically permit showing only a limited number of the regressions. But the reader should nevertheless be informed of exactly what produced the result for the variable of interest in the preferred model specification.

WHICH VARIATION?

Macro-comparative analysts who use pooled cross-section time-series regression often fail to make clear what variation they aim to explain. There are three main options. One is variation in levels across countries. Here one can estimate cross-sectional effects averaged over multiple time periods (years, business cycles, decades). An example might be the impact of left government on welfare state generosity across 20 countries, averaged over the 1980s and 1990s. A second is variation over time within countries. Here regression can estimate an average over-time effect for a set of countries – for instance, the effect of left government on change in welfare state generosity in the 1980s and 1990s, averaged over 20 nations. A third is cross-country variation in change over time. We might, for example, be interested in the impact of left government on cross-country differences in change in welfare state generosity in the 1980s and 1990s.

Pooled regressions usually focus on one or the other of the first two of these, and most commonly on both. Following Larry Griffin, Walters, O'Connell, and Moor (1986) and Kittel (1999), Shalev rightly notes that a common problem with use of pooled regression in macro-comparative research is that researchers combine these two types of variation without (apparently) considering whether it is reasonable to expect that the

causal process will be the same for both. Often that assumption is questionable.

Suppose cumulative left government is a major determinant of cross-country variation in welfare state generosity across 20 OECD countries as of 1980. But suppose it then has little or no effect on developments within these countries during the 1980s and 1990s. Perhaps over-time changes during these two decades are dominated by budget pressures and globalization. A pooled regression that does not distinguish between the determinants of cross-sectional variation vs. over-time variation will miss something very important in this type of situation.

Explaining cross-country variation in over-time changes is something different altogether. Suppose changes in budget pressures and globalization account for a significant portion of the longitudinal variation in welfare state generosity within each country in the 1980s and 1990s but that neither varies much across the countries. These two factors will not, then, help in explaining the differences between the countries in the direction and degree of over-time change. Those differences might instead be due to catch-up effects or to variation among the countries in public support for generous benefits or in the structure of the political system.

Fig. 1 illustrates these hypothetical differences in types of variation, using data on public social expenditures as a share of GDP. Setting aside the

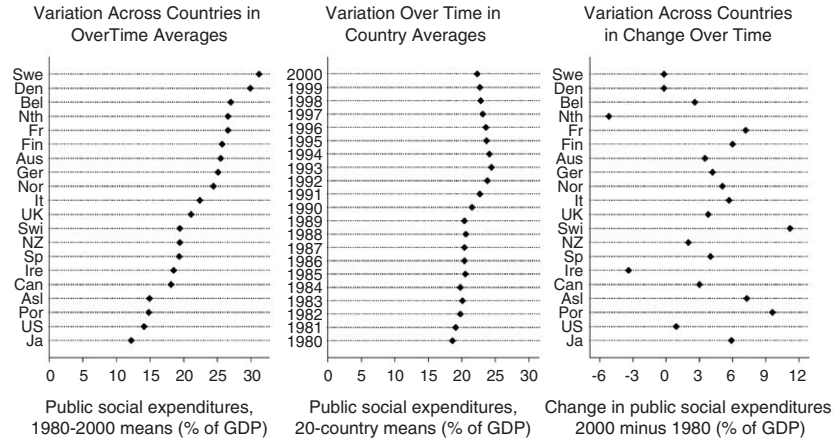

Fig. 1. Three Types of Variation in Public Social Expenditures. *Note:* Author's calculations from data in OECD (2004). The ordering of countries in the third chart follows that in the first chart, to highlight the contrast.

question of whether this is a useful measure of social policy generosity, the three charts show clearly that there are sizeable differences. This suggests the possibility of differing causal processes.

Macro-comparative researchers need to be clear about the type of variation to which their theory applies. And for empirical analysis the default assumption should be that causal patterns for cross-sectional and over-time variation differ. The utility of pooling should be demonstrated rather than presumed.

LONG-TERM VS. SHORT-TERM EFFECTS

Many pooled regression analyses use annual data. As best I can tell, most of the time that is because such data are available and because using them increases the number of observations, allowing for inclusion of more independent variables and enhancing statistical power. But many of the theories such analyses test imply medium-to-long-term effects. Sometimes analyses with annual data can pick up such effects, but that hinges on getting the lag structure correct. More often than not, using annual data to examine hypothesized medium-run or long-run associations will obscure rather than clarify.

But using longer time periods reduces the number of observations, heightening concern about omitted variable bias. What to do? There is no ideal solution. My preferred strategy is to examine all possible combinations of a "reasonable" number of independent variables (Kenworthy, 2004, 2007). For an N of 15 or so, that means perhaps three or four. This by no means eliminates worry about biased results due to improperly specified models. But inclusion of more independent variables is not inherently better in this regard (Lieberson, 1985; Achen, 2002). And in any event, having a better specification is not an improvement if the time period is wrong.

STATISTICAL SIGNIFICANCE?

Over the past decade much of the methodological debate in quantitative macro-comparative research has focused on how to properly estimate standard errors in pooled regressions. But in most instances such analyses include the full population of affluent countries in the time period considered. Where a sample is used, the sample is almost always dictated

by data availability; there is no pretense that it is representative of the population.

Statisticians disagree about whether there is a rationale – based on the "superpopulation" notion – for considering statistical significance in this type of circumstance (Berk, 2004, offers a useful discussion). At the very least, however, analysts who believe standard errors are important to consider should offer an argument in favor of doing so, instead of simply doing so because it is conventional practice. Either way, many macro-comparative analyses would be substantially improved by paying more attention to the direction, size, and robustness of regression coefficients and less to statistical significance.

REGRESSION AS THE ANALYTICAL STARTING POINT

Because we often are dealing with the full population, macro-comparativists should treat analyses less as a means of drawing generalizable inferences and more as a means of understanding the cases (Ragin, 2001). In the prototypical quantitative macro-comparative article, the regressions are the starting and ending point of the analysis. I would like to see more papers in which regression is used to inform discussion of cases. What do the regression results tell us about why country A or regime-type B turned out as it did or changed in the way it did? Discussion of cases can then, of course, be used to question and/or further explore the regression results.

REFERENCES

Achen, C. H. (2002). Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science*, *5*, 423–450.

Berk, R. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician*, *56*, 121–130.

Griffin, L. J., Walters, P. B., O'Connell, P., & Moor, E. (1986). Methodological innovations in the analysis of welfare-state development: Pooling cross sections and time series. In: N. Furniss (Ed.), *Futures for the Welfare State* (pp. 101–138). Bloomington: Indiana University Press.

Hall, P. A., & Franzese, R. J., Jr. (1998). Mixed signals: Central bank independence, coordinated wage bargaining, and European Monetary Union. *International Organization*, *52*, 505–535.

Kenworthy, L. (2004). *Egalitarian capitalism*. New York: Russell Sage Foundation.
Kenworthy, L. (2007). *Jobs with equality*. New York: Russell Sage Foundation (forthcoming).
Kittel, B. (1999). Sense and sensitivity in pooled analysis of political data. *European Journal of Political Research*, *35*, 225–253.
Lieberson, S. (1985). *Making it count*. Berkeley: University of California Press.
OECD (Organization for Economic Cooperation and Development) (2004). *OECD social expenditure database: 1980–2001*. Paris: OECD.
Ragin, C. C. (2001). Case-oriented research. In: *International encyclopedia of the social and behavioral sciences* (pp. 1519–1525). Amsterdam: Elsevier.
Rothstein, B. (1990). Labour market institutions and working class strength. In: S. Steinmo, K. Thelen & F. Longstreth (Eds), *Structuring politics: Historical institutionalism in comparative analysis* (pp. 33–56). Cambridge: Cambridge University Press.
Shalev, M. (2007). Limits and alternatives to multiple regression in comparative research. *Comparative Social Research*, *24*, 259–308.
Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
Wilkinson, L. (2001). Graphical methods: Presentation. In: *International encyclopedia of the social and behavioral sciences* (pp. 6368–6379). Amsterdam: Elsevier.

HOW TO GET AT CAUSALITY IN THE SOCIAL SCIENCES: MULTIPLE REGRESSIONS VERSUS CASE STUDIES

Bo Rothstein

I am grateful to the editors of this journal to be given the possibility to comment on Michael Shalev's article. Although I have some minor disagreement with his general argument, I am also grateful to Michael Shalev for taking up what I think is an important question in comparative social science. I find myself in the curious position of being a target of a general critique that I mostly agree upon, namely that too much energy is going into sophisticated methodological techniques at the expense of substantive knowledge about individual cases and theoretical reasoning about causality. However, and probably not surprisingly, I find Shalev's critique of my particular venture into this area far from convincing.

Let me start with the former problem. When I look out from the window at my office, I can see an interesting regularity of behavior, namely that cars make a stop when there is a red light. Since I sometimes work late when there is very little traffic, the case becomes really puzzling because the cars stop even if there are no other cars at the intersection or any pedestrians that want to cross the street. The question is how we should explain this strange pattern of behavior. So, I went over the street to my colleagues at the

Capitalisms Compared
Comparative Social Research, Volume 24, 351–360

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24011-0

department of economics and asked. ''Piece of cake'' they say. The drivers stop because there is a clear and well-known incentive structure – they will get fined if they don't make a stop. As usual the economists' answer did not convince me since I can see that cars stop even if it is in the middle of the night when the chance of being caught by the police is miniscule.

My colleagues at the Göteborg Institute of Technology gave a very different answer. According to their findings, the cars came to a halt because there is a complicated technical device in the car linking the downward movement of the brake pedal to the brakes connected to the four wheels. Without this technical device, the cars would not make a stop. But then my friend, the brain neurologist, argued that the reason the cars came to a halt was that there is a neurological link in the brain that connects what the drivers' eyes register to a certain movement of the drivers' leg that makes them move the leg when they register that the light shifts from green to red. Without this neurological link from brain to leg, drives would not be able to stop the car when the light turns red. ''This is all very silly'', my friend the sociologist later told me. The reason that the cars stop is that drivers in Sweden have been socialized since childhood to internalize a norm that when there is a red light, the appropriate thing to do is to stop. However, drivers in some other countries are much less socialized into this norm. They would drive even if there was a red light. We should carry out a comparative project on this. Later, at a conference, I met some colleagues working on evolutionary game theory. Their answer was that by trial and error, drivers had found stopping at red light a useful ''convention'' that would change what used to be a ''chicken-race'' type of game to an ''assurance game''. But to the question how this all started they had no answer that was based on any empirical research.

My point is this – causality in the social sciences is a hard thing since it operates on different levels (the individual, the organizational and the societal). Finding out exactly why agents do what they do is very difficult. If it is large aggregates of agents (the working class, the employers, the politicians) it becomes even more complicated. The most common explanations are either self-interest, or social norms, or values. When it comes to self-interest, there is now tons of empirical work showing that this explanation is of limited value (Fehr & Fischbacher, 2002; Gintis, 2004; Jones, 1999). Social norms are certainly important, but that begs the question why people in different social settings have so different social norms. Explanations that are based on values tell us that people do things because they want to do them. This is probably correct but not much of an explanation because it is not very far from saying that people do things because they do them. People

vote on the parties the like, marry the partners the love and join or do not join unions because the like or dislike what the unions do. In other words, explanations based on values are close to just repeating the behavioral data. The modern fad in this is that people do things because they want to handle risks. But then we lack an explanation why are people in different societal settings so different when it comes to how they perceive risks?

While I agree with much of Michael Shalev's critique against the use of multiple regressions, it is far from new. When I wrote my dissertation about how the Swedish Social Democrats implemented to of their major social reforms (the active labor market policy and the comprehensive educational system), instead of a quantitative approach I used the comparative historical case study method. I was then inspired by the methodological work by Alexander George who argued that what happens in the few cases many variables situation is that one adds the one independent variable after the other in order the increase the r2 (George, 1979). Very soon, what happens is that one thereby describes each case a unique (Switzerland? add referendums; United States? add the Supreme court; the Netherlands? add consociationalism, Germany? add Federalism). Alexander George's argument, which I followed and still think is the best, was that a theoretically motivated selection of a few cases for which the researchers tries to trace the process of how the main variables has been connected over time was superior to statistical methods such as multiple regressions.

I also agree with Shalev's argument that no statistical or econometric technique can replace theoretical reasoning about causality. This is of course based on a meta-theoretical standpoint – I happen to be a silent but card-carrying member of what is known as the "realist" school in the social science. This implies that you want theories that not only make good predications but also are in line with reality (MacDonald, 2003). The problem is that such theories become pretty complicated if you are going to explain human behavior because to a large extent, what people do depends on what they think that "the others" are going to do (Shapiro & Wendt, 1992). For example, you may join the union of you think that enough other workers also will join, because it makes no point to be a member and pay dues to a union that is weak or ineffective. My argument in this specific case is that if there is some kind of institution in place that makes it likely that most workers believe that most other workers will join the union, you will get strong unions regardless of social norms, values or self-interest. The French case is maybe telling – it seems that many French workers are likely to act in solidarity with other workers when there is a "grand issue". However, because the French unions lack an institutional device as described above,

the degree of unionization in France is among the lowest in the OECD countries.

As for the particular article of mine that Shalev criticize, my comments are that his critique is unfair, unconvincing and misses the main point. It is unfair because readers get the impression that my whole argument about the importance of the Ghent system for explaining degrees of unionization is based solely on using the multiple regression method. The fact is that eighty percent of the article contains exactly what Shalev is asking for, namely a detailed historical case study based on original archival data explaining under what social and political circumstances it was possible to introduce this type of unemployment insurance system in Sweden, the rationale of why the agents did it and what political effects it had in this particular case compared to other cases. Using secondary sources, the particularities of the Swedish case is then briefly compared to other the historical situation in countries such as the UK, Denmark, France and the Netherlands. It is also unfair to criticize me for not having a ''well-developed'' theory of why workers would be more inclined to join unions if the unions control who will get support if claiming to be unemployed. In this case, I constructed a combination of three very ''well-developed'' theories, namely Michael Lipsky's theory of the importance of decisions made by ''street-level bureaucrats'', Mancur Olson's theory of the problem of collective action and Marxist theory that the institutions that influence the ''buying and selling'' of labor force should be the most important if one want the understand power relations in a capitalist society. Since there are literally thousands of institutions in any given society, one has to have a theory why some institutions are more important than others. Anyway, these three theories operate on different levels (the societal, the organization and the individual) and I connected them according to a model I had developed in my earlier work (Rothstein, 1986, 1996). The theory goes as follows: Since what constitutes being unemployed can always be questioned (what type of work at what wage should the jobless person have to accept or how far should he/she have to move to find work without risking to loose the benefit), the decisions made by the ''street-level'' bureaucrats that implement the unemployment insurance becomes essential. Since workers in a country with a Ghent system ''know'' this, they are likely to join unions since it is union officials who make these decisions. Secondly, this power over the process of implementation gives the unions what Mancur Olson named a ''selective incentive'' that makes it easier for them to overcome the problem of collective action. Lastly, the Marxist theory would tell us that more unions have power over institutions that influence the ''buying and selling''

of labor power, they will be a stronger force in society. My theory may of course be inaccurate, but Shalev's statement that my argument lacks "causal meaning" and therefore is not "theoretically plausible" is simply not valid. It should be added that I substantiated my argument for how the causal mechanisms operate by referring to two surveys that were carried out in Sweden during the 1970s. My interpretation of the results from both these surveys supported the causal argument I made. If Shalev wanted to criticize the way I theoretically specified the causality, he should have criticized my theory as it was presented and/or come up with an argument that refuted my interpretation of the results from these surveys.

