

# *THE EMERGENCE OF PROGRAM EVALUATION: EXPERIENCES IN THE NETHERLANDS*

Henk A. Becker,\*  
Dik van Kreveld and Frans A. van Vught

## INTRODUCTION

Impact assessment program evaluation (PE) research is a rapidly growing field (Crónbach et al., 1980; Raizen and Rossi, 1981; Becker, 1986). The Netherlands and other countries have witnessed an increase in demand for program evaluation. This issue of the *Bulletin* addresses the topic of PE by presenting papers delivered on the subject at the 1988 IAIA European Conference in Leiden and Delft. In this, the opening article, we present a review of PE in The Netherlands with reference to developments in other countries as well as a brief overview of papers found in this issue.

A working definition of a program is “a well defined complex of activities, structured in time with regard to periods, undertaken to accomplish one or more well defined targets” (Schnabel, 1984). “A complex of activities” implies that more than one intervention is at stake while “. . . structured in time with regard to periods” means that at least two periods can be discerned in the implementation of the complex of activities.

We can define PE as “impact assessment and subsidiary policy-oriented research accompanying a program.” On a small scale this may imply an ex post evaluation of two periods of a program. More elaborate PEs may include a baseline analysis and ex ante and ex post evaluations of each period and of the program as a whole. An overview is given in Table 1 of an elaborate program and an accompanying PE.

---

\*Keynote paper presented at the European IAIA Chapter Conference, Leiden and Delft, The Netherlands, June 16-17, 1988

Table 1: Model of an Intervention Program and Accompanying Evaluation\*

PROGRAM	PROGRAM EVALUATION
1. Period of preliminary activities	1. Baseline analysis, scenarios and forecasts, general ex ante evaluation
2. Period of policy formation	2. Focussed ex ante evaluation selection of priorities
3. Period of program implementation, first phase	3. Ex post evaluation (formative) ex ante evaluation of phase 2
4. Period of program implementation, second phase	4. Ex post evaluation (formative)
5. Period of policy evaluation, reconsideration; continuation or termination of intervention program	5. Ex post evaluation (summative) ex ante evaluation (of new policy options)

\* This model gives a generalized overview; in practice, periods may overlap

This introduction sketches the emergence of PE in The Netherlands and links these developments to trends in other countries. It then summarizes an investigation of the utilization of evaluation research results in The Netherlands by the national and local governments. The micro and macro dimensions of PE are then elaborated. Finally, the contributions which comprise this volume are introduced and placed in perspective.

## TOWARD PROGRAM EVALUATION

In The Netherlands the pioneering period in policy-oriented social research lasted from the nineteen thirties until the fifties. During the 1930s a number of new polders (land reclaimed from the sea) were built (e.g., Haarlemmermeer, Wieringermeer). Their colonization was preceded by social research and planning. Prospective farmers were interviewed and new villages were designed according to ideas taken from sociology and

demography. When the polders had been in operation for a number of years, ex post evaluation projects were carried out. Also during the pioneering period, industrial recruiting policies were evaluated and worker absenteeism was analyzed. A small number of researchers, usually with little experience, worked on an ad hoc basis with little research methodology to go on. This picture is not very different from the experiences of other Western countries.

In the early sixties the scene changed. Economic prosperity enabled The Netherlands to embark on a number of large-scale social intervention programs which were accompanied by Social Research. Special emphasis was given to forecasts, ex post evaluations and experimental innovations. There are many examples of these interventions. The Delta Works were initiated in the South-Western part of the country, changing the regions behind the new dikes, for instance, by giving better access to main roads and railways. Large-scale reallotments changed the agricultural areas.

It was a bold period for social intervention in The Netherlands, one similar to that in the United States which saw the "War on Poverty," "Big Cities Program," and "Program Headstart." Policy-oriented social research adopted American techniques and organizational practices. In The Netherlands when the social interventions could not meet their ambitious targets, subsequent developments closely resembled those in the United States. In the political arena, in government departments and in business firms alike, policy-oriented social research lost credibility.

From the mid-seventies to the early eighties, a period of stagnation and prolonged criticism reigned. Large scale intervention was restricted to areas with little capacity for self-defense, for instance universities and other parts of higher education. Policy-oriented social research played only a restricted part. However, experience and methodological sophistication grew.

In the early eighties, in the United States, the United Kingdom and the Federal Republic of Germany, governments tried to stem the tide of economic recession by applying no nonsense politics: "deregulation," "privatisation" and "decentralization" were implemented on relatively large scales. In the beginning this approach appeared hostile to policy-oriented social research. However, the political attitude changed. The opposition was strong enough to impose a significant amount of impact assessment, and the parties in power were forced to play ball. Environmental impact assessment emerged and was institutionalized by law in 1987. Technology

Assessment followed shortly. More and more new laws contained clauses demanding *ex ante* as well as *ex post* evaluation.

In the long run, *ex ante* and *ex post* evaluations on an *ad hoc* basis proved to be unsatisfactory. A stimulation program of the Department of Education and Sciences (1985) included heavy emphasis on "program evaluation" (see Hofstee & Laros, to follow). This type of policy-oriented research was defined primarily as *ex post* evaluation. However, in an intervention program structured over time, an *ex post* evaluation of "period a" often evolves into an *ex ante* evaluation of "period b." In this way a "philosophy" of program evaluation for major interventions emerged that considers both *ex ante* and *ex post* evaluation.

#### THE USE OF PROGRAM EVALUATION

One might ask if the Dutch parliament utilizes the results of evaluation research? Candidly, although expectations sometimes are high, results are only partially used.

To investigate how the parliament deals with evaluation, a content analysis of parliamentary documents was performed (Van Driel, 1988). First, a collection of all parliamentary documents using the word "evaluatie" in the summary and appearing between July 1, 1986 and March 30, 1987 was made. There were 92 such documents.

Initially, these documents were superficially analyzed on several dimensions. In 71 cases, documents dealt with a recent evaluation: the remaining 21 used "evaluation" only superficially. In 54 cases, the initiative to evaluate had been taken by a Ministerial Office, in 31 it came from parliament, and in the remaining seven, it was unclear who took the initiative. Eighty-seven of the evaluations were *ex post*, three appeared to be *ex ante* and twice the type was unclear. Of the 87 *ex post* evaluations, 57 were goal free, effect evaluations, 27 dealt with achievement of a goal. Of these, five were cost analyses. The focus of the remaining three was unclear. In most effect evaluations, criteria were not formulated explicitly, perhaps to avoid the conclusion that the content of the program was useless. In 70 of the 92 cases, the goal could be categorized in more detail. In 54 of these 70 cases, the goal was to adjust specific policy measures. Strictly summative evaluations were absent.

The study also focussed on who performed the evaluation. This was indicated in only 57 cases. In 33, the Ministerial Office undertook the

evaluation, in 12 cases it was performed by independent investigators. In 55 cases, we could not determine how or even if the evaluation was to be used. Of the remaining 37 cases, 32 were intended to be used to adjust policy, one to gain experience, and four were intended both to gain experience and to adjust policy. Documents also were examined to determine the Ministerial Office involved. Nine of the documents involved combinations of several offices, while involvement in two was unclear. In most cases (see Table 2), the Office of Welfare, Health and Culture was involved. In general, the well known triad of health, education and welfare clearly can be seen (see also Van Vught, 1982).

Table 2: Ministerial Office with which 81 evaluations dealt

<u>OFFICE</u>	<u>N</u>
Welfare, Health and Culture	24
Social Affairs and Employment	13
Education and Sciences	9
Economic Affairs	9
Housing, Land Planning and Environment	7
Foreign Affairs	5
Finances	4
Traffic, Roads, Dikes	3
Agriculture and Fishery	3
Internal Affairs	2
Developmental Aid	1
Dutch Antilles	1

In 47 cases it could not be determined how the evaluation was placed in record, in 45 there was a report with descriptions, figures or both. Most of the 92 documents dealt only superficially with evaluation research as a scientific enterprise. In only 14 was an evaluation report analyzed in detail. Of these 14, seven were internal evaluations (self-evaluations) and seven were performed by an external person or organization. In nine cases there was an explicit governmental instruction. Two evaluations which had summative natures were internally performed.

Evaluation supposedly implies a norm to which a situation is compared; differences from the norm then can be tested for significance. Such tests apparently were performed in only eight cases. The impression given is that "evaluation" is often used quite broadly. Of the 14 cases which

could be strictly considered as evaluations, only two tested an instrument. The remaining 12 evaluations did not apply true experimental design: three were quasi-experiments, the remaining nine were of a non-experimental nature.

Of course, in many situations true experiments or even quasi-experiments are hard to design. On the other hand it can be concluded that a sincere analysis of research design possibilities often was not undertaken. This is disappointing. It is questionable to what extent the conclusions of non-experimental evaluations can deal with effects, as opposed to merely describing situations or processes, irrespective of the causes. Nevertheless, such descriptions often can be quite valuable.

Another analysis was performed of the use of evaluation as an instrument for management. This analysis was primarily focused on central and local government and the non-profit sector (Van der Leun, 1988). It was performed because the impression existed that little use is made in management of the many evaluations performed in The Netherlands. Management was perceived as independent: the implementation of policy decisions.

Reports of recent evaluations were collected in two ways. First, reports were reviewed through the documentary systems of the Universities of Utrecht and Wageningen and the report pool of the SWIDOC at Amsterdam, as well as informally through newspapers, etc. This search yielded 25 reports. Second, requests for recent evaluation reports were sent to organizations seen as likely sources. Types of organizations contacted included: Ministerial offices, provinces, municipalities, non-profit organizations, organizational advice agencies, academic institutions and industries. Ten organizations in each category were contacted. Altogether, 13 public reports (and one non public report not considered here) were obtained. Reports came from all types of institutions except industries which, unfortunately, sometimes were unwilling to cooperate. A total of 38 evaluation reports were gathered by the two methods. A survey of seven can be found in Table 3.

To a large extent, the potential use of an evaluation depends on the relationships between the three parties involved: the evaluated, the evaluator and the principal who commissions the evaluation. Five important situations can be distinguished:

1. Autonomous self-evaluation

- 2. Obligatory self-evaluation
- 3. Invited external evaluation
- 4. Autonomous external evaluation
- 5. Obligatory external evaluation

Table 3: Survey of Seven of the Evaluation Reports

1. Chernobyl:	The response of the central government to the short term effects of the accident.
2. Scholarships:	A study of how the new system of student scholarships worked?
3. Departmental Division:	The introduction of the new division of governmental offices.
4. Mental Health:	A requested self evaluation of the functioning of the national council of public health.
5. Agricultural Organization:	A self initiated evaluation of the functioning of the central agricultural organization.
6. Environmental Policy:	An account of the environmental policy of the province of North Holland.
7. Debt Sanitation:	A project to sanitize private debts in the city of Rotterdam.

These five situations reflect decreasing independence on the part of the evaluