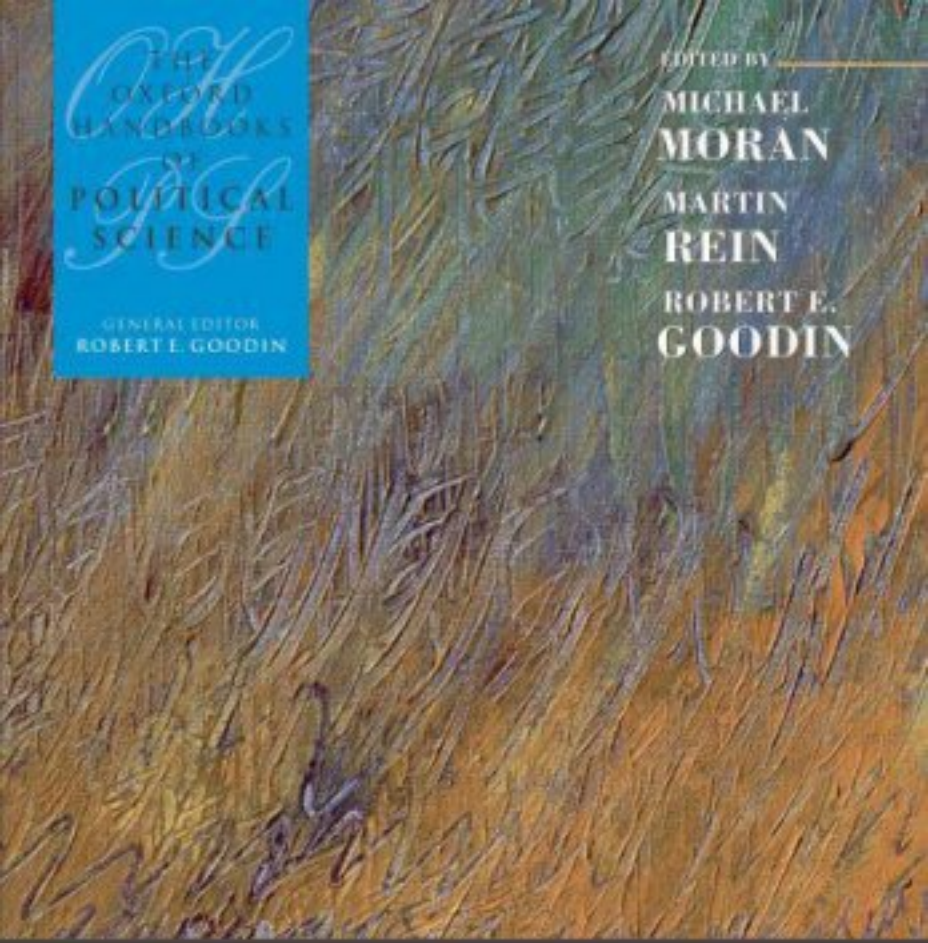

The Oxford Handbook *of* PUBLIC POLICY

Heckman, J., Lalonde, R., and Smith, J. 1999. The economics and econometrics of active labor market programs. Pp. 1865 2097 in *Handbook of Labor Economics*, vol. iii, ed. A. Ashenfelter and D. Card. Amsterdam: Elsevier.

Immervoll, H., et al. 2004. The effect of taxes and transfers on household incomes in the European Union. Paper accepted for the conference "The distributional effects of government spending and taxation," 15 16 Oct., Levy Institute of Bard College.

Jencks, C. 1992. *Rethinking Social Policy: Race, Poverty, and the Underclass*. Cambridge, Mass.: Harvard University Press.

Johansson, L., Sundstrom, G., et al. 2003. State provision down, offspring's up: the reverse substitution of old age care in Sweden. *Ageing and Society*, 23 (3): 269 80.

Korpi, W., and Palme, J. 1998. The paradox of redistribution and strategies of equality: welfare state institutions, inequality, and poverty in the western countries. *American Sociological Review*, 63: 661 87.

Künemund, H., and Rein, M. 1999. There is more to receiving than needing: theoretical arguments and empirical explorations of crowding out. *Ageing and Society*, 19: 93 121.

Lingsom, S. 1997. *The Substitution Issue: Care Policies and their Consequences for Family Care*. Report 6. Oslo: NOVA

Mahler, V., and Jesuit, D. 2004. *State Redistribution in Comparative Perspective: A Cross national Analysis of the Developed Countries*. Luxembourg Income Study Working Paper No. 392. Syracuse, NY: Syracuse University.

Mitchell, D. 1991. *Income Transfers in Ten Welfare States*. Aldershot: Avebury.

Murray, C. 1984. *Losing Ground: American Social Policy 1950 1980*. New York: Basic Books.

Nelissen, J. H. M. 1993. *The Redistributive Impact of Social Security Schemes on Lifetime Labour Income*. Tilburg: Proefschrift Katholieke Universiteit Brabant.

OECD 2002. *Benefits and Wages*. Paris: OECD.

Oxley, H., Dang, Th. Th., Förster, M. and Pellizari, M. 2001. Income inequalities and poverty among children and households with children in selected OECD countries. Pp. 371 405 in *Child Well Being, Child Poverty and Child Policy in Modern Nations: What do we Know*, ed. K. Vleminckx and T. Smeeding. Bristol: Policy Press.

Pechman, J., and Timpane, P. M. 1975. *Work Incentives and Income Guarantees*. Washington, DC: Brookings Institution.

Peltzman, S. 1975. The effects of automobile safety regulation. *Journal of Political Economy*, 83 (4): 677 726.

Penning, M., and Keating, N. 2000. Self , informal and formal care: partnerships in community based and residential long term care settings. *Canadian Journal on Aging*, 19 (Suppl. 1): 75 100.

Ragin, C. 1987. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ringen, S. 1989. *The Possibility of Politics: A Study in the Political Economy of the Welfare State*. Oxford: Clarendon.

Schoeni, R., and Blank, R. 2000. *What has Welfare Reform Accomplished? Impacts on Welfare Participation, Employment, Income, Poverty and Family Structure*. NBER Working Paper No. W7627. Cambridge, Mass.: National Bureau of Economic Research.

Sutherland, H. 2001. *EUROMOD: An Integrated European Benefit Tax Model*. Euromod Working Paper No. EM9/01. Cambridge: Department of Applied Economics, University of Cambridge.

Titmuss, R. 1974. *Income Distribution and Social Change*, London: Allen and Unwin.

Van den Bosch, K. 2002. Convergence in poverty outcomes and social income transfers in member states of the EU. Paper for the XV World Congress of Sociology, Brisbane, July.

VERBIST, G. 2004. Herverdeling door de fiscus. Effecten van de personenbelasting op de inkomensongelijkheid in België en andere OESO landen. *Kwartaaltijdschrift Economie*, 3: 284 303.

WAGSTAFF, A., et al. 1999. Redistributive effect, progressivity and differential tax treatment: personal income taxes in twelve OECD countries. *Journal of Health Economics*, 72: 73 98.

WEST PEDERSEN, A. 1994. *The Welfare State and Inequality: Still no Answer to the Big Questions*. Luxembourg Income Study Working Paper No. 109. Syracuse, NY: Syracuse University.

CHAPTER 15

THE POLITICS OF POLICY EVALUATION

MARK BOVENS
PAUL ´T HART
SANNEKE KUIPERS

1. Evaluation between "Learning" and "Politicking"

In this chapter policy evaluation refers to the *ex post* assessment of the strengths and weaknesses of public programs and projects. This implies we shall not address the voluminous literature on *ex ante* policy analysis, where methods to evaluate policy alternatives are developed and offered to policy makers and other stakeholders as decision-making aids (see, e.g., Nagel 2002; Dunn 2004). We shall argue that policy evaluation is an inherently normative act, a matter of political judgement. It can at best be informed but never fully dominated by scholarly efforts to bring the logic of reason, calculation, and dispassionate truth seeking to the world of policy making. Policy analysis's mission to "speak truth to power" (Wildavsky 1987) is laudable, and should be continued forcefully, but scholars should not be naive about the nature of the evaluation game they participate in (Heineman et al. 1990, 1). In the ideal world

of policy analysis, policy evaluation is an indispensable tool for feedback, learning, and thus improvement. In the real world of politics, it is always at risk of degrading into a hollow ritual or a blame game that obstructs rather than enhances the search for better governance.

When public policies are adopted and programs implemented, the politics of policy making do not come to an end. The political and bureaucratic controversies over the nature of the problems to be addressed and the best means by which to do so that characterize the policy formulation and policy selection stages of the policy cycle do not suddenly abate when "binding" political decisions are made in favour of option X or Y. Nor do the ambiguities, uncertainties, and risks surrounding the policy issue at stake evaporate. They merely move from the main stage, where political choices about policies are made, to the less visible arenas of policy implementation, populated by (networks of) bureaucratic and non-governmental actors who are involved in transforming the words of policy documents into purposeful actions. At one time or another, the moment arrives to evaluate what has been achieved. This moment may be prescribed by law or guided by the rhythm of budget or planning and control cycles. It may, however, also be determined by more political processes: the replacement of key officials, elections that produce government turnovers, incidents or figures that receive publicity and trigger political calls for an investigation, and so on.

Whatever its origins, the ideal-typical structure of a formal evaluation effort is always the same: an *evaluating body* initiates an investigation with a certain *scope* (what to evaluate: which programs/projects, policy outcomes, and/or policy-making processes, over which time period?); it employs some—explicit or implicit—evaluation *criteria*; it gathers and analyzes pertinent *information*; it draws *conclusions* about the past and *recommendations* for the future; and it *presents its findings*. Beneath this basic structure, tremendous variations exist in evaluation practices (Fischer 1995; Vedung 1997; Weiss 1998; Weimer and Vining 1999; Nagel 2002; Dunn 2004). They differ in their analytical rigor, political relevance, and likelihood to produce meaningful learning processes (cf. Rose 1993).

Bodies that conduct evaluations range from scientific researchers acting on their own accord to consulting firms to public think tanks, and from institutionalized watch dogs such as ombudsmen or courts of audit, to political bodies such as parliamentary commissions. Some of these evaluations are discreet and for direct use by policy makers; others occur in a blaze of publicity and are for public consumption and political use. One and the same policy program or episode may be evaluated by several of these bodies simultaneously or over time. It frequently happens that one type of evaluation exercise triggers others. For instance, the crash of a Dutch military cargo plane at Eindhoven airport in 1996 and the subsequent disaster response by the military and local authorities led to no less than fifteen separate investigation efforts by various government bodies, courts, and think tanks. This cascading effect was partly caused by the fact that both the cause of the accident

and the adequacy of the response were subject to speculation and controversy, including the taking of provisional disciplinary sanctions against military airport officials. Moreover, different evaluation bodies may even compete overtly: government-initiated versus parliamentary evaluations, different chambers of parliament with different political majorities each conducting their own investigations into some presumed policy fiasco, governmental versus stakeholder evaluations, national versus IGO evaluations, and so on. The Reagan government's so-called Iran-Contra affair (which included the selling of arms to Iran in the hope of securing the release of American hostages held by Shi'ites in Lebanon) set in motion three evaluation efforts: one by a blue-ribbon presidential commission, one by the Senate, and one by the House of Representatives. Not surprisingly, the three reports were all critical of the course and outcomes of the policy, but differed markedly in the attribution of responsibility for what happened (see Draper 1991).

In the ideal world of the positivist social scientist, we stand to gain from this multiplicity: presumably it results in more facts getting on the table, and thus a more solid grasp of what happened and why. In the real world, multiple evaluations of the same policy tend to be non-cumulative and non-complementary. Their methods and findings diverge widely, making it hard to reach a single authoritative or at least consensual judgement about the past and to draw clear-cut lessons from it.

In this chapter we shall approach the politics of policy evaluation in two ways. First we shall elaborate on the roles and functions of policy evaluation in the broader politics of public policy making. Then we shall look at how key schools of policy analysis propose to deal with the essentially contested, inherently political nature of evaluation. Each, we argue, has crucial strengths and shortcomings. In the final section, we offer our own view of how policy analysis may cope with the conundrum of *ex post* evaluation.

2. The Politics of Policy Evaluation

It is only a slight exaggeration to say, paraphrasing Clausewitz, that policy evaluation is nothing but the continuation of politics by other means. This is most conspicuous in the assessment of policies and programs that have become highly controversial: because they do not produce the expected results, because they were highly contested to begin with, because they are highly costly and/or inefficient, because of alleged wrongdoings in their implementation, and so on. The analysis of such policy

episodes is not a politically neutral activity, which can be done by fully detached, unencumbered individuals (Bovens and 't Hart 1996). The ominous label of "failure" or "fiasco" that hovers over these policies entails a political statement. Moreover, once policies become widely viewed as failures, questions about responsibility and sometimes even liability force themselves on to the public agenda. Who can be held responsible for the damage that has been done to the social fabric? Who should bear the blame? What sanctions, if any, are appropriate? Who should compensate the victims? In view of this threat to their reputations and positions, many of the officials and agencies involved in an alleged fiasco will engage in tactics of impression management, blame shifting, and damage control. The policy's critics, victims, and other political stakeholders will do the opposite: dramatize the negative consequences and portray them as failures that should, and could, have been prevented (cf. Weaver 1986; Gray and 't Hart 1998; Anheier 1999; Hood 2002).

The pivotal importance of blaming entails the key to understanding why the evaluation of controversial policy episodes itself tends to be a highly adversarial process. The politics of blaming start at the very instigation of evaluation efforts: which evaluation bodies take on the case, how are they composed and briefed (Lipsky and Olson 1977)? It is highlighted especially by the behaviour of many stakeholders during the evaluation process. To start with, the very decision to have an incident or program evaluated may be part of a political strategy. Penal policy constitutes an interesting example of this. In most countries, prison escapes take place from time to time, and in some periods their incidence increases. But there appears to be no logical connection between objectifiable indicators of the severity of the problem such as their frequency, their success rate, the number of escapees per annum, and the likelihood of major evaluation and learning efforts being undertaken at the political level. In the Netherlands, for example, political commotion about prison escapes rose to peak levels at a time when all penal system performance indicators were exceptionally good after an earlier period of problems and unrest. Rather, the scale, scope, and aims of a post-escape investigation seem to be a function of purely coincidental factors such as the method of escape and the level of violence, as well as the nature of the political climate regarding criminal justice and penal policy at any given time (Boin 1995; Resodihardjo forthcoming).

Even seemingly routine, institutionalized evaluations of unobtrusive policy programs tend to have political edges to them, if only in the more subterraneous world of sectoral, highly specialized policy networks. Even in those less controversial instances, policy evaluations are entwined with processes of accountability and lesson drawing that may have winners and losers. However technocratic and seemingly innocuous, every policy program has multiple stakeholders who have an interest in the outcome of the evaluation: decision makers, executive agencies, clients, pressure groups. All of them know that apart from (post-election) political turnovers or crucial court cases, evaluations are virtually the only moments when existing policy trajectories can be reassessed and historical path dependencies may be broken (cf. Rose and Davies 1994). Evaluations hold the promise of a reframing of a program's

rationale and objectives, a recalibration of the mix of policy instruments it relies on, a reorganization of its service delivery mechanisms, and, yes, a redistribution of money and other pivotal resources among the various actors involved in its implementation. Hence in the bulk of seemingly "low-politics program" evaluations, the stakes for the circle of interested parties may be high (Vedung 1997, 101–14; Pawson and Tilly 1997; Radin 2000; Hall and Hall 2004, 34–41).

Astute players of the evaluation game will therefore attempt to produce facts and images that suit their aims. They will produce—or engage others to produce—accounts of policy episodes that are, however subtly, framed and timed to convey certain ideas about what happened, why, and how to judge this, and to obscure or downplay others. They will try to influence the terms of the evaluation, in particular also the choice and weighting of the criteria by which the evaluators arrive at their assessments. Evaluating bodies and professional policy analysts will inevitably feel pressures of this kind building up during the evaluation process. The list of tactics used by parties to influence the course and outcomes of evaluation efforts is long, and somewhat resembles the stratagems of bureaucratic and budgetary politics: evaluators' briefs and modus operandi may be subject to continuous discussion; key documents or informants may prove to be remarkably hard, or sometimes remarkably easy, to encounter; the drafting and phrasing of key conclusions and recommendations may be a bone of contention with stakeholder liaisons or in advisory committees; there may be informal solicitations and *démarches* by stakeholders; reports may be prematurely leaked, deeply buried, or publicly lambasted by policy makers. In short, even the most neutral, professional evaluators with no political agenda of their own are likely to become both an object and, unwittingly or not, an agent of political tactics of framing, blaming, and credit claiming (see Bovens et al. 1999; Brändström and Kuipers 2003; Pawson and Tilley 1997; Stone 1997).

3. Dealing with the Political in Policy Evaluation

Policy scientists have long recognized these political ramifications of policy evaluation, but have found it impossible to agree on how to cope with them. The cybernetic notion of evaluation as a crucial, authoritative "feedback stream" that enhances reflection, learning, and thus induces well-considered policy continuation, change, or termination, has ceased to be a self-evident rationale for elaborating evaluation theory and methodology. The political realities have simply been too

harsh. "The field of evaluation is currently undergoing an identity crisis," lamented two advocates of the positivist approach to policy analysis twenty years ago (Palumbo and Nachimas 1983, 1). At that time, a multitude of alternative approaches had taken the place of the single methodology and assumption set of the classical, first-generation policy analyst of the science-for-policy kind. The mood of optimism and its belief in planned government intervention that had characterized for instance Johnson's "Great Society Program" in the United States was replaced by a mood of scarcity and skepticism (Radin 2000; see also Rossi and Freeman 1993, 23). The focus in policy analysis shifted from *ex ante* evaluation to *ex post* evaluation, because the creation of large public policies became less fashionable than the scrutiny of existing programs (Radin 2000, 34). As Dye (1987, 372) put it, it became "exceedingly costly for society to commit itself to large-scale programs and policies in education and welfare, housing, health and so on, without any real idea about what works." Instrumental policy evaluation continued to be a stronghold in the field of policy analysis, although it was now increasingly exploited as a tool to measure *ex post* cost–benefit ratios to support retrenchment efforts by New Right governments (Radin 2000; Fischer 1995).

At the same time, the value trade-offs and political controversies involved in the scrutiny of existing public policies raised questions about the neutrality assumptions of policy analysis. The apolitical, quantitative assessments of policy outcomes that were supposed to support optimal decision making in the 1950s and 1960s became the subject of increasing criticism. The judgemental character of policy evaluation provoked discussion about its inherently normative, political nature, and about the initial stubbornness among policy analysts steeped in the rationalistic tradition to deny that evaluating policy impact is "an activity which is knee-deep in values, beliefs, party politics and ideology, and makes 'proving' that this policy had this or that impact a notion which is deeply suspect" (Parsons 1995, 550). A new generation of policy analysts came up, and rejected the fundamental assumption that it is possible to measure policy performance in an objective fashion. Like Hugh Heclo, they argued that "a mood is created in which the analysis of rational program choice is taken as the one legitimate arbiter of policy analysis. In this mood, policy studies are politically deodorized—politics is taken out of policy-making" (Heclo 1972, 131). Several approaches to policy evaluation were developed to "bring politics back in" (Nelson 1977; Fischer 1980; Majone 1989).

The diversity of evaluation approaches that has developed since will be discussed here in terms of two traditions. The dividing line between those traditions will be based on the way norms, values, interests, and power are accommodated in evaluation. The *rationalistic tradition* with its strong emphasis on value neutrality and objective assessments of policy performance tries to save evaluation from the pressures of politics, by ignoring these pressures or somehow superseding them. In contrast, the *argumentative tradition* sees policy evaluation as a contribution to the informed debate among competing interests and therefore explicitly incorporates politics in the *ex post* analysis of policy performance.

3.1 Rationalistic Policy Evaluation

The rationalists advocate a rigorous separation of facts and values and explicitly strive to produce apolitical knowledge (Hawkesworth 1988; Lynn 1999; Mabry 2002). Policy analysis is rooted in positivism and strives to produce factual data about societal structures and processes by employing concepts and methods borrowed from the natural and physical sciences. Policy analysis serves to bring about rational decision making in the policy process. Judgements about a program's or project's effectiveness and efficiency have to be based on reliable empirical data. It is the task of the policy analyst to produce information that is free from its psychological, cultural, and linguistic context. Because such information transcends historical and cultural experiences, it is assumed to have political and moral neutrality.

Rational methods can be used to construct theoretical policy optimums (in terms of both efficiency and efficacy); in evaluation one can then measure the distance of actual policy outcomes from this optimum. Evaluation thus yields policy-relevant information about the discrepancies between the expected and factual policy performance (Dunn 2004). According to Berk and Rossi (1999, 3) evaluation research is "essentially about providing the most accurate information practically possible in an even-handed manner." Political decisions and judgements require testimonies based on generally applicable and scientifically valid knowledge for "it is rarely prudent to enter a burning political debate armed with only one case study" (Chelimsky 1987, 27). The effort to "remedy the deficiencies in the quality of human life" requires continuous evaluation directed at the improvement of policy programs, based on valid, reliable empirical information (Rossi, Freeman, and Lipsey 1999, 6).

This form of policy evaluation assumes the existence of an exogenously produced, i.e. given, set of clear and consistent policy goals and/or other evaluation standards. It also assumes intersubjective agreement on which indicators can be identified to measure the achievement of these goals. Some rationalistic evaluators might acknowledge that evaluation is in essence a judgement on the value of a policy or program and therefore goes beyond the realms of empirical science (Dunn 2004), or that policy evaluation takes place in a political context with a multitude of actors and preferences involved. For example, Nagel's (2002) approach to *ex ante* policy evaluation includes political considerations to the extent that it proposes a "win-win analysis" to be made: a survey and assessment of the preferred alternatives of political actors involved to find among them an alternative that exceeds the best initial expectations of representatives of the major viewpoints in the political dispute. But their bottom line is clear: Dunn (2004), for instance, asserts that the outcome of policy evaluation is a value judgement, but that the process of evaluation nevertheless has to provide unbiased information. Likewise, the Rossi et al. (1999) handbook self-consciously advocates the systematic application of social research procedures, emphasizing the analysis of costs and benefits, targets, and effects. Earlier, they did not only argue that evaluation should provide value-neutral information to political

decision makers, but also that context-sensitive, biased, and argumentative evaluators are "engaged in something other than evaluation research" (Rossi and Freeman 1993, 33).

A remarkably influential institutionalized manifestation of the rationalistic approach to policy evaluation is the Organization for Economic Co-operation and Development (OECD). The OECD aims to foster good governance by monitoring and comparing economic development, deciphering emerging issues, and identifying "policies that work" (according to its own website at www.oecd.org). Its country reports have gained considerable authority over the years and its standardized comparisons are used as verdicts on national policy performance.

3.2 Argumentative Policy Evaluation

This brings us to the other camp. The argumentative critics of the rationalist approach complain that the positivist world view is fundamentally distorted by the separation of facts from values. Policy intervention with respect to social and political phenomena is an inherently value-laden, normative activity which allows but for a biased evaluation (Fischer and Forester 1993; Guba and Lincoln 1989). The so-called "post-positivists" or social constructivists understand society as an organized universe of meanings, instead of a mere set of physical objects to be measured. It is not the objects per se that are measured, but the interpretation of the objects by the scientist. The system of meanings shapes "the very questions that social scientists choose to ask about society, not to mention the instruments they select to pursue their questions" (Fischer 1995, 15). Facts depend on a set of underlying assumptions that give meaning to the reality we live in. These assumptions are influenced by politics and power, and empirical findings based on these underlying assumptions "tend to reify a particular reality" (Fischer 1998, 135). The first evaluation of the "Great Society's" Head Start program for socially deprived children was a measurement of the participating children's cognitive development shortly after the program's implementation. This measurement was a relatively simple quantitative assessment of only one of the program's possible positive effects. It showed a lack of improvement in the children's cognitive capacities and that, compared to the total costs of the government intervention, the program had been an expensive failure. If only the evaluators had accepted the program's underlying assumptions that children would benefit from their participation by gaining social experience that would teach them how to function successfully in middle-class-oriented educational institutions, they would have awaited the results of long-term monitoring. The short-term evaluation outcomes were very welcome to the new Nixon administration as an argument to cut down on Head Start considerably (Fischer 1995). The short-term cost–benefit analysis that befitted Nixon's attack on large-scale government planning efforts served to prove him right.

Likewise, the standardized comparison of budgetary and performance figures employed by think tanks such as the OECD leaves open much interpretative and therefore contested ground. One ground for dispute concerns the construction of the categories. In the OECD's report, the Belgian unemployment rate was put just above 8 per cent of the total labor force; in contrast, the Belgian unemployment agency's (www.rva.be) own reports state that it pays unemployment benefits to more than a million people monthly, i.e. 23.5 per cent of the labor force (Arents et al. 2000). The disparity can only be explained by examining closely the definitions of "unemployment" used in studies such as these.

To post-positivists this is just one example among many. They claim it is an illusion to think that separation between values and facts is possible. Moreover, it is impossible to create a division of labor between politics and science where politicians authoritatively establish policy values and scientists can neutrally assess whether the policy outcomes meet the prior established norms (Majone 1989). Policy analysts should actively engage in and facilitate the debate on values in policy making and function as a go-between for citizens and politicians. By attempting to provide "the one best solution" in *ex ante* policy analysis and the "ultimate judgement" in *ex post* evaluation, the ambition of most (rationalist) policy scientists has long been to settle rather than stimulate debates (Fischer 1998).

The advocates of the argumentative approach see yet another mission for policy analysis, including evaluation. Knowledge of a social object or phenomenon emerges from a discussion between competing frameworks (Yanow 2000). This discussion—or discursive interaction—concerning policy outcomes can uncover the presuppositions of each framework that give meaning to its results from empirical research. Policy analysts can intervene in these discussions to help actors with different belief systems understand where their disagreements have epistemological and ethical roots rather than simply boiling down to different interests and priorities (Van Eeten 1999; Yanow 2000). If evaluations can best be understood as forms of knowledge based on consensually accepted beliefs instead of on hard-boiled proof and demonstration (Danziger 1995; Fischer 1998), it becomes quite important to ascertain whose beliefs and whose consensus dominates the retrospective sense-making process. Here, the argumentative approach turns quite explicitly to the politics of policy evaluation, when it argues that the deck with which the policy game is played at the evaluation can be stacked as a result of institutionalized "mobilization of bias." In that sense evaluation simply mirrors the front end of the policy process (agenda setting and problem definition): some groups' interests and voices are organized "in" the design and management of evaluation proceedings, whereas other stakeholders are organized "out." Some proponents of argumentative policy evaluation therefore argue that the policy analyst should not just help expose the meaning systems by which these facts are being interpreted; she should also ensure that under-represented groups can make their experiences and assessments of a policy heard (Fischer and Forester 1993; Dryzek 2000).

DeLeon (1998) qualifies the argumentative approach's enthusiasm about "consensus through deliberation." He cautions that the democratic ambitions of