Moreover, when the article was published, I had already published two articles and two books which in detail described the historical particularities of the Swedish (Rothstein, 1986, 1987, 1990, 1992a, 1996). In these books and articles, I do exactly what Shalev asks for, namely describe how the strength of the union movement increased the political support of the Social Democratic party which in its turn used its political power to strengthen the union movement, and so forth. In the article he criticize, I could of course have presented more historical material of this kind, but these things take a lot of space and I am probably not the first author who have had to limit what can be done in an article for an edited volume. In any case, the references to my earlier work are there so the existence of this research of mine cannot come as a surprise to Shalev. In sum, the argument that I should have made the particularities of the Swedish case "invisible" by concentrating my research on using multiple regressions is not born out by the facts, neither in the specific article Shalev criticize, nor on my other work. The argument that I have simply used "linear models" and been insensitive to the dialectics of social processes (so-called feed-back mechanisms, or to use Shalev's term, "interactions effects") is simply not true. On the contrary, this has been a main point both in the specific article he criticize and in my previous work.

One part of Shalev critique is simply impossible to understand. He writes that instead of showing the results from the regression, I should have shown "tabular or graphical presentation of the dataset with named observations" which would have permitted selecting "outlayers" for further historical analysis. Exactly such a table is presented in the article *before* the results of the regression is shown. My comment to this table and the regression in the paper reads:

> Taking the "visual" result from table 2.1. Into consideration, we can say that it is possible to have a fairly strong union movement without a Ghent system, but that in order to have really strong unions, such a system seems necessary. It must be recalled,

> however, that this statistical analysis does not help us understand how the causal link operates. It might very well be true that already very strong labor movements have introduced Ghent systems, rather than vice versa. In order to get a handle on this problem, we must go from static comparison to diacronic comparative analysis. (Rothstein, 1992b, p. 42f)

It is strange to be criticized for having omitted things that are in the article. Shalev's critique becomes even more puzzling since he argues that I make a mistake by refuting two other explanations for the variation in union density. The first one – the size of the potential membership – had been put forward by Michael Wallerstein and the reason I refuted it was because I could not find a theoretically plausible argument for how the logic at the micro-level could operated. This is exactly what Shalev asks for, namely that one should not believe in the causality of independent variables no matter how statistically significant they show up if there is not "a well-developed theory". The other variable that I doubted was the strength of left party participation. The reason I did this is the by now well-known "feed-back" mechanism problem that my earlier historical work had shown to be at work. Left parties in government are likely to enact laws, policies and regulations that will strengthen the unions and the unions will in their turn use this strength to support the electoral campaigns of left parties, and so on. What explains what (is it strong unions that give rise to left party governments or is it left party governments that explains strong unions) can as I wrote in the article not be solved by using multiple regression. When Shalev states that I disregard that the effects of one of the variables can "be conditional on the value of other variables" he is simply making things up. I find somewhat puzzling that Shalev criticizes me for things I actually have done that he argues should be done.

Moreover, Shalev's argument that the importance of a Ghent system is overblown in my article is unconvincing. His argument is that even in "well-matched" countries like Belgium versus the Netherlands or Sweden/Denmark/Finland versus Norway, there may be other factors that explain the huge differences in unionization. His argument for why Ghent system Belgium at that time had a 74 percent degree of unionization while the non-Ghent system Netherlands had only 29 percent is that Belgian unions are stronger at the work place. But the reason a union movement is stronger at the work place is very likely due to the fact that it has control over the unemployment insurance. For the difference between the otherwise very similar Nordic countries (were non-Ghent Norway has a degree of unionization 26 percent below the other Nordic countries), Shalev argues that this may be because the Norwegian unions had "lesser effectiveness" in

recruiting new members from the 1960s an onward. But according to my theory, this may be precisely because Norwegian unions lacked the type of "selective incentive" that I argued the Ghent system provides. Lastly, Shalev is not in line with the historical facts that I report in the article when he states that "only countries with a substantial left had the Ghent system". As I show in the article, in Denmark it was a conservative government that introduced the Ghent system while in Norway and the Netherlands it was left governments who replaced such systems with government controlled unemployment schemes. I find it problematic that Shalev argues for research that is more historically contingent, but when this research does not fit his argument, his simply dismiss the facts.

What is behind all this nonsensical critique from a seasoned social scientist like Michael Shalev? My guess is that Shalev wants to rescue the so-called "power resources" model that he and others had developed and that my research showed to be incomplete for explaining the variation in "working class strength". In short, the power resource model argues that the stronger the political power of Social Democratic parties, the more social policies for equality would be enacted. What has been omitted in this theory is a simple yet important question, namely why do some countries have "more Social Democracy" than others? This is where Shalev's critique misses the point of my article. He portrays the argument as if the question was about how many fractions of a percentage of the degree of unionization the existence of a Ghent-system can explain. As is clear from the article (and the volume it was published in), this was never the main question. Instead, the problem was how the dramatic variation in organizational strength of the working class could be explained. Or in other words, why some countries are "more Social Democratic" than others? I found it problematic that the cherished power-resource theory was (and still is) silent on this central question. Assuming that the thirst for Social Democracy is not genetic in some populations or has to do with inherited ancient social norms, the question I posed was what role *political institutions in general* could have played in the development of this astonishing variation in union strength. This also had to do with the wider theoretical agency-structure debate in the social science, since the question I posed was s if it was possible to find political agents within the labor movement that had the power to establish institutions that would increase the future organizational strength of the working class. Moreover, the question was if these agents also created such institutions with this strategic goal. This is a central question in institutional analysis, namely if institutions just evolve as functionalist responses to diverse and unconnected societal forces, or if they can be designed by agents so

as to alter power relations in a society (Thelen, 1999). For any type of social science that wants to be policy relevant and not just give "after the facts" explanations, such knowledge is of course central. The power resource model is cleansed of such political agency – the power of the organization strength of a county's working class just rains down (or not) like manna from heaven. The power resource model has been quite successful for explaining that politics matters for policy, but not for explaining why different countries have different (read more or less Social Democratic) politics.

What I was able to show was that in just one particular case (Sweden) it was possible to find a political agent that both understood the future logic of the institution he enacted and had the power and political skill to establish it. But I could also show that most agents misunderstood the long-term consequences of the institutional devices they debated. For example, while the leadership of the powerful Metal Workers union in Sweden wanted to support the introduction of the Ghent system, its more left-oriented members voted against the system at no less than three union congresses during the late 1930s.

The argument I made was thus that Ghent system was but just an example of this institutionalist theory. Moreover, I underlined that there could be other types of institutions that would have the same effect (think about a society in which the unions have control over the health insurance system and where union officials decided what types of medical treatments would be covered by the insurance). In my own (now almost twenty year old) words: "we should concentrate on *political institutions directly affecting the* relations *of production*. In common language this means labor-market institutions or policy taken in a broad sense, including such things as rules governing the right of labor to organize and take collective action against capitalists, unemployment policies, training programs, etc." (Rothstein, 1992b, p. 23). For example, the comparatively high degree of unionization in non-Ghent Norway can in all likelihood be explained by existence of other institutional devices the unions has influence over and that to some extent plays the same role as a Ghent system. An example if how this worked can be found in Svein Andersen's fine comparative study of how industrial relations evolved differently in the British and Norwegian off-shore industry (Andersen, 1988). I can also add that after the Social Democrats lost the election in Sweden in September 2006, the first major quarrel between the union movement and the new Conservative led government is .... quess what? The unions strongly oppose the new governments' policy to make the unemployment insurance mandatory and thus disconnect it from the unions' control (see for example *Dagens Nyheter*, 2006-10-24).

To go back to the general problem with using multiple regressions, I think Shalev misses one of the major problems. I am thinking about the argument put forward by Peter Hall that ''the ontologies of comparative politics have substantially outrun its methodologies'' (Hall, 2003). Hall's first argument is that we often assume unit homogeneity while we know that this is not the case. For example, six years of Social Democratic rule in the 1930s are not equivalent to six years of Social Democratic rule in the 1980s. Secondly, the development of an ontology that recognizes strong feed-back mechanisms and lock-in effects between variables over time (such as the relation between union strength and Social Democratic electoral success), is not compatible with the idea that the world consists of variables than can be clearly distinguished by labeling them ''independent'' and ''dependant''. Strategic interaction or institutionally induced pay-offs that serve to strengthen the reproduction of that very institution, are but two examples of this problem. Thirdly, we have observations that the event(s) that ultimately puts a system on to a specific historical ''path'' leading to a unique equilibrium (such as the establishment of a Ghent system), may have occurred at ''formative moments'' very early in the process. Hall's point is well taken, namely that such ultimately important variables that are to be found in a ''distant past'' are hard to capture by using the standard regression method. Hall's main recommendation for aligning ontology and methodology in comparative politics is that analysis should be centered on the tracing of processes so that we can uncover how the causal mechanisms operate. This is what I have tried to do in my work and therefore Shalev's way of portraying what I have done is not only inaccurate in its details but also a misleading description of my research.

REFERENCES

Andersen, S. S. (1988). *British and Norwegian offshore industrial relations: Pluralism and neo-corporatism as contexts of strategic adaptation*. Aldershot: Avebury.

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter – The impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal*, *112*, C1–C33.

George, A. L. (1979). Case studies and theory development: The method of structured, focused comparisons. In: P. Gorden Lauren (Ed.), *Diplomacy: New approaches in history, theory and policy*. New York: Free Press.

Gintis, H. (2004). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.

Hall, P. (2003). Aligning ontology and methodology in comparative politics. In: J. M. Mahoney & D. Rueschmeyer (Eds), *Comparative historical analysis in the social sciences* (pp. 373–407). New York: Cambridge University Press.

Jones, B. D. (1999). Bounded rationality. *Annual Review of Political Science*, *2*, 297–321.
MacDonald, P. K. (2003). Useful fiction or miracle maker: The competing epistemological foundations of rational choice theory. *American Political Science Review*, *97*, 551–565.
Rothstein, B. (1986). *Den socialdemokratiska staten. Reformer och förvaltning inom svensk arbetsmarknads – och skolpolitik*. Lund: Arkiv Förlag.
Rothstein, B. (1987). Corporatism and reformism: The social democratic institutionalization of class conflict. *Acta Sociologica*, *30*, 295–311.
Rothstein, B. (1990). Marxism, institutional analysis and working-class strength. *Politics and Society*, *18*, 317–345.
Rothstein, B. (1992a). *Den korporativa staten. Intresseorganisationer och statsförvaltning i svensk politik*. Stockholm: Norstedts.
Rothstein, B. (1992b). Labor-market institutions and working-class strength. In: S. Steinmo, K. Thelen & F. Longstreth (Eds), *Structuring politics: Historical institutionalism in a comparative perspective* (pp. 33–56). Cambridge: Cambridge University Press.
Rothstein, B. (1996). *The social democratic state: The Swedish model and the bureaucratic problem of social reforms*. Pittsburgh: University of Pittsburgh Press.
Shapiro, I., & Wendt, A. (1992). The difference that realism makes. *Politics and Society*, *20*, 197–223.
Thelen, K. (1999). Historical institutionalism in comparative perspective. *Annual Review of Political Science*, *2*, 369–404.

WHAT COMPARATIVISTS REALLY DO

Duane Swank

Michael Shalev has made several important contributions to the fields of comparative political and social research; among his most prominent work is his seminal and more recent contributions to the study of the welfare state (e.g., Shalev, 1983, 1996). The current essay on the problems inherent in the use of multiple regression techniques to test rich, complex theories in comparative politics adds another important article to this set of works. Indeed, Shalev is certainly right in arguing that, in some portion of comparative research that relies principally on multiple regression (hereafter MR), the ''real world'' cases of comparative analysis (e.g., national states, institutions, collective actors) have been at least partially ignored and that tests of contingent and conjunctural causal arguments have been notably oversimplified within the linear, additive analytical framework of regression analysis. Many other problems mentioned in the essay, namely, that generic problems of statistical control and inference abound in quantitative comparative analysis and that the promises of ''high powered'' techniques such as pooled time-series cross-section (or panel) analysis are overly optimistic, are too frequently ignored or underestimated by researchers. Yet, on the other hand, a careful review of the actual work of contemporary political analysts – the very scholars Shalev singles out as among the leading practitioners of MR and pooled time-series cross-section (hereafter PTSCS) analysis – suggests

Capitalisms Compared
Comparative Social Research, Volume 24, 361–372

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24012-2

that a substantial number of current comparativists are much more attuned to the central problems Shalev highlights than his essay admits.

Shalev makes a series of criticisms against the general use of MR in comparative political research. He also argues that PTSCS analysis offers little escape from these general weaknesses of regression analysis and that PTSCS analysis presents quantitative comparativists with additional difficulties that make their plight worse rather than better. Ultimately, his answer is to use a visually oriented case-variable method where cases, properly named, are brought center stage for (non-statistical) analyses of (co)variations among two to three complex variables; this analysis is conducted primarily through use of tabular or graphic devises (e.g., tree diagrams; two-dimensional graphs). In the following pages, I will focus in some detail on what I believe are two of his most important critiques of the use of MR in comparative analysis; I will offer only a brief commentary on what I believe are less important or pressing issues connected to the use of MR, and on Shalev's critique of PTSCS analysis. Again, the thrust of my comments on Shalev's generally useful discussion is that he misses a lot of what contemporary researchers actually do (i.e., how they design and execute research). In essence, comparative political analysts recognize most of the problems and difficulties he highlights in his essay and they actually offer quite sophisticated responses to these problems much of the time.

INVISIBLE AND COMPLEX CASES

Perhaps one of the most important points Shalev makes is that Przeworski and Tuene's (1970) admonition to comparative political analysts to replace the proper names of cases with concepts and variables, or to pursue what Ragin (1987) dubs variable-oriented research, has gone too far. In Shalev's view, cases (typically nation states for the purposes of this discussion) have become all but invisible. This is particularly troublesome in Shalev's mind because, at least as far as comparative analysis of developed democratic capitalist systems is concerned (and we can say the same for political research on Latin America, Africa, or Asia), the cases are few enough to know quite well and bring to the forefront of sophisticated analysis.[1] In addition, Shalev makes the distinct point that the theories we seek to test in comparative political research entail complex and often non-linear causal sequences: causes of particular political outcomes are commonly contingent on the presence of other forces, or conjunctural with temporally and spatially bound forces and contexts. In fact, in comparative theory, it is fair to

argue (as Shalev does) that causal explanations of important political outcomes are often put forward in terms of complex configurations of multiple factors. Moreover, in theory and in practice, we are often confronted with the prospect of multiple configurative paths of causation of the same outcome. In the end, Shalev believes that the linear and additive logic of general MR analysis, as well as the more sophisticated versions with non-linear specifications and interaction terms, cannot adequately test our complex theories.

The Usual Suspects

I wish to address these two important sets of issues by focusing on the work of scholars that Shalev identifies as visible and sophisticated practitioners of MR analysis. In the introduction to his essay, he identifies Carles Boix, Robert Franzese, Geoffry Garrett, Tobern Iversen, and myself, scholars associated with the Cambridge Studies in Comparative Politics series of Cambridge University Press, as the ''usual suspects'' who practice advanced MR analysis (including PTSCS) and who use analysis of cases only ''in a subsidiary role.'' Although Shalev offers a detailed critique of one example of Franzese' work (Hall & Franzese, 1998) as well as Garrett's (1998) well-known book on globalization and national policy autonomy, he mentions only in passing other key works of the rest of this group. As in the world of film, the usual suspects have been rounded up and charged without a full examination of the evidence.

In addition, while he discusses the contributions of Alex Hicks to the literature on PTSCS analysis, Shalev ignores Hicks (1999) highly visible book on the political economy of the welfare state. This is a particularly important work for current purposes in that it is at the center of Shalev's substantive field of vision and that the book was awarded the 2000 Luebbert Award of the American Political Science Association for the Best Book in the Field of Comparative Politics. Generally, one could argue that the representative works of Boix, Iversen, and myself cited by Shalev as well as Hicks' award-winning book might be good indications of whether Shalev's critique of contemporary comparativists who practice MR is a fair one at that.[2]

For each of the works by Boix, Iversen, and myself, as well as for Hicks' 1999 book, I succinctly outline the content and method of the work. I assume many if not most readers are loosely familiar with these books and I focus on Shalev's charges of ''invisble cases,'' inattentiveness to complex

causal processes, and the absence of utilization of Shalev's preferred visually oriented case-variable method. As to the work of Carles Boix, I focus on his 1998 work (mentioned in passing by Shalev), *Political Parties, Growth and Equality*. Boix is most centrally concerned with whether or not social democratic parties can pursue distinct policy strategies (compared to parties of the Center and Right) to promote economic growth and material equality in the age of globalization. Through formal theoretical analysis, Boix hypothesizes that social democratic parties are likely to promote growth and equity through distinct interventionist supply-side policies. Parties of the Left combine active supply-side policies targeted to public infrastructure development with education, training, and related policies to simultaneously promote growth and equity in a world of economic internationalization; conservative parties in contrast prefer market allocation of investment and income. Boix initially evaluates these hypotheses through extensive MR analysis of cross-national data during one time period as well as PTSCS analysis (e.g., of 1960s–1990s annualized data from roughly 16 nations).

How does the comparative political analysis of Boix (1998) stack up when it comes to Shalev's critique? In terms of cases (developed democratic capitalist nation states), Boix not only includes multiple case references and synoptic illustrative case analysis during theory development and the interpretative stages of quantitative analysis, but the entire second portion of the book consists of rigorous analytic case studies, structured by the central theoretical questions at hand, of the formation and implementation of distinct policy strategies of the Spanish Socialist and British Conservative parties. What is particularly interesting is that Boix, during both theory development and quantitative analysis, repeatedly utilizes Shalev's own visually oriented case-variable method of graphic and tabular display of the positions of (virtually all) the developed democracies on two or three key political economic dimensions. For instance, nearly duplicating illustrative analysis in Shalev's article, Boix (Figure 2.3) maps individual developed democracies into a two dimensional space defined by educational attainment and unemployment rates; individual cases are also labeled as to their position on levels of the social wage (i.e., unemployment income replacement rates). The point of this analysis, in combination and dialogue with some simple regression analyses, is to provide a concrete, case-based initial empirical evaluation of key propositions about the influence of education and the social wage on economic performance (i.e., unemployment rates). Many further examples of Boix's (1998) utilization of Shalev's preferred method could be offered.[3]

As to analysis of complex causal processes, Boix judiciously uses carefully formulated empirical models and estimating equations to test key, formally derived hypotheses; interaction terms are used to assess contingent causal effects in quantitative analysis. The author further enriches the empirical assessment of theory with the aforementioned extensive case material on the Spanish and British cases. The causal basis of the strategic choices of Socialist and Conservative parties are teased out in the context of a finely grained analysis of the historic, institutional, and macroeconomic contexts of party choices in the post-OPEC Spanish and British political economies. The end product is a balanced, theoretically driven yet case-sensitive and multimethod analysis of an important set of questions in comparative political economy. Few if any readers, in my view, would vote to convict Boix of the crimes of MR purportedly so pervasive among the usual suspects.

The second work is the influential 1999 Cambridge University Press book by Torben Iversen (1999), *Contested Economic Institutions*. In this well-known analysis, Iversen seeks to understand which combinations of wage bargaining institutions and macroeconomic policy orientations and institutional infrastructures promote full employment and, in turn, how post-industrialization and globalization have altered the political and economic underpinnings of such successful configurations. Generally, the design and execution of Iversen's work is quite similar to Boix's (1998) research. Formal theorizing generates central hypotheses, which are evaluated with MR (especially PTSCS) analysis of 1960s to 1990s data from most of the developed capitalist democracies. Also similar to Boix's work, Iversen utilizes extensive synopses of case experiences and detailed, rigorous comparative case analysis of five key countries – Austria, (West) Germany, Norway, and especially Denmark and Sweden – to enrich quantitative analysis. Formal theory, quantitative empirical modeling and case analysis are in dialogue with each other throughout the book. And, perhaps most interesting with respect to Shalev's current critique and alternative, Iverson makes extensive use of the very tabular and graphic techniques recommended by Shalev to further enrich his analysis of core relationships (e.g., see Fig. 3.4 in which countries, differentiated by the non-accommodating or accommodating character of their monetary policy regime, are mapped in the two dimensional space of wage inequality and bargaining centralization). Overall, rich, complex comparative theory is generated and comprehensively assessed with multiple methods and a strong sense of the individual experiences of each of the developed capitalist democracies. In my view, the jury of readers would render yet another acquittal of a ususal suspect.

The third work to be considered is my own 2002 Cambridge University Press book, *Global Capital, Political Institutions, and Policy Change in Developed Welfare States*. In this research, I seek to systematically assess conventional globalization theory on the roles of economic internationalization in welfare state retrenchment as well as my alternative argument, namely, that domestic political and institutional contexts condition the policy impacts of rises in international capital mobility and trade openness. In one core chapter (Chapter 3), I provide the bulk of the quantitative analyses (principally PTSCS) of core theoretical propositions. The design and execution of the quantitative analysis is particularly sensitive to the non-linear and contingent nature of causal arguments. The bulk of the empirical portion of the book (Chapters 4–6) is an in-depth analysis of four Nordic political economies (Denmark, Finland, Norway, and Sweden), three continental welfare states (France, Italy, and Germany), and, in less depth, five Anglo liberal welfare states (Australia, Canada, New Zealand, the United Kingdom, and the United States). This qualitative analysis consciously recognizes the limits of large-N quantitative analysis to address causal sequence, collective actors' motivations, and rich historical and institutional contexts of strategic choices. Similar to Boix and Iversen's use of qualitative case studies, my use of case analysis is carefully structured to address central theoretical questions and to draw on the strengths of comparative case analysis (including process tracing within cases) in order to address the aforementioned shortfalls of quantitative analysis and to engage in a consistent dialogue with quantitative findings.[4] Overall, for these reasons, I would hope I would join the ranks of those usual suspects acquitted by the jury of readers.

As a final representative work of contemporary comparative research, I turn to the winner of the American Political Science Association's 2000 Luebbert Award for the Best Book in the Field of Comparative Politics: Alex Hicks' 1999, *Social Democracy and Welfare Capitalism*. In this book, Hicks seeks to explain the early adoption and later consolidation of the basic income maintenance programs of the modern welfare state. He also seeks to advance our knowledge about the determinants of post-World War II expansion and, ultimately, retrenchment of these core programs of social protection. While conscious of the continuing controversies over determinants of origins, expansions and contractions of 20th century income transfer policies, Hicks offers a set of core theoretical arguments that combines the insights from power resources and political institutional theories of welfare state development.

To test core and alternative theoretical explanations of welfare state origins, expansions and contractions, Hicks innovatively combines Ragin's

(1987) Qualitative Comparative Analysis (aka Boolean algebra), carefully designed PTSCS regression analysis and a battery of synoptic case analyses from the last decades of the 19th century to the last decades of the 20th century. Hicks' utilization of Boolean analysis of the conditions necessary and sufficient for early (circa 1920) and later (1930s/1940s) consolidations (i.e., comprehensive adoption) of income maintenance programs is bolstered by a comprehensive use of country names (e.g., in the truth tables of Boolean analysis) and constant reference to causal sequences in individual countries. So too is the MR (especially PTSCS) analysis of temporal and cross-sectional variations in welfare expansion and retrenchment. As in the historical analysis of determinants of welfare state consolidation, the quantitative analysis of expansion and contraction is especially sensitive to the adequacy of tests of complex non-linear and contingent causal effects. In fact, a major thrust of Hicks' analysis of 20th century consolidation of income maintenance programs is to theorize and assess the presence of multiple configurative paths to welfare state development. As with the works of Boix, Iversen, and myself, Hicks' impressive analysis of the 20th century welfare state development in democratic capitalism, which extensively employs MR analysis, arguably does a notably better job in making cases visible, combining multiple methods, and adequately assessing complex comparative theory on an important set of substantive questions than Shalev's critique of this body of quantitatively oriented work would predict. As in the world of film, most if not all of the usual suspects are innocent of purported crimes of which they are charged, or at least they are innocent of the felonies that may have been laid at their door.[5]

OTHER ISSUES

Shalev raises a number of additional issues with the use of MR in comparative political and social research. One set of issues consists of those connected with statistical control and inference as well as the use of apparent populations. This set of concerns, which encompasses much of quantitative social science, is far beyond the scope of this response. There are large and sophisticated literatures on these topics and Shalev, himself, barely scratches the surface. On the other hand, Shalev makes some rather specific claims about further problems with the use of MR generally, and PTSCS analysis specifically, that I would like to address.

First, Shalev argues that while MR is useful in disciplines such as economics where researchers are interested in the marginal effects, let us say, of

prices on economic output, political scientists and sociologists analysis are commonly interested in the impact of the presence or absence, let us say, of corporatism on economic growth. MR is purportedly less appropriate for these comparative political analysts. I find this criticism to be without merit. MR is perfectly suited for precisely estimating effects (i.e., mean differences) of a theoretically relevant variable such as the presence or absence of corporatism (and loads of other categorical variables) on continuous variables such as economic growth rates. Moreover, a particular family of MR-type estimators, event history (duration or hazard) models, are especially useful in estimating the determinants of categorical variables (for instance, see Hicks and Zorn's (2005) expert utilization of Cox hazard models of repeated events to assess the causes of the occurrence of welfare retrenchment).

Second, Shalev argues that the aforementioned problems of MR are magnified by conceptual ambiguity and imprecision in measurement. On the first part of this observation, I see no reason why quantitatively oriented political analysts should worry anymore about conceptual confusion and contention than more qualitatively oriented researchers. The dangers of fuzzy concepts for good research seem universal. As to the question of imprecise measurement, I would argue in response to Shalev that students of comparative politics now have access to many more databases and highly improved measures in many areas of research relative to comparativists in the late 1970s and early 1980s. To use an area for illustration that Shalev often invokes, the ability of researchers to measure across countries and time the degree to which interest representation and, in turn, national policy making is corporatist has vastly improved. Reliable and valid publicly available measures of employer and union organization (e.g., density, centralization) and incorporation of peak associations into national policy making forums is available in several databases, most notably, Golden, Wallerstein, and Lange (No date) and Traxler, Blaschke, and Kittle (2001).

Finally, Shalev is particularly critical of the proponents of PTSCS analysis. Beyond those problems already identified for MR, Shalev argues that researchers who use PTSCS analysis are commonly insensitive to the general question of whether one should pool cross sections of time series (e.g., the problem of potential parameter heterogeneity is ignored), and to the question of whether causal dynamics are different across temporal and cross-sectional dimensions of causal factors. In addition, Shalev questions whether the technical expertise required for PTSCS analysis generally, and for adjudicating contemporary debates and assessing technical advances specifically, is worth the investment for researchers given questionable payoffs. On the first point, this is indeed an important admonition for

researchers that is too often ignored. On the other hand, I would point out that it is now increasingly common for researchers to offer tests for parameter homogeneity in dynamic relationships across cross-sections (e.g., Swank & Steinmo, 2002) or test for theoretically predicted dynamic parameter heterogeneity (e.g., Swank, 2002, 2006). Researchers also increasingly test for cross-national parameter homogeneity at different time points and do so for explicit theoretical reasons (e.g., Kwon & Pontusson, 2002). Few if any of the scholars mentioned above are insensitive to this set of issues.

As to the last point, I am not convinced that technical complexity, alone, should deter researchers from learning and employing complex methods when substantive and theoretical questions suggest doing so. Many newer, increasingly utilized quantitative methods have generated important new findings (e.g., multilevel modeling) and are technically complex; few researchers would consider abandoning them.[6] Relatedly, Shalev cites eminent econometrician G. S. Maddala (1998) to cast doubt on technical advances in PTSCS, namely, the panel-correct standard error approach for Ordinary Least Squares regression developed by Beck and Katz (1995, 1996). It is the case, however, that Shalev's quote of Maddala to the effect that Beck and Katz procedures are "not, strickly speaking, correct" is misleading. What Maddala (1998, p. 61) says is "Some of the statements made in the Beck–Katz articles are not, strictly speaking correct, but these are minor issues and do not affect their analysis … The idea of using OLS with panel corrected standard errors is fine …" What Maddala is concerned about is a classic problem of the use of lagged dependent variable (which Beck and Katz recommend to explicitly model temporal dynamics) in the presence of autocorrelation. Maddala likes instrumental variables as the solution, Beck and Katz do not and will take the risk of inconsistency of the OLS estimators over the uncertainty of generating good instrumental variables. Overall, these technical complexities and debates characterize most areas of quantitative social science and, in my view, should not be regarded in anyway as a justification for abandoning techniques appropriate for many substantively and theoretically important questions.

CONCLUDING THOUGHTS

To sum up, my own view is that many of Shalev's admonitions to comparative political analysts who utilize on MR and PTSCS analysis are well worth taking to heart. It is certainly true that too often we lose track of cases

and oversimplify tests of complex theories. In addition, the challenges and technical difficulties of PTSCS are often minimized or ignored. On the other hand, most of the scholars cited by Shalev as leading practioners of MR and PTSCS analysis think about these problems and offer relatively effective methodological designs to advance research. As I hope I have demonstrated, the books (and many of the articles) of the usual suspects that Shalev cites combine sophisticated MR and PTSCS analysis with complementary quantitative and qualitative techniques in an effort to produce comprehensive assessments of the core substantive and theoretical questions at hand; several authors actually make use of Shalev's preferred technique of bringing cases with proper names to the center of the stage of analysis. This technique, as noted above, is most appropriate to exploratory analysis during theory development or to initial tests of simple hypotheses (and this is how the aforementioned authors use it). Overall, while all quantitative comparativists would benefit from a careful reading of Shalev's article, many contemporary scholars do a much, much better job than Shalev admits in designing and executing research.

NOTES

1. As such, Shalev's central point here both invokes and goes beyond the quantitative (variable-oriented) versus qualititative (case-oriented) methods debate over the best approach to testing causal theories in comparative politics. For excellent introductory overviews and discussions, see contributions to the 1995 *American Political Science Review* symposium, King, Keohane, and Verba (1994), Ragin (1987), and the excellent synoptic discussion in Chapter 2 of Rueschemeyer, Stephens, and Stephens (1992).

2. I do not include additional works of Franzese or Garrett (or others), or a defense of the focal works critiqued by Shalev, because length considerations suggest a discussion of a more limited set of representative works. Franzese and Garrett are, of course, more than capable of effectively responding to the crimes laid at their doorstep.

3. It should be noted that Shalev's alternative case-variable method is well-suited to the assessment of the plausibility of initial theoretical suppositions, or to provide exploratory tests of initial hypotheses. The problem with making this technique central to the analytical framework is that once a researcher is concerned with tests of hypotheses much beyond three variables, the technique becomes unmanageable. For instance, in Shalev's Chart 1, we map countries in a tree diagram by country size, strong or weak left parties, and the presence of the Ghent system. If we had one or two further dimensions (say economic openness, or openness and industrial concentration), the simple graphical exposition would resemble an organizational chart for the financial accounts of Enron.

4. For a formal schema for "nesting" case analysis in large-N quantitative work, and engaging the two methods in productive dialogue, see Evan Leiberman's (2005) important recent contribution in the *American Political Science Review*. While Shalev seems skeptical about the potential power of nested analysis, or more broadly "triangulation" (Ragin, 1987), I am much more impressed by its potential for both assessing and improving general theory as well as comprehensively understanding cases.

5. Shalev implicitly seems to recognize that the authors of books have more capacity to offset problems of MR analysis; the quantitative comparative analysts writing for journals (because of space considerations if nothing else) seem to be less able to make cases "visible," to adequately test complex theory and so forth. A brief response to this notion is simply to point readers to increasing numbers of articles in leading journals such as the *American Political Science Review* and *World Politics* that balance sophisticated uses of MR with relatively developed case-oriented material. For just two examples, see Iversen and Wren's (1998) combination of PTSCS and synoptic cases analysis of the "trilemmas" of the service economy in the *World Politics,* and Martin and Swank's (2004) multi-level, multi-method analysis of employers' preferences for social policy interventions in the *APSR*.

6. For an overview of method and applications of multilevel modeling, see among others Goldstein (2003) and Luke (2004).

REFERENCES

Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, *89*(3), 634–647.

Beck, N., & Katz, J. N. (1996). Nuisance vs. substance: Specifying and estimating time-series cross-section models. *Political Analysis*, *6*, 1–36.

Boix, C. (1998). *Political parties, growth and equality: Conservative and social democratic economic strategies in the world economy*. New York: Cambridge University Press.

Garrett, G. (1998). *Partisan politics in the global economy*. New York: Cambridge University Press.

Golden, M., Wallerstein, M., & Lange, P. (No date). Union centralization among advanced industrial societies. Electronic database available at www.shelly.polisci.ucla.edu/data

Goldstein, Harvey. (2003). *Multilevel statistical models*. New York: Oxford University Press.

Hall, P. A., & Franzese, R. J. (1998). Mixed signals: Central bank independence, coordinated wage bargaining, and European Monetary Union. *International Organization*, *52*(3), 505–535.

Hicks, A. (1999). *Social democracy and welfare capitalism: A century of income security politics*. Ithaca, NY: Cornell University Press.

Hicks, A., & Zorn, C. (2005). Economic globalization, the macroeconomy, and reversals of welfare: Expansion in affluent democracies, 1978–1994. *International Organization*, *59*(Summer), 631–662.

Iversen, T. (1999). *Contested economic institutions: The politics of macroeconomics and wage bargaining in advanced democracies*. New York: Cambridge University Press.

Iversen, T., & Wren, A. (1998). Equality, employment, and budgetary constraint: The trilemma of the service economy. *World Politics*, *50*(4), 507–546.

King, G., Keohane, R., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton: Princeton University Press.

Kwon, H. Y., & Pontusson, J. (2002). Welfare spending in OECD countries revisited: Has the salience of partisanship really declined? Paper presented at the annual meeting of the American Political Science Association, Boston, MA, August 29–September 1.

Lieberman, E. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, *99*(3), 435–452.

Luke, D. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Maddala, G. S. (1998). Recent developments in dynamic econometric modeling: A personal viewpoint. *Political Analysis*, *7*, 59–87.

Martin, C. J., & Swank, D. (2004). Does the organization of capital matter? Employers and active labor market policy at national and firm levels. *American Political Science Review*, *98*(4), 593–612.

Przeworski, A., & Tuene, H. (1970). *The logic of comparative social inquiry*. New York: Wiley.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Rueschemeyer, D., Stephens, E. H., & Stephens, J. (1992). *Capitalist development and democracy*. Chicago: University of Chicago Press.

Shalev, M. (1983). The social democratic model and beyond: Two 'generations' of comparative research on the welfare state. *Comparative Social Research*, *6*, 315–351.

Shalev, Michael (Ed.) (1996). *The privatization of social policy? Occupational welfare and the welfare state in America, Scandinavia, and Japan*. London: Macmillan.

Swank, D. (2002). *Global capital, political institutions, and policy change in developed welfare states*. New York: Cambridge University Press.

Swank, D. (2006). Tax policy in an era of internationalization: Explaining the spread of neoliberalism. *International Organization*, *60*, 847–882.

Swank, D., & Steinmo, S. (2002). The new political economy of taxation in advanced capitalist democracies. *American Journal of Political Science*, *46*(3), 642–655.

Traxler, F., Blaschke, S., & Kittle, B. (2001). *National labour relations in international markets*. New York: Oxford University Press.

NEW METHODS FOR COMPARATIVE RESEARCH?

Claude Rubinson and Charles C. Ragin

INTRODUCTION

Shalev's (2007) critique of the use of multiple regression in comparative research brings together and synthesizes a variety of previous critiques, ranging from those focusing on foundational issues (e.g., the persistent problem of limited diversity), to estimation issues (e.g., the unrealistic assumption of correct model specification), to narrow technical issues (e.g., the difficulty of deriving valid standard errors for regression coefficients in pooled cross-sectional time-series models). Broadly speaking, these concerns can be described as epistemological, theoretical, and methodological, respectively. While the distinctions among these three are not always clear-cut, the tripartite scheme provides a useful way to map the different kinds of critiques that may be directed at the use of regression analysis in comparative research. In the first half of this essay we build upon Shalev's discussion to clarify the conditions under which regression analysis may be epistemologically, theoretically, or methodologically inappropriate for comparative research. Our goal is to situate Shalev's specific critiques of the use of multiple regression in comparative work within the context of social research in general.

In the second half of this essay, we focus on Shalev's proposed solutions. We commend Shalev for offering constructive solutions to the problems he

Capitalisms Compared
Comparative Social Research, Volume 24, 373–389

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24013-4

raises. Too often, critiques end without solutions being offered, demoralizing those who are committed to empirical research. However, we feel that Shalev has overlooked the fact that the issues he raises are addressed more completely and fully in the growing literature on Qualitative Comparative Analysis (hereafter, QCA) and fuzzy-set analysis.[1] We argue that QCA and related methods both encompass and extend Shalev's proposed solutions and provide a strong foundation for systematic, case-oriented comparative research.

THE CRITIQUE OF MULTIPLE REGRESSION

One of the central themes of Shalev's critique of multiple regression is its incongruence with case-oriented analysis. Case-oriented approaches might be preferred for several reasons, in addition to the simple fact that comparativists tend to study small *N*s. For example, case-oriented methods are better suited for the types of questions that comparative researchers typically ask. Unlike multiple regression, case-oriented techniques such as QCA and fuzzy-set analysis are specifically designed to address questions about necessary or sufficient conditions that often motivate comparative research. Furthermore, case-oriented techniques can be used to address causal complexity. Finally, case-oriented methods such as QCA, fuzzy-set analysis, and those recommended by Shalev are more closely aligned with the epistemological orientations held by many comparative researchers. This orientation identifies the case – rather than the variable – as the fundamental unit of interest to social researchers.

The Epistemological Critique

The epistemological critique of the use of regression analysis in comparative research is straightforward: the method results in unproductive representations of social phenomena. Social research is best described as the construction of scientific representations of social life (Ragin, 1994). To the extent that applications of regression analysis result in representations that do not resonate with researchers' understandings, the method is called into question. The primary reason that these representations are found lacking is that case-oriented comparative researchers keep cases, not the net effects of variables, at the forefront of their analyses. It is not that case-oriented researchers dismiss variables, but rather that they perceive that it is not

variables but cases that have relationships with one another. The variable-oriented researcher shows, for example, that poverty is correlated with crime and that economic development is correlated with democracy. The case-oriented researcher observes that criminals tend to be poor, especially the ones that get caught, and that economically developed countries tend to be democratic. Although subtle, this distinction entails fundamentally different views of social phenomena. Where variable-oriented researchers view the social world as a manifestation of the myriad relationships among variables, case-oriented researchers see many different kinds or sets of cases.

For the case-oriented researcher, the problem with regression analysis is that it veils cases. Regression analysis describes the relationships between independent and dependent variables which, from the vantage of case-oriented research, is a limited and fragmented picture of reality. Note that this is not a technical critique of the method's capabilities but, rather, a reaction to the world-view inherent in the method. In regression analysis, cases do not constitute anything in and of themselves; they are merely carriers of information about the relationships among variables.[2] The case-oriented researcher, however, requires methods that maintain the constitution and integrity of the cases under observation. Stated simply, regression is incapable of doing this and, therefore, is an inadequate platform for conducting case-oriented research. Although strongly worded, such a statement should not be controversial. It is not a deficiency of regression that it fails to meet the needs of case-oriented researchers but simply reflects the fact that it meets the needs of variable-oriented researchers so well. Its strength is also its weakness. While comparative researchers tend to be case-oriented, this coupling is not mandatory. Regression analysis is perfectly suitable in the hands comparative researchers who see cases as instances of relationships between variables.

The Theoretical Critique

Regression analysis is best at answering theoretically framed questions about the net effects of competing independent variables on a dependent variable (Ragin, 2006). In a multiple regression model, it is assumed that any single variable is sufficient by itself for achieving an impact on the outcome and that no variable is necessary. Thus, regression analysis is not well-suited for the analysis of causal complexity. As Shalev (2007) notes: "The results will be ambiguous because they will be unable to distinguish between additive effects, conditional effects, and multiple causal pathways." These limitations are especially problematic when comparative researchers attempt

to use regression to answer theoretical questions that it is not designed to answer. Unfortunately, such attempts are far too common. Shalev provides an example in his review of Rothstein's analysis of union membership. Rothstein hypothesizes, in essence, that the presence of the Ghent system is a necessary condition for high levels of unionization. However, he attempts to test this hypothesis by regressing percent union on presence of Ghent, strength of left government, and potential union membership. While a statistically insignificant effect for Ghent might undermine the hypothesis, Shalev (2007) points out that "the regression could not make a positive case for Rothstein's argument." Regression coefficients report the partialed effects of each independent variable on the dependent variable. There is no basis in this type of analysis for privileging the effect of Ghent as a necessary condition for high levels of unionization.

Sometimes it is difficult to correctly identify the method most appropriate for answering a given theoretical question. Consider another study that Shalev reviews, Hall and Franzese's (1998) investigation of the effect of central bank independence on unemployment rates. Shalev (2007) identifies two central hypotheses:

> In nations where wage coordination is high, an increase in the independence of the central bank is associated with a very small increase in the rate of unemployment.
>
> Where wage coordination is low, however, an increase in the independence of the central bank is associated with a substantial increase in the rate of unemployment.

These hypotheses encompass both case-oriented and variable-oriented questions. On the one hand, Hall and Franzese are asking about nations with high versus low levels of wage coordination, a question that is case-oriented in nature. On the other hand, they are asking about the association between central bank independence and unemployment rates, a question that is variable-oriented in nature. This disjuncture results from Hall and Franzese's recognition of the contextual effect of wage coordination. In regression analysis, contextual effects are operationalized through interaction terms. Indeed, from a variable-oriented perspective, the two hypotheses are one. It would be more precise to test the single hypothesis that "as wage coordination decreases, the strength of the positive association between central bank independence and the unemployment rate increases," and, in fact, this is the hypothesis that they test, using an interaction term (Hall & Franzese, 1998, p. 519). Dichotomizing the wage coordination measure permits Hall and Franzese to interpret their results as applying to nations with high versus low levels of wage coordination. Strictly speaking, however, their results only describe relationships among variables.

Case-oriented techniques, by contrast, downplay the relationships among variables and instead emphasize how the countries in the data set constitute 18 combinations of the variables of interest. Shalev (2007) presents a rudimentary configurational view in his Chart 2. Examining Chart 2, what stands out is not the relationship between unemployment rates, central bank independence, and wage coordination but that, with the exception of Australia, all of the states with relatively low levels of wage coordination also have relatively high rates of unemployment. Furthermore, again with the exception of Australia, all countries with relatively low rates of unemployment have relatively high levels of wage coordination. Setting Australia aside, these results indicate that having high levels of wage coordination is a necessary condition for low levels of unemployment and, correspondingly, that low levels of wage coordination are sufficient for high rates of unemployment. Such conclusions naturally lead to a focus on the anomalous case of Australia: why is Australia's unemployment rate so much lower than one would otherwise expect? How does Australia differ from France, which has similar scores on wage coordination and central bank independence? Similar questions are raised regarding Germany, Denmark, Finland, and Norway: these countries all have high levels of wage coordination; why do they also have high unemployment rates? In short, the use of case-oriented techniques disposes the researcher to focus on the characteristics of the cases under investigation rather than the relationships among the variables.

The Methodological Critique

The most well-known critique of the use of regression analysis in comparative research is methodological, the so-called "small-*N* problem." Since comparative researchers generally study only a handful of cases, sophisticated regression techniques quickly exhaust available degrees of freedom. Comparative researchers often emphasize that case-oriented analytic techniques can better address issues of causal complexity than variable-oriented analytic techniques (see, e.g., Rueschemeyer & Stephens, 1997); however, Shalev (2007) correctly points out that:

> The problem is not that MR does not have or could not invent technologies for dealing with such complexities. Nonlinear functional forms, interaction effects and (in time-series analysis) complex lag structures immediately come to mind. The point is that because such techniques are either difficult to employ or impose a steep statistical penalty due to the "small n problem," they are rarely or insufficiently used.

To the extent that regression analysis is ill-suited for comparative analysis due to limited degrees of freedom, then, there are three possible solutions. One is to increase the number of observations available for analysis through additional data collection or a reformulation of the research question (see King, Keohane, & Verba, 1994). A second is to use a technique such as factor analysis or metric scaling to reduce the dimensionality of the model's vector space. The third is to turn to case-oriented analytic techniques that are not directly constrained by considerations of degrees of freedom.

Pooled cross-sectional designs are a common example of the first option. Shalev (2007) cogently describes the methodological problems that can accompany this technique, noting that such designs reflect the limitations of both cross-sectional and longitudinal studies: "pooled designs are the worst of both worlds." Rather than artificially increasing the number of observations by using pooled designs, Shalev recommends reducing the number of causal variables, as illustrated in his discussion of Esping-Andersen's analysis of welfare regimes. Reduction of vector space dimensionality is a long standing recommendation (Berg-Schlosser & De Meur, 1997). However, comparative researchers should be cognizant of the limitations of data reduction techniques such as factor analysis. Specifically, in order to create an index from several causal conditions factor analysis rescales correlated conditions and then sums their scores. The assumption, in effect, is that the different conditions that go into an index are partially substitutable such that any one condition may compensate for any other condition. Thus, factor analysis, like regression analysis, masks causal complexity and veils case specificity. The application of vector reduction techniques such as factor analysis and metric scaling demands theoretical justification; they should not be used as easy, technical solutions to problems associated with limited degrees of freedom. This leaves the third solution to the degrees-of-freedom shortage, which is to use case-oriented methods such as QCA, fuzzy-set analysis, or the methods Shalev proposes. While attractive, this path is not necessarily a panacea, for it carries with it a world-view that emphasizes similarities and differences among cases, not relationships among variables.

Discussion

Shalev reviews a number of arguments as to why regression analysis is often inappropriate for comparative research. The methodological critique is most commonly made, and it is easiest to understand: Comparative researchers frequently study a limited number of cases; under such conditions, the

assumptions of regression analysis are very difficult to meet. The theoretical and epistemological critiques are less frequently made, but are more important because they are directed at more fundamental concerns. The theoretical critique observes that regression analysis is best-suited for answering only certain types of questions regarding relationships among variables. Frequently, however, these are not the types of questions that interest comparative researchers. The epistemological critique observes that regression analysis carries with it a variable-oriented world-view that is incongruent with the case-oriented world-view, which is common, though certainly not universal, among comparative researchers. Taken together, these three critiques point to the need for methods that meet the specific requirements – epistemological, theoretical, and methodological – of case-oriented comparative researchers.

Shalev (2007) writes "MR remains by far the predominant mode of numerical data analysis and most of its critics sees qualitative analysis (whether formal or not) as the only real alternative. This paper seeks to promote a third way." We share Shalev's concern. The conventional division between quantitative and qualitative research techniques tends to hinder – rather than benefit – social research by implicitly limiting case-oriented researchers to qualitative tools and variable-oriented researchers to quantitative tools.

ALTERNATIVES TO REGRESSION ANALYSIS

Shalev largely neglects QCA and fuzzy-set analysis in favor of his proposed methods. In the second half of this essay, we contest this oversight. QCA and fuzzy-set analysis are, in fact, more elaborate and refined versions of the methods Shalev recommends.

Shalev's Critique of QCA

Shalev (2007) suggests that, with regard to comparative research, QCA and fuzzy-set analysis suffer some of the same shortcomings as regression analysis:

> Ragin's methods are not "qualitative" in the sense of relying on the interpretive skills of analysts wading knee-deep in thick description. If anything, as Griffin and Ragin (1994, p. 10) have insisted, QCA is more like MR: both apply rules that are independent of the researcher, and both treat cases as "discrete, multiple instances of more general phenomena."

It is true that QCA shares a few characteristics with regression analysis. Both are formal methods and, as such, are characterized by the application of procedures that are independent of the researcher. But whereas regression is an application of linear algebra rooted in matrix theory (Marcus & Minc, 1988), QCA and fuzzy-set analysis are applications of set theory (Whitesitt, 1995). Set theory is used, very simply, to formalize the logic of comparative analysis, as practiced by case-oriented researchers. The primary goal of the formal procedures implemented in QCA is to prevent researchers from drawing illogical conclusions from comparative evidence, especially when the *N* of cases is more than a handful. Consider the truth table, which forms the foundation of both QCA and fuzzy-set analysis. Superficially, it appears similar to a conventional data set in that it utilizes a "cases-by-variables" format. But the rows of a truth table are not observations as they are in a conventional data set. Rather, each row represents a logically possible combination of causal conditions.[3] It is up to the researcher to determine which of these combinations map onto real-world cases. Frequently, the process of mapping the causal configurations onto real-world cases will prompt researchers to revisit and revise their classification schema, based on in-depth analysis of cases. From the same Griffin and Ragin (1994, p. 10) article:

> To resolve the contradictions[4] in their data, the authors intensively reexamined both the configurations producing contradictory outcomes and the cases in those configurations. They searched for errors in their original classification, thought more deeply about whether their dichotomous measure of labor management practices was too crude, looked anew at their interviews with personnel managers, and strategically compared mills in contradictory configurations with mills in configurations free of contradictory outcomes. All of this interpretive work on classification – really, on the meaning of their outcome factor and its applicability to several of their cases – was but a prelude to their explanatory analysis.

It is through the construction, revision, and refinement of truth tables that QCA and fuzzy-set analysis rely "on the interpretive skills of analysts wading knee-deep in thick description" (see also Ragin & Rihoux, 2004 and the commentary and exchanges it generated in a special issue of *Qualitative Methods* devoted to QCA; see also the three-way exchange on QCA versus regression analysis published in *Studies in Comparative International Development*: Achen, 2005; Seawright, 2005; Ragin, 2005). As with the alternative techniques that Shalev proposes, effective application of QCA and fuzzy-set analysis depends directly upon the researcher's substantive knowledge of the cases under investigation.

Shalev's Tabular Technique and QCA

Shalev's first proposed alternative to regression analysis makes use of tabular techniques. Reanalyzing Rothstein's model of union membership, Shalev clusters countries according to the different combinations of causal conditions that they exhibit. This technique has two benefits. First, it provides a direct test of a hypothesis that conventional regression analysis could not provide, that is, Rothstein's hypothesis that "the highest levels of unionization have been reached only in countries where [the Ghent system] is in place" (Shalev, 2007). Shalev's tabular technique clearly identifies the combinations of conditions linked to high rates of unionization, confirming Rothstein's hypothesis. The second benefit of Shalev's (2007) tabular technique is that, in moving the individual countries to the forefront of the investigation, it also "point[s] the interested researcher to the most fertile questions for selective case comparisons." Shalev's tabular technique leads naturally to the investigation of similarities and differences among cases in a way that analysis of regression residuals does not.

The insight underlying Shalev's tabular technique – that it is combinations of conditions that matter – is the same insight that underlies QCA (Ragin, 1987). QCA is built around the analysis of a "truth table" that delineates the various combinations of conditions linked to the presence/absence of an outcome. As with Shalev's tabular technique, QCA permits the investigation of necessary and sufficient conditions and keeps the individual cases at the forefront of the analysis. QCA has a number of advantages over Shalev's tabular technique. The most obvious advantage is conciseness. As the number of causal conditions increases, tabular analysis quickly becomes unmanageable. A truth table, however, can accommodate a large number of causal conditions. Furthermore, the existence of truth table reduction algorithms provided in software packages such as fs/QCA simplifies the accompanying analysis. Shalev (2007) notes that determining the proper setup of the table "requires some forethought" due to the complexity of the analysis. Researchers using QCA need only define and measure the relevant causal conditions. The method's algorithm identifies the causal configuration(s) linked to the outcome under investigation.

Shalev emphasizes that an advantage of his tabular technique over regression analysis is that it places cases in the foreground of the analysis. QCA does this as well. Using either technique, researchers will not simply determine that high union membership is present in countries with a combination of small size, medium or strong left parties, and the presence of the Ghent system; they can explore additional avenues of investigation: "In

particular, it must be questioned whether the Ghent system alone can explain the very large differences in density between otherwise well-matched countries: Belgium versus the Netherlands, and Sweden and Denmark versus Norway'' (Shalev, 2007). But QCA has an additional advantage in that it forces investigators to resolve contradictions (cases with similar causal configurations that produce divergent outcomes). For example, Shalev simply ignores the contradiction between Switzerland and Ireland: both are small countries with weak left parties and no Ghent system, but union membership is weak in Switzerland and strong in Ireland. QCA forces the researcher to confront such contradictions and decide how to deal with them (see Ragin & Rihoux, 2004; Ragin, 2005). In this way, QCA structures a close interaction between researcher and cases.

Another advantage of QCA over Shalev's tabular technique regards the analysis of counterfactual cases. Shalev (2007) correctly points out that the social world is characterized by limited diversity:

> In cross-national quantitative research the situation is very different [than in survey research]. We often analyze the entire universe of cases, and if not it is usually because of lack of data rather than sampling considerations. For the most part then, *if a particular configuration of attributes does not exist in a cross-national data set, it does not exist at all* (emphasis in original).

Shalev (2007) raises the issue of limited diversity within his critique of regression analysis and comments that ''it cannot be denied that one of the tests of a useful causal model is that it be capable of answering counterfactual questions.'' QCA provides just such a capability, for the analysis of limited diversity is a long-standing focus of the approach (see Ragin, 1987). As detailed in Ragin and Sonnett (2004), QCA includes tools especially designed for the analysis of ''remainder'' causal combinations (that is, logically possible combinations of conditions that lack empirical instances). Such analyses formalize the thought experiments proposed by Weber (1905) by treating the remainder combinations as counterfactual cases. By incorporating the analysis of remainders into QCA, the researcher can better assess the causal role that specific conditions play in bringing about the outcome in question.

Shalev's Three-Dimensional Plots and Fuzzy-Set Analysis

A second technique that Shalev utilizes in his reanalyses of Hall and Franzese (1998) and Garrett (1998) is that of three-dimensional scatterplots, with the third dimension represented as proportionately sized ''bubbles.''

These scatterplots can be seen as nascent fuzzy-sets. Reanalyzing Garrett's data, Shalev (Chart 4) clusters countries according to their degree of capital restriction and trade openness. In set-theoretic terms, these clusters represent subsets. Shalev identifies three subsets: a set of countries with low levels of both capital restriction and trade openness,[5] a set of countries with high levels of capital restriction and middling levels of trade openness, and a set of countries with low levels of capital restriction and high levels of trade openness. Shalev observes that the countries in Garrett's analysis exhibit limited diversity: there are specific regions of the property space that are void of cases. In particular, he notes that there are no cases for the combinations of (a) high left-labor power with low capital restriction or (b) low left-labor power with high trade openness. As Shalev notes, Garrett conducted tabular analyses that included precisely these combinations.

Shalev's critique of Garrett reflects the distinctive manner by which comparative researchers often measure their variables. Garrett (1998, p. 84) employs relative measures of left-labor power, capital restriction, and trade openness: "Low (high) levels of trade and capital mobility refer to the 20th (80th) percentile scores in the sample. Low (high) levels of left-labor power refer to the 20th (80th) percentile scores on left-labor power index." For Garrett, scores are low or high relative to the median; indeed, it would be more accurate to use the labels "lower" and "higher" to reflect this operationalization. For comparative researchers, adjectives such as "low" and "high" generally describe qualitative conditions measured against a defined standard. Consider the set of Western European countries. Although there is certainly variation in GDP per capita among these countries, all may reasonably be considered rich – depending upon how the researcher defines "rich." Case-oriented methods do not evaluate variation in the same way that variable-oriented methods do. In case-oriented research, it is the substantive meaning of the scores that is most important; scores must be calibrated relative to some standard, not simply relative to a measure of central tendency. In qualitative work, measurement is an interpretive process, based on the researcher's theoretical and substantive knowledge.

From this viewpoint, Shalev makes the same general error as Garrett. When constructing his scatterplots, Shalev does not consider the substantive meaning of the various scores but simply accepts them at face value. Examining Chart 4, for example, Shalev (2007) identifies a cluster of social democracies consisting of Sweden, Denmark, Austria, and Norway. Finland is not included in this cluster, presumably due to its higher level of capital restriction. But does Finland's exclusion make sense? Garrett's measure of capital restriction is simply a count of four types of government restrictions

on capital mobility. Excluding Finland from the social-democratic cluster assumes that the raw number of restrictions matters. It is not clear that this is true. For example, in an investigation of foreign exchange market turbulence, Eichengreen, Andrew, and Wyplosz (1995) operationalize capital restriction simply as a dummy variable indicating the presence or absence of *any* capital controls. This operationalization indicates that the researchers believe that capital restrictions are substitutable for one another and, furthermore, that their effects are not necessarily additive. It may be that the difference between Finland's level of capital restriction and those of Sweden, Denmark, Austria, and Norway amounts to nothing more than irrelevant variation, and Finland should be included in the social-democratic cluster.

Shalev's reanalysis of Hall and Franzese (1998, Chart 2) displays the same shortcoming. Hall and Franzese do not justify the dichotomization of their institutional variables, and Shalev appropriately criticizes this oversight. But it is by no means clear that Shalev's strategy of disaggregating the variables is better. Both actions are arbitrary. Shalev's approach assumes that the data – and the variation in the data – speak for themselves. But researchers must always interpret scores and evaluate what they mean. Because Shalev does not find a pattern in Chart 2, he concludes that Hall and Franzese's findings are an artifact of their dichotomizing their measures. But it is also possible that Shalev's lack of findings is a result of his failure to properly calibrate his measures, using theoretical and substantive knowledge to guide the interpretation of scores.

Fuzzy-set analysis forces researchers to calibrate their measures carefully; the resulting fuzzy membership scores must be substantively meaningful. In fuzzy-set analysis, scores indicate the degree of membership of cases in a given set. A country may be classified as fully, partially, or not belonging to the set of countries with, for example, high left-labor power or high capital restriction.[6] After the researcher calibrates membership scores, formal fuzzy-set techniques can be applied to determine the subset relationships that exist among the cases. Shalev derives his clusters using *ad hoc* procedures; fuzzy-set analysis applies set theory to the same end, based on the researcher's interpretation of each case's degree of membership in the relevant sets.

A further difference between Shalev's clustering technique and fuzzy-set analysis concerns the role that the derived subsets play in the subsequent analysis. Shalev's clusters are primarily descriptive. By keeping the cases in the foreground of his reanalysis of Garrett, Shalev's technique permits him to distinguish a social-democratic subset, an autarchic subset, and a

small-state subset. In fuzzy-set analysis, however, subsets are not merely descriptive but also provide a foundation for the analysis of causality. Through the application of set theory and fuzzy algebra, fuzzy-set analysis provides formal methods for evaluating necessary and sufficient conditions.

Fuzzy-set analysis is a variant of QCA; as such, it shares QCA's advantages. Like QCA, fuzzy-set analysis can accommodate a substantial number of causal conditions. Shalev's scatterplots are useful, but it is difficult to visualize a plot with more than three dimensions. Reflecting the fact that his technique grants explanatory primacy to just two dimensions at a time, Shalev is unable to incorporate level of left-labor power into his clusters. Fuzzy-set analysis, on the other hand, locates each case's position in a vector space with a much larger number of dimensions. (In practice, most researchers use from four to nine.) Also like QCA, fuzzy-set analysis makes use of truth tables and provides formal techniques for identifying the various causal configurations linked to the outcome under investigation and for the analysis of counterfactuals.

Discussion

Case-oriented comparative researchers seek explanation by exploring the similarities and differences among cases. The problem with variable-oriented techniques such as multiple regression is that they render cases invisible. At the heart of Shalev's tabular and scatterplot techniques is an attempt to bring cases to the foreground of the analysis in order to facilitate the researcher's case-oriented analysis. We are surprised that Shalev positions his techniques as alternatives to QCA and fuzzy-set analysis when they are in fact rudimentary versions of QCA and fuzzy-set analysis. Perhaps the formality of QCA and fuzzy-set analysis makes these techniques appear inappropriate for case-oriented research. With regard to formal quantitative methods, Shalev (2007) cautions against such a reaction: "provided they fit researchers' theoretical assumptions, there is no reason why inductive multivariate statistical methods should not be exploited by comparativists." We extend this astute guidance to formal qualitative methods. It would be unfortunate if comparative researchers dismissed QCA and fuzzy-set analysis simply due to their formality.

As formal methods, QCA and fuzzy-set analysis provide useful ways of simplifying many of the common tasks that comparative researchers face. In constructing the tabular presentation of Rothstein's data, Shalev faced two tasks: developing the measures of the various causal conditions and building

a useful table showing key patterns. QCA formalizes the latter task, freeing researchers to concentrate on the former. Similarly, in developing his scatterplot of Garrett's data, Shalev had to measure the various indicators, build the scatterplot, and identify the relevant subsets. Fuzzy-set analysis frees researchers to concentrate on the measurement and calibration of set memberships; set-theoretic analysis of configurations of set memberships is accomplished using software. Shalev suggests that QCA and fuzzy-set analysis distance comparative researchers from their cases; in fact, the opposite is true. By formalizing the most difficult analytic tasks involved in comparative research – the comparison of cases as configurations of similarities and differences – these methods free researchers to direct their time and energy toward getting to know their cases well.

QCA and fuzzy-set analysis enhance comparative research by facilitating case comparisons. The analytic process brings contradictions to light and reveals conditions of limited diversity, providing avenues for further study. As noted above, QCA also offers procedures for the consideration of counterfactual cases. Perhaps most important, QCA and fuzzy-set analysis provide methods for the analysis of necessary and sufficient conditions as well as multiple conjunctural causation. These procedures, while formal, remain under the control of the researcher. In this manner, QCA and fuzzy-set analysis offers the transparency desired by comparative researchers while remaining faithful to the theoretical and substantive expertise of the researcher.

CONCLUSION

Michael Shalev's essay is an important contribution to the continuing debate on appropriate methods for comparative research. Drawing upon previously published research, he demonstrates a variety of ways in which the inappropriate application of multiple regression has compromised comparative work. Shalev proposes a number of alternative research strategies better suited to the needs of case-oriented researchers. It is important to note that Shalev's recommendation is not that comparative researchers abandon regression analysis or quantitative methods altogether, but instead that they learn to better match research questions and techniques. We strongly endorse this recommendation.

In the first half of this essay we clarify the various ways in which the choice of method matters. Different research methods embody different epistemological world-views. These different world-views shape the

questions that scholars may ask and, consequently, their results. Changing the research technique, then, can fundamentally alter the research project. In the second half of the essay, we address Shalev's critique of QCA and fuzzy-set analysis. Contrary to Shalev's assessment, QCA and fuzzy-set analysis are case-oriented techniques finely tuned to the needs and practices of comparative researchers. We demonstrate that QCA and fuzzy-set analysis incorporate and extend the insights and techniques of Shalev's recommended methods. Although Shalev positions his methods as alternatives to QCA and fuzzy-set analysis, we find greater similarity than difference among the approaches.

Comparative researchers frequently find themselves in the gulf between small-*N* qualitative studies and large-*N* quantitative studies (Ragin, 2000). Most of the studies that Shalev reviews involve between 14 and 18 countries, numbers small enough to constrain the available degrees of freedom but large enough to hinder in-depth analysis of each case. Case-oriented techniques such as QCA, fuzzy-set analysis, and those developed by Shalev permit the pursuit of both breadth and depth of understanding by assisting comparative researchers in their search for commonalities and differences across cases.

NOTES

1. See, for example, the extensive international bibliography on comparative methodology, QCA, and fuzzy sets at www.compasss.org, which lists more than 250 applications of QCA.

2. It is important to note that this critique does not apply to all quantitative methods. Social network analysis, for example, is both quantitative and case-oriented. Network analytic methods can be used to describe not only the cases within a network but also the overall network (the network itself, constituting a case). Reflecting the case-oriented researcher's concern with the relationships among sets, methods exist to assess the intersections, unions, and divisions within and between social networks. The point here is simply that one should not assume that case-oriented research is necessarily qualitative. Likewise, there is no reason to assume that variable-oriented research is necessarily quantitative.

3. In the article that Shalev references, Griffin and Ragin (1994, p. 10) overstate the resemblance between regression and QCA when they write "So similar are QCA and logit regression in causal epistemology, for example, that the very same data matrix can serve both kinds of analyses." Logit regression would be applied directly to the data set; QCA (or fuzzy-set analysis) would be applied to a truth table derived from the data set. Popular software applications such as fs/QCA automate the transformation of a conventional data set into a truth table, further obscuring this distinction.

4. A "contradiction" occurs when there are cases with identical causal configurations, except that some of the cases exhibit the outcome under investigation and others do not. Notice how the problem of contradictions highlights the difference between a truth table (in which rows represent configurations of causal conditions) and a conventional data set (in which rows represent observations).

5. The text indicates that this subset includes seven countries but only six are presented. We assume that France – which was included in Garrett's original analysis – was inadvertently omitted from this subset and would not change the results of the analysis.

6. Fuzzy scores range between 0.0 and 1.0. A score of 0.0 indicates that a case is fully out of the set of interest while a score of 1.0 indicates that a case is fully in the set. Scores between 0.0 and 0.5 indicate that a case is "more out than in" while scores between 0.5 and 1.0 indicate that a case is "more in than out."

REFERENCES

Achen, C. H. (2005). Two cheers for Charles Ragin. *Studies in Comparative International Development*, *40*(1), 27–32.

Berg-Schlosser, D., & De Meur, G. (1997). Reduction of complexity for a small-N analysis: A stepwise multi-methodological approach. *Comparative Social Research*, *16*, 133–162.

Eichengreen, B., Andrew, K. R., & Wyplosz, C. (1995). Exchange market mayhem: The antecedents and aftermath of speculative attacks. *Economic Policy*, *10*, 251–312.

Garrett, G. (1998). *Partisan politics in the global economy*. Cambridge: Cambridge University Press.

Griffin, L., & Ragin, C. C. (1994). Some observations on formal methods of qualitative analysis. *Sociological Methods & Research*, *23*, 4–21.

Hall, P. A., & Franzese, R. J., Jr. (1998). Mixed signals: Central bank independence, coordinated wage bargaining, and European Monetary Union. *International Organization*, *5*, 505–535.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Marcus, M., & Minc, H. (1988). *Introduction to linear algebra*. New York: Dover.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Ragin, C. C. (1994). *Constructing social research*. Thousand Oaks, CA: Pine Forge Press.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: The University of Chicago Press.

Ragin, C. C. (2005). *From fuzzy sets to crisp truth tables*. http://www.compasss.org/Raginfztt_April05.pdf

Ragin, C. C. (2006). The limitations of net-effects thinking. In: B. Rihoux & H. Grimm (Eds), *Innovative comparative methods for policy analysis* (pp. 13–42). New York: Springer.

Ragin, C. C., & Rihoux, B. (2004). Qualitative comparative analysis: State of the art and prospects. *Qualitative Methods*, *2*(2), 3–13.

Ragin, C. C., & Sonnett, J. (2004). Between complexity and parsimony: Limited diversity, counterfactual cases, and comparative analysis. In: S. Kropp & M. Minkenberg (Eds), *Vergleichen in der Politikwissenschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rueschemeyer, D., & Stephens, J. D. (1997). Comparing historical sequences – a powerful tool for causal analysis. Reply to John Goldthorpe's current issues in comparative macro-sociology. *Comparative Social Research*, *16*, 55–72.

Seawright, J. (2005). Qualitative comparative analysis vis-a-vis regression. *Studies in Comparative International Development*, *40*(1), 3–26.

Shalev, M. (2007). Limits and alternatives to multiple regression in comparative research. *Comparative Social Research*, *24*, 259–308.

Weber, M. (1905). Objective possibility and adequate causation in historical explanation. In: E. A. Shils & H. A. Finch (Eds), *The methodology of the social sciences* (pp. 164–188). Glencoe, IL: The Free Press.

Whitesitt, J. E. (1995). *Boolean algebra and its applications*. New York: Dover.

REJOINDER: AFFIRMING LIMITS AND DEFENDING ALTERNATIVES TO MULTIPLE REGRESSION

Michael Shalev

I greatly value the readiness of the eminent scholars participating in this symposium to debate the issues raised in my paper on the use of multiple regression (MR) in the comparative political economy of the OECD countries. Their thoughtful and often detailed commentaries testify to the existence of healthy differences of opinion alongside a shared commitment to methodological advance. These are encouraging signs of vitality and integrity. At a practical level, I believe our students will learn a lot from the symposium. Readers will of course need to make their own judgments. My comments focus on either clarifying my position where it seems necessary, or identifying what I believe are limits to some of the counter-suggestions made by the symposium contributors.

Most commentators interpreted my paper as calling for a blanket boycott of MR in small-N cross-national research. My intended message was that the costs and limitations of MR outweigh its benefits in comparison with alternative ways of analyzing numeric data. I first summarized the limitations of MR from a case-oriented perspective. Then, building on the existing critical literature on the popular pooled design (the merging of timeseries for multiple countries), I contended that it complicates analysis in often unacknowledged ways without overcoming difficulties that are inherent in using

Capitalisms Compared
Comparative Social Research, Volume 24, 391–409

ISSN: 0195-6310/doi:10.1016/S0195-6310(06)24014-6

linear and additive models to evaluate the effects of many variables on few cases. I noted some theoretically appealing uses of pooled datasets (such as testing how and why cross-country differences alter over time, or why temporal dynamics vary between countries or families of nations), but raised doubts about their viability.

The majority of my paper was devoted to illustrating the limits of MR in a variety of previous works. Through reanalysis of these works, I tried to show the advantages of technically simple exploratory methods of data analysis to be used where appropriate with synthetic variables created by methods of data reduction like factor analysis. These methods were chosen to maximize the potential for dialog between cases and explanations by keeping the cases visible during the data analysis. I argued that such visibility benefits comparativists (both producers and consumers of research) by allowing them to employ their knowledge of individual cases in judging the adequacy of measurement and the fit between data and conclusions, and also in identifying cases that merit closer study.

Most of the papers in this symposium fall into three groups. In two cases, the contributors largely or fully agree with my criticisms of standard regression approaches but believe that there are better ways of obtaining the benefits claimed for my proposed methods of analysis. Rubinson and Ragin advocate formal analytical methods based on Boolean or fuzzy-set algebra. Esping-Andersen proposes sidelining the conventional approach to MR and instead tapping the power of regression diagnostics. Two other contributors are sympathetic to some of my criticisms of MR but contend that practitioners are actually much more aware of these difficulties than I admit. They also believe that recent work using the pooled design has made successful efforts to overcome the problems which I identified, tapping the power of pooling to address theoretically interesting questions. Pontusson shows that modified pooled regressions have been insightfully used to address important topics like variation in causal processes over time and across types of countries. Similarly, Swank demonstrates that a number of major studies have succeeded admirably in cross-fertilizing both theory and qualitative case studies with pooled regression analysis.

Two other commentators also believe that MR has merit in comparative research, but (like Esping-Andersen) contend that it needs to be used more appropriately. Lane Kenworthy contributes to the project of improving the best practice of regression users by offering practical advice on how to deal with a variety of important issues, many of which were also raised in my article. I endorse Kenworthy's suggestions, but believe that some of them are more radical – and therefore less likely to be adopted – than he seems to

assume. In addition, while appreciating his efforts to bridge the divide between case and variable-oriented scholarship, I am less confident than Kenworthy that it is possible to accommodate the case-oriented critique within the framework of MR. Scruggs clearly thinks otherwise. At the same time, and in contrast to Kenworthy, he mounts a vigorous critique of my advocacy of case-oriented methodological principles, such as intimacy between analysis and cases and sensitivity to causal complexity. Because we differ on crucial points, a substantial part of my response will be devoted to Scruggs' paper.

The remaining commentary, by Rothstein, is distinctive in that unlike the other symposium authors, his work was one of the targets of my critique on the use of MR in comparative research. For this reason, his readiness to take part in the debate is especially welcomed. Moreover, there is much methodological and theoretical wisdom in Rothstein's remarks. Unfortunately, though, he misinterprets my treatment of his classic 1990 article on the effect of the "Ghent system" on union membership and writes as if my intention had been to offer a definitive answer to that perplexing question. (The same error motivates Scruggs' comments on this aspect of my article.) However, I was pursuing the more modest goal of showing that the purpose for which Rothstein originally invoked MR would have been better served by an exploratory tabular analysis. Rothstein points out that other sections of his paper, as well as other publications, offer a more nuanced and persuasive account of the causal role of the Ghent system. I am sure that he is correct, but the question at issue in my article was what can and cannot be learned from the use of MR in the specific study that I reviewed.[1]

THE SAME THING, ONLY BETTER?

At first sight the reader may find it odd that I have grouped the commentaries by Esping-Andersen and Rubinson/Ragin together, given that while the former proposes revamping the use of MR in comparative research, the latter offers a radical alternative. Nevertheless, while suggesting different solutions, both largely agree with my diagnosis of the problems. Also, my principal response to both is that their practical proposals look promising, yet are difficult to judge. That would require, (a) a user-friendly guide to implementing the advocated techniques; and (b) side-by-side comparison of the results obtained by their favored methods, the conventional MR method, and my own suggestions. Let me hasten to add that given (a), I would be ready to undertake the work involved in generating (b).

Rubinson and Ragin fill an acknowledged lacuna in my article by providing a valuable exposition of QCA/FSA (Quantitative Comparative Analysis and Fuzzy-Set Analysis). In addition, their paper elegantly and forcefully recapitulates and enriches my critique of MR, helpfully distinguishing between epistemological and methodological problems. They also believe that there is a genuine theoretical conflict between case and variable-oriented questions in comparative research, but I find this less convincing. In the background of variable-oriented research linking abstract causes and effects across many countries, there is often a burning desire to test generalizations inspired by the historical record of a particular country (Sweden is a favorite). At the same time, the sought-after end products of Ragin's methods of analyzing cases are generalizations, which specify values of the independent variables that predict a given value of the dependent variable. The theoretical questions driving case and variable-oriented research are thus not as distinct as Rubinson/Ragin suggest, but their discussion of this point is nevertheless illuminating in clarifying the comparative advantages of each approach.

I largely concur with Rubinson and Ragin's main claim that the methods which they advocate are "more elaborate and refined versions" of the exploratory techniques which I favor. There are two different reasons why my article gave only passing attention to the comparative-analytical methods developed by Ragin. First, I wanted to showcase an alternative to MR that was not already well-recognized in the literature. Second, never having actually worked with QCA or FSA, I felt unqualified to discuss them in any depth. I am still undecided about the potential value of these methods. It is not easy to learn them, because of limited documentation and buggy software.[2] Fortunately a growing number of studies in the general area of comparative political economy illustrate the use of QCA/FSA in practice, and a few usefully include the results from regression analyses as well (Ebbinghaus & Visser, 1999; Katz, vom Hau, & Mahoney, 2005; Nelson, 2004). However, some of the available examples of Ragin's methods, including his own studies of welfare state variation, have yielded rather disappointing results. In an early study using QCA, Ragin (1994b) was placed in the awkward position of having to assign one-third of his countries to a "spare" category, which effectively excluded them from the analysis. A subsequent effort using FSA (Ragin, 2000, Chapter 10) yielded a bewildering variety of different results, leaving this reader at least with the impression that the method was stretching the capacity of four explanatory variables arrayed against 18 cases.

Rubinson and Ragin level a similar charge against the tabular and graphical methods which I support, pointing out that they would be

hard-pressed to deal with more than three independent variables. It may be that the two approaches can be profitably employed in parallel, and not only for this reason. Combining both methods would prevent us from having to choose between case visibility and an impartial method of linking causes to cases. Judging from Rubinson and Ragin's description of how their methods are used, cases are only brought into the picture if the software yields surprises like finding more than one causal configuration for a single case. The actual process of discovering causal relationships is a black box. While they regard it as advantageous for researchers to leave it to software to come up with the answers, I am less certain of this. Given our rather primitive abilities to theorize and measure, it would be unwise to sacrifice human ability to judge the empirical adequacy of causal generalizations by applying expertise and common sense. Researchers (and their readers too) will be greatly helped in this respect by being able to view cases in relation to one another and their presumed causes, rather than relying on an automated algorithm to make the decisions. Accordingly, although QCA/FSA and the exploratory methods advocated in my paper may well represent two different methods of implementing the same approach, in contrast to Rubinson and Ragin I believe that each method has both advantages and disadvantages. I hope that future studies will exploit this complementarity.

Before leaving Rubinson and Ragin, it should be noted that we disagree regarding a specific but important element of my paper, namely the role which it advocates for methods of statistically summarizing affinities between multiple indicators. In principle, factor analysis and related techniques could be used for what they call "reducing the number of causal variables". However, that was not the purpose of the factor analysis which I carried out on Esping-Andersen's original welfare states dataset. Instead, my declared intention was to *validate* his typological classification of the *dependent* variable (the three welfare regimes). Moreover, Rubinson and Ragin exaggerate when they suggest that variables which load similarly on the same factor are "substitutable" and that, as a result, this technique suffers from the same problem of insensitivity to configurations that bedevils MR. First, even after rotation is performed it is not uncommon for the same indicator to load on more than one factor, and for these multiple factors to have quite different meanings (and causes). Second, configurations can be effectively described using combinations of factors. For instance, my analysis of welfare regimes showed that social-democratic and liberal countries share similarly negative values on "corporativism" while being located at opposite ends of the institutional/residual continuum famously described by Titmuss (1958).

Esping-Andersen himself seems to agree that my factor analysis-based replication of the analysis in his 1990 book is superior to the MR-based original. He also clearly shares my jaundiced view of the way that MR is actually practiced in cross-national studies. Nevertheless, he claims that there is a better way of doing regression, one that puts a priority on the use of diagnostic techniques. In the process, his essay offers an excellent survey of problems that challenge causal analysis of *any* kind in cross-national research, and how econometricians try to deal with them.

Esping-Andersen's recommendations can be divided into two different categories. Much of his commentary advises us to utilize sophisticated statistical techniques to identify causality issues and, where possible, to employ equally sophisticated techniques for resolving them. For instance: given causal feedback, correct for endogeneity; given selection bias, generate the missing counterfactual causal configurations using "statistical distributions".

Is this a practical agenda for comparativists? Diagnosing causality issues turns out to be far from trivial. It is discouraging to read, for instance, that statistical techniques identifying simultaneous causation cannot be used in small-N studies. Equally worrying is Heckman's observation, quoted by Esping-Andersen, that selection bias can only be reliably corrected in circumstances where it is not a problem to begin with. Another example, as Esping-Andersen points out, is that inferring causal relationships from cross-sectional comparisons is problematic if in the course of time the dependent variable feeds back onto the explanatory variable. The problem is that cross-national researchers lack sufficient data points to evaluate simultaneity using timeseries analysis, and in any event, many of their key (institutional) variables do not change much. This is of course only one of the many limitations of timeseries analysis for cross-national research. Indeed, the larger paper cited in Esping-Andersen's contribution to this symposium (Esping-Andersen & Przeworksi, 2001) provides a comprehensive catalog of the perils involved, including those noted in my article.

In sum, Esping-Andersen's first line of defense leaves me more pessimistic than ever about the utility of MR. Problems are astutely identified. Regression may or may not be able to uncover the problems, and workable solutions are especially hard to find. Fortunately, though, Esping-Andersen has a second string to his bow, which is the suggestion that MR users pay less attention to regression coefficients and more to residuals. By purposefully adding "dialogue with the cases" to the power of MR, this approach seems to overcome the tension between case and variable-oriented analytical strategies. However, when Esping-Andersen writes that "residual plots are a

minefield of information", he inadvertently raises the question of whether they are indeed a mine of gold rather than a field of buried explosive devices! It is difficult to answer this question without seeing concrete examples of how residual analysis is or could be used in comparative research, but some problems can be anticipated.

One of Esping-Andersen's strongest arguments on behalf of analyzing residuals is that if countries influence one another or share common traditions, their prediction errors are likely to cluster. While this is undoubtedly true, I suspect that informal familiarity with cases is more likely than MR to *inspire* insights like the diffusion hypothesis or the existence of "families of nations". At the same time, if the theory underlying MR is accepted, working with residuals may not provide a convincing *test* of such propositions. The virtue of the probabilistic approach built into the conventional way of doing regression analysis is precisely that it expects to uncover only broad *tendencies*. The assumption is that residuals may be influenced by measurement error, omitted explanatory variables and idiosyncratic features of the cases. In contrast, residual analysis of the kind that I understand Esping-Andersen is advocating could encourage reading too much into prediction errors (an issue raised by Scruggs). This is the kind of fear that motivates criticism of techniques like QCA, which strive for a perfect fit between cases and explanations. A further limitation of residual analysis is that if explanatory variables are inter-correlated, as they often are in cross-national research, the apparent predictive ability of each of the causal variables depends on the sequence in which they are entered into the regression.

These considerations point to advantages of the approach adopted in my reanalyzes of Rothstein and Hall/Franzese. This approach forfeits both the benefits and the burdens of precision by using broad categorical measures. It seeks out configurations of explanatory variables by cross-tabulating these measures. The resulting tables or charts make it possible to identify what combinations of attributes actually exist, how they apparently influence outcomes, and which anomalous cases or focused comparisons are worth pursuing in greater detail. The first and last of these three benefits seem more likely to emerge from the type of exploratory analysis that I advocate than from analysis of residuals.

POOLED REGRESSION REHABILITATED?

The contributions by Pontusson and Swank both argue that while pooling over-time and cross-country data has its problems, a large body of

sophisticated work in comparative political economy uses this technique responsibly and effectively. Pontusson highlights extensions to the pooling approach, which have greatly enhanced its value to comparativists. Swank shows that leading researchers actually follow many of my suggestions and prescriptions in conjunction with the use of pooled regression models. Both of their contributions are a useful antidote to my pessimistic view of the value of MR, and they may be correct that it is a more valuable tool than I admit.

In reviewing the use of pooled timeseries cross-section (PTSCS) models, my article emphasized that: (1) it cannot be assumed that dynamic and comparative-static (longrun) causality are identical, (2) the statistical advantages of simultaneously analyzing multiple country timeseries may be illusory, and (3) the technical complications inherent in pooling have spawned a Sisyphian spiral of critique and refinement, forcing practitioners to face an ever-rising learning curve.

Jonas Pontusson counters the first of these criticisms by contending that it must logically be the case that causes which hold over time also hold at the level of enduring cross-national differences. A supportive example would be the Swedish story as told by Korpi (1983), in which a country consistently dominated by left governments developed a comprehensive welfare state, beginning with a historic rise in left power that permanently altered the parameters of policymaking. Pontusson's argument that longrun effects logically embody the accumulated shortrun effects of the same explanatory variable does not necessarily hold, however. As Esping-Andersen suggests in his contribution to this symposium, both social democracy and the welfare state in Sweden could be the result of a common historical antecedent (cf. Therborn, Kjellberg, Marklund, & Ohlund, 1978). Alternatively, contemporary welfare state diversity may mirror the path-dependent effects of differential responses to one-off events like the Great Depression or World War II (e.g. Klausen, 1998). On the other side of the equation proposed by Pontusson, shortrun fluctuations in social expenditure may be driven by forces (such as election cycles or incrementalism) that have no causal relevance to the question of why some countries have enduringly bigger welfare states than others. Consistent with these reservations, Kenworthy's contribution to the symposium provides a convincing empirical illustration of the fact that different types of variation in welfare state spending may indeed have different causes.

Pontusson's main contention is that pooling potentially opens up new lines of empirical enquiry that allow us to tap what are arguably the most interesting types of questions confronting comparative researchers. Do

causal dynamics vary across different families of nations? Does the weight of factors that explain cross-national differences vary between different time-periods? Can we explain why causal relationships at the individual level vary across different national contexts? I agree with Pontusson in this respect. My paper applauded Western for trying to address the first question and criticized Hall and Franzese for failing to address the second. But I also emphasized that there are reasons to be skeptical whether pooling is more beneficial than simply inspecting the coefficients obtained from independently estimating regressions for different countries or time-points. Indeed, I suggested that "borrowing strength" could result in the statistical invention of non-existent effects.

My article did not address multilevel (also known as hierarchical or random-coefficient) models, the third type of suggestion made by Pontusson. These models facilitate the use of national characteristics to explain cross-country differences in individual-level effects.[3] The same reservations that my article raised in connection with Western's hierarchical pooled modeling apply to this design as well. One obvious concern is whether the heavy artillery of multilevel modeling is worth the effort. The first study to fully implement a complex multilevel design on Luxembourg Income Study data yielded findings of great importance for the study of welfare states (Mandel & Semyonov, 2005). However, reading the article in question one discovers that the elaborate statistical analysis produced results that are essentially no different from, and if anything less informative than, those presented in simple tables and scatterplots. This example sums up my overall response to Pontusson's commentary. While the extensions of pooling to which he draws attention indeed address important questions and have generated notable findings, I am not convinced that the pooling *technique* was a necessary means to this end. Pooled models may be useful for concisely summarizing contextual effects established by less sophisticated and parsimonious methods, such as "manually" comparing country-by-country or year-by-year regressions. But the credibility of these summary results depends on the strength of the underlying evidence.[4]

Duane Swank's paper is similar in spirit to Pontusson's but different in substance. Swank provides an enlightening survey of five major works in comparative political economy, which in his view were bypassed or undersold in my article. For the record, I should state that in preparation for a much earlier iteration of my article I drafted critiques based on close readings of three of these five books, those by Boix, Iversen and Garrett. For reasons of space, except for comments on Garrett's study these critiques were not included in the published version, although they were shared

privately with the authors and portions were presented at conferences and seminars. Briefly, in my view all three books suffer from an exaggerated belief in the power of PTSCS models. One reservation, discussed above in my response to Pontusson, concerns the importance of distinguishing between longrun and shortrun causality. Another is the dubious validity of using regression models (pooled or otherwise) to predict outcomes that represent non-existent causal configurations.[5] Boix and Garrett derived their most important empirical evidence on the basis of both of these questionable practices; Iversen relied only on the first.

Although Swank points out that Boix made effective use of charts, and also that his book dwells at length on two cases that drive his key quantitative findings, this does not alter the fact that Boix's pooled regressions and simulations suffer from the very same flaws which I claim are typical of pooled regression analyses. Moreover, while both Boix and Iversen enhanced their books by including case study chapters, Swank and I disagree on their importance. I find it striking that rather than building up generalizations from individual case studies and targeted case comparisons and then testing them statistically at lower resolution (across many cases), the authors relegated their qualitative material to later chapters, after the quantitative evidence was presented. However, I concede that my impression could have been mistaken. Swank may be correct that, particularly when the books by Hicks and himself are considered, major studies that relied on pooling have utilized case materials in fruitful dialogue with their statistical inferences. Needless to say, this in itself does not necessarily mean that their conclusions are correct,[6] but it does suggest that triangulation is more widely practiced than I acknowledged in my paper, and that as a result my pessimism on this score may not have been justified.

IS THERE A CASE FOR BEING CASE-ORIENTED?

Lyle Scruggs' contribution to the symposium offers a spirited defense of MR in comparative research. He claims that the problem is not MR itself, but the fact that it is practiced poorly. Indeed, he writes "statistical analysis is used atrociously in a lot, if not most, comparative social science". According to Scruggs, standard textbooks on regression already offer solutions to pseudo-problems that I raise, or else issue clear warnings against committing genuine errors in the practice of MR to which I draw attention.

In addition, and more fundamentally, Scruggs rejects the core ontological assumptions of case-oriented comparative analysis, claiming that they

"really undermine any attempt at explanation or verification in the sciences". However, what these assumptions actually challenge, and indeed seek to undermine, are over-simplistic explanations and inappropriate methods of verification. Specifically, my critique of MR practitioners is twofold. First, within the terms of their own epistemological discourse, they nearly always overstate their causality claims. Second, and more importantly, core features of the MR method are likely to burden rather than benefit macro-comparative researchers seeking to uncover or test for causality. I interpret Scruggs as agreeing with the first claim but strongly rejecting the second.

Scruggs takes issue with the view, most clearly articulated in Charles Ragin's work, that macro-comparative researchers should prefer methodologies which take it for granted that (a) a given outcome may be located at the end of more than one causal path, and (b) causal effects may be conjunctural (dependent on the broader constellation of conditions in which they are embedded). Scruggs distorts the first of these assumptions, asserting that it "makes any causal explanation largely irrefutable". In fact, the notion of multiple causal paths simply means that a given causal condition (or configuration) may be sufficient without being necessary. Scruggs also misinterprets the logic of conjunctural causation, contending that it is ultimately bound to lead to particularistic explanations ("irreducible differences in the cases themselves"). However, the main point is that adjectives matter. For example, capitalism may be authoritarian or democratic, and democracies can be two-party or multiparty. In each case, the nature of the coupling between capitalism and the political system could alter how "generic capitalism" affects the size of the social budget or the likelihood of a general strike. Because the real world of OECD countries contains a limited number of bundles of attributes, Scruggs' fears are unfounded. There may be main effects along with interaction effects (authoritarianism may exacerbate capitalism's tendency to immiserate the capital-poor), or there may only be interaction effects (it could be that proportional-representation systems inherently check capitalism's inegalitarian nature while parliamentary systems do not). None of this means that every country requires a unique explanation.

Scruggs warns that considering causes to be embedded in bundles could degenerate into giving up the aspiration to prioritize causes and identify decisive factors. This connects to a concern raised by Pontusson, that in-depth case studies should not be sacrificed in favor of using the results of case studies solely to establish superficial patterns in the data (whether via MR or QCA). I agree with both comments. In this spirit, when discussing

Rothstein's work on the causes of variation in trade union density, I emphasized that a logical next step after drawing conclusions from the visual clustering of cases and variables generated by my reanalysis would be to carry out paired case comparisons, which hold the promise of clarifying how much the Ghent system matters. A second type of enhancement advocated elsewhere in my paper (when discussing triangulation) was historical process-tracing, which has the unique promise of pinning down sequentially what it is that brings bundles of attributes together in the first place.

Oddly enough, after critiquing my efforts to draw attention to conjunctural causation and causal heterogeneity, Scruggs goes on to contend that MR is perfectly capable of handling these complications. He also claims that alternative methods, including those which I propose, are less rather than more appropriate than MR. I believe that he is mistaken. By his own admission, for MR to uncover the effects of causal configurations, researchers must know in advance what they are looking for. Herein lies the problem. Our theories can sometimes flag promising interactions, but they could turn up anywhere. Both informal qualitative comparisons like mine and Ragin's formal methods suggest that this problem should be addressed through a collaborative dialog between case evidence and received theory. However, as a committed deductivist Scruggs cannot accept the contaminating effects of such an approach. In line with the recommendations of King, Keohane, and Verba (1994), he proposes that scholars should either divide their data between the exploratory and testing phases of research, or else mobilize hitherto unexploited data that are also capable of addressing testable implications of their theories. The former counsel is often unhelpful to small-N researchers, and the latter is what has led many of them into a misguided romance with pooling.

Scruggs also errs in his critique of the specific reasons why I claim that MR is poorly suited for a world in which causes are bundled and a given cause can have varying effects in different contexts. Contrary to his assertion, case-oriented analysis does not suffer from a burden analogous to the loss of degrees of freedom that occurs when MR models add interaction terms. The reason is that it is only in the latter that interactions must be tested by adding what Scruggs calls another "moving part" to the model. Consider again Rothstein's study of unionization. My approach, which searches for configurations actually present in the data and does not attempt to make inferences to non-existent combinations of the explanatory variables, reveals that the effect of the Ghent arrangement can only be assessed in two specific clusters of countries, which are small and characterized by either medium or strong left party power. We can easily calculate from my

Chart 1 that the mean difference between Ghent and non-Ghent countries is larger in the countries with a medium left, suggesting an interaction. This inference is based on comparing four pairs of countries.[7] A regression modeler with a hunch that an interaction might be found along these lines, would presumably need to add three interaction terms to the model (Ghent∗size, Ghent∗government, and Ghent∗size∗government). This would offer the dubious benefit of allowing her to estimate the effects of all possible combinations of these three causal variables. Dubious, because as already noted, Ghent is only found in a very limited region of the parameter space. Since the case-centered approach does not aspire to explain empty cells, treating causation as conjunctural therefore amounts to anchoring our empirical generalizations in *more* cases (only those countries that actually populate each constellation of variables).

Scruggs' claims that causal heterogeneity "is precisely what multiple regression allows for" is especially misguided. He uses a hypothetical example to show that two explanatory variables, each of which correctly predicts the outcome for only some cases, may yield poor predictions alone while together explaining most or all of the variance. On the face of it, this is a resounding vindication of the additive model. However, the only reason this example works is that Scruggs conveniently uses only two independent variables and forces all cases to take on dichotomous values. (Rather ungraciously, Scruggs then criticizes non-MR researchers – presumably, QCA users – for dichotomizing their variables and taking a deterministic view of causation!) Furthermore, the coefficients on which regression analysts rely for their conclusions tell us the relative weight of each predictor, but not which cases it predicts. This is of course precisely why Esping-Andersen suggests in his commentary that when doing MR we should pay less attention to coefficients and more to residuals.

In Scruggs' fictitious example the two explanatory variables, A and B, have identical values in three cases while each of them uniquely predicts two other cases. In these unusual circumstances, analysis of residuals would work very well. But supposing, as case-oriented analysts do, that where there are multiple causal paths, each represents not one single variable but a configuration of several. Imagine also, as these analysts would, that at least one of the explanatory variables appears in several different causal configurations, each time with different effects. (See Ragin (1994a) for a textbook example along these lines.) *Given* such a model, an MR equation with appropriate interaction terms could be a parsimonious way of summarizing and verifying it. But not *revealing* it. Inspection of residuals is equally unhelpful under these circumstances unless we have a good idea in advance,

which are the conditioning variables and what cutoff points should be used to build the sub-groups for which meaningful comparisons of residuals could be conducted.

Finally, Scruggs seeks to undermine the credibility of my overall claims by exposing alleged errors and omissions in my reanalyses of earlier works. This is not the place for an itemized defense, but I do want to make two general points. The first is that a number of Scruggs' criticisms result from misreading my intentions. He chides me for not testing competing explanations on new ("external") data. However, the purpose of revisiting the works discussed in my article was to show that, using the same data, (a) their questions could be better answered with alternatives to MR (Rothstein; Esping-Andersen); or (b) technical innovations intended to overcome the limits of MR in comparative research actually fail to do so (Hall and Franzese; Garrett; Western). In evaluating Esping-Andersen's study, I concentrated on his conceptualization of the dependent variable, testing for the existence of welfare regimes by submitting a large number of policy indicators to factor analysis – in the process, turning "few cases/many variables" into a benefit instead of a burden. Disappointingly, Scruggs has nothing to say about the merits of this approach. Instead, he disputes the accuracy and validity of Esping-Andersen's original indicators and my decision not to utilize some of them in the factor analysis.

Overall, Scruggs' commentary exemplifies the chasm that continues to divide so many variable-oriented and case-oriented researchers. For example, he believes that residual analysis is a good thing, but only as a way of "examining assumptions" and not (as suggested by Esping-Andersen) in order to identify countries (with names!) that have something in common, which is unpredicted by the regression or is at odds with its assumption of independent errors. Similarly, Scruggs prefers broadly applicable generalizations to conditional ones and fears "ad hoc'ism". Are these simply differences in epistemological "tastes", or is it correct to believe (as I do) that a case orientation is fundamentally better suited to the type of investigations carried out by comparative political economists? Scruggs' preference for playing it by the (econometrics) book sometimes results in helpful reminders that we need to discipline our inferences. But it more often strikes me as stubborn loyalty to principles unsuited to cross-national research. To reiterate, a small and finite universe and what Ragin summarily calls causal complexity frequently make it impossible or impractical to separate exploration from testing, to hope for prediction errors that contain no systematic biases, or to aspire to discover novel explanations that hold broadly across all cases.

Ironically enough, along with what I regard as his exaggerated optimism concerning the suitability of MR to comparative research, Scruggs offers a distressing characterization of the realities of quantitative comparative research, in which practitioners unaccountably fail to abide by the rules clearly inscribed in econometrics textbooks. Scruggs concurs with my contention that repeated cross-sections in pooled models often add no useful information. He also agrees that it is a misuse of regression coefficients to "predict cases outside of the range of observed X's". Similarly, Scruggs makes it clear that much of the scientific veneer of "official" presentations of regression results is illusory. It is typically not the case that the model was born pristine from theory before ever meeting the data that underlie the reported findings, without any trial-and-error process of refinement. Similarly, when an observed coefficient is consistent with a researcher's hypotheses, few of them resist the temptation to refer to this as evidence of a "causal relationship".

Responding to the prevalence of regression malpractice, Scruggs suggests more than once that it is "an interesting question" why reputable scholars fail to use MR responsibly, but he offers no answers. My own seat-of-the-pants explanation is threefold.

1. Researchers who use regression often forget about fundamentals because they are preoccupied with technique. Some hope to gain attention by introducing something that is hot in econometrics but which has not yet reached their own discipline. Others worry too much about getting their standard errors right, lest they be unveiled as charlatans or fools sometime in the future.
2. Peer review does not always work because it focuses too much on whether state-of-the-art techniques are being applied, and not enough on whether the results are robust to varied methodological assumptions, or whether they make sense when scrutinized against the data used (which are usually not made available to reviewers anyway).
3. Whatever their private skepticism, for the sake of their own professional reputation and the prestige of their entire occupational community, participants in the production of scholarly literature have an interest in preserving an image of scientific respectability.

As Scruggs points out, more than one econometrician (my personal favorite is Leamer) have exposed the emperor's nudity. But few of us have had the temerity to advise him to put on a bathrobe, because we are too well socialized and have strong vested interests in not doing so. (I need to confess

to my own status as an occasional participant in this game, as well as a critic.)

No doubt good advice and better training would help to close the gap between the theory of MR and its practice in comparative research. Yet, given the realities just described, I am less optimistic than Scruggs on this score. Some of Kenworthy's recommendations also strike me as somewhat unrealistic, particularly his suggestion that researchers reveal the iterative process by which they use MR to arrive at their final model. Since this would publicly expose the backstage data-mining that goes on behind most front-stage hypothesis-testing, it seems unlikely to take root. Interestingly enough, because Kenworthy agrees that "understanding the cases" is the key task of macro-comparative research, he is not bothered by the prospect of giving up the deductive pretensions of mainstream quantitative studies. Indeed, his own recent work relies mainly on effective use of descriptive and exploratory methods of data analysis (Kenworthy, forthcoming), unlike many of his earlier publications that were based on advanced MR techniques without transparency of the kind that he now advocates (e.g. Kenworthy, 2002, 2003). Similarly, Pontusson freely admits that regression coefficients are "not really about … causal relationship[s]" and that "they themselves must be explained". However, the impressive papers he has coauthored that are cited in his commentary follow the convention that causal effects have been confirmed or disconfirmed.

In contrast to the gaps that both Scruggs and Kenworthy portray between the theory and practice of MR, when reading Pontusson's and especially Swank's upbeat reports on methodological advances in comparative political economy, one gets the impression that the errors and excesses of earlier implementations of pooled regression are now a thing of the past. In this view, recognizing that regression can offer no more than a crude representation of complex realities, researchers now routinely fill in gaps by other means, nuance their findings and theories, and use qualitative research to corroborate quantitative results. Against this, I continue to be struck by the excessive faith in econometric technique that is evident in published papers and books relying on MR, and indeed in many of the contributions to this symposium. It should not be forgotten that the leading political science journals in Europe and the USA repeatedly publish articles (which I cited) claiming to expose egregious statistical errors and omissions that allegedly invalidate what was previously thought of as state-of-the-art research. At the same time, we should also remember that even the wisdom of econometricians is not infinite. In this connection, I again urge that attention be paid to Maddala's pointed warnings to political methodologists a decade

ago (Maddala, 1997). Swank contends that I misquoted Maddala and he also downplays Maddala's criticism of Beck and Katz, presenting it as a run-of-the-mill difference of opinion between experts on a fine point of technique. Both claims are unfounded.[8]

In conclusion, while encouraged by the actual or potential advances in the use of MR that have been reported or advocated by symposium participants, my own view, like Maddala's, is that "I do not think the uncritical adoption of econometric methods in [comparative] political methodology is a good development" (Maddala, 1998, pp. 81–82). Exploratory techniques may be better suited to cross-national research than formal methods of analysis, including MR. At the very least, they should be utilized alongside these other methods and treated as the legitimate and uniquely appropriate tool that they are.

NOTES

1. It should be emphasized that my paper made it clear that "Rothstein's article was primarily based on comparative-historical analysis". It was also noted that he himself had "low expectations" from the regression analysis. A more specific instance of how both Rothstein and Scruggs misinterpret my intentions is their belief that I criticized Rothstein for not persuasively theorizing the Ghent-unionization relationship. Actually, observing that Rothstein himself questioned the theoretical adequacy of several of his independent variables, I pointed out (following an elementary principle of regression analysis) that if he did not believe these claims were true he should not have included them in the model.

2. Anecdotal evidence to this effect emerged from my experience in teaching comparative methods at the Oslo Summer School in July 2006. Although several of the participating graduate students had previous experience with the fs/QCA software my class and I were unable to implement an elementary example. It is especially unfortunate that no tutorials are yet available.

3. See the special issue of *Political Analysis* (Vol. 13, No. 4, 2005) on multilevel analysis, edited by Kedar and Shively.

4. The carefully crafted studies by Pontusson and his collaborators cited in his present paper furnish good examples of this axiom. Pooling only countries that belong to the same "variety of capitalism" (Rueda & Pontusson, 2000) is a definite advance on studies that aggregate the entire OECD bloc, but the article in question lacks evidence that the differences in effects "between-varieties" indeed exceed the "within-varieties" variation. Similarly, periodizing effects on the basis of a moving window analysis (Kwon & Pontusson, 2005) introduces welcome temporal conditionality, but in the absence of year-by-year evidence the validity of the periods chosen is difficult to judge.

5. The results of such simulations might become valuable when embellished by additional evidence and reasoning, as in the type of counterfactual thought

experiments advocated by Fearon (1991) and helpfully recalled in this symposium by Esping-Andersen.

6. In this connection it is sobering that three different case studies suggest that Boix's reading of the Spanish case, which was close to Socialist Party's interpretation, was deeply flawed (Etchemendy, 2004; McVeigh, 1999; Perez, 1999).

7. Readers who carry out this exercise for themselves will notice a further advantage of the transparency of the cases in the type of exploratory analysis illustrated in my Chart 1. Although the comparison of means affords strong evidence of interaction, the countries in one of the averaged pairs (comprising the Netherlands and Australia) are seen to have very different values on the dependent variable, alerting us to the fact that the apparent interaction holds only for the Dutch case.

8. After enumerating the many "basic issues to be tackled" in pooled models, Maddala (1998) stated: "These issues have been discussed in Beck and Katz (1996) for the benefit of political scientists. But their prescriptions are not, strictly speaking, correct." (p. 60). He went on to describe Beck and Katz's famed (among political scientists) method of panel-corrected standard errors as "sweeping the problems under the rug", adding however that some other errors made by Beck and Katz' are only "minor issues" (p. 61). The citation in my article accurately indicated that my quotations from Maddala were drawn from these two different pages of his article.

REFERENCES

Ebbinghaus, B., & Visser, J. (1999). When institutions matter: Union growth and decline in Western Europe, 1950–1995. *European Sociological Review*, *15*(2), 135–158.

Esping-Andersen, G., & Przeworksi, A. (2001). Quantitative cross-national research methods. In: N. J. Smelser & P. B. Bates (Eds), *International encyclopedia of the social and behavioral sciences* (pp. 12649–12655). New York: Elsevier Science.

Etchemendy, S. (2004). Revamping the weak, protecting the strong, and managing privatization—governing globalization in the Spanish takeoff. *Comparative Political Studies*, *37*(6), 623–651.

Fearon, J. D. (1991). Counterfactuals and hypothesis testing in political science. *World Politics*, *43*(2), 169–195.

Katz, A., vom Hau, M., & Mahoney, J. (2005). Explaining the great reversal in Spanish America: Fuzzy-set analysis versus regression analysis. *Sociological Methods & Research*, *33*(4), 539–573.

Kenworthy, L. (2002). Corporatism and unemployment in the 1980s and 1990s. *American Sociological Review*, *67*(3), 367–388.

Kenworthy, L. (2003). Do affluent countries face an incomes-jobs trade-off? *Comparative Political Studies*, *36*(10), 1180–1209.

Kenworthy, L. *Jobs with equality*. New York: Russell Sage Foundation, forthcoming.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Klausen, J. (1998). *War and welfare: Europe and the United States, 1945 to the present*. New York: St. Martin's Press.

Korpi, W. (1983). *The democratic class struggle*. London: Routledge and Kegan Paul.

Kwon, H. Y., & Pontusson, J. (2005). The rise and fall of government partisanship: Dynamics of social spending in OECD countries, 1962–2000. Princeton, NJ: Department of Politics, Princeton University, Unpublished paper.

Maddala, G. S. (1997). Recent developments in econometric modelling: A personal viewpoint. Paper presented at the annual meeting of the Political Methodology Group, Columbus, OH, July 24–27, http://wizard.ucr.edu/polmeth/working_papers97/madda97.html.

Maddala, G. S. (1998). Recent developments in dynamic econometric modelling: A personal viewpoint. *Political Analysis*, *7*, 59–87.

Mandel, H., & Semyonov, M. (2005). Family policies, wage structures, and gender gaps: Sources of earnings inequality in 20 countries. *American Sociological Review*, *70*(6), 949–967.

McVeigh, P. (1999). Globalisation and national economic strategy: The case of Spain. *Journal of European Area Studies*, *7*, 73–90.

Nelson, K. (2004). The last resort: Determinants of the generosity of means-tested minimum income protection in welfare democracies. Paper presented at the European Social Policy, ESPANet conference, Stockholm, Unpublished paper, http://www.apsoc.ox.ac.uk/Espanet/espanetconference/papers/ppr%5B1%5D.18.KN.pdf.pdf.

Perez, S. A. (1999). From labor to finance. Understanding the failure of socialist economic policies in Spain. *Comparative Political Studies*, *32*(6), 659–689.

Ragin, C. C. (1994a). *Constructing social research: The unity and diversity of method*. Thousand Oaks, CA: Pine Forge Press.

Ragin, C. C. (1994b). A qualitative comparative analysis of pension systems. In: T. Janoski & A. M. Hicks (Eds), *The comparative political economy of the welfare state* (pp. 320–345). Cambridge: Cambridge University Press.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Rueda, D., & Pontusson, J. (2000). Wage inequality and varieties of capitalism. *World Politics*, *52*(3), 350–383.

Therborn, G., Kjellberg, A., Marklund, S., & Ohlund, U. (1978). Sweden before and after social democracy: A first overview. *Acta Sociologica*, *21*, 37–58.

Titmuss, R. (1958). *Essays on the 'Welfare State'*. London: Allen and Unwin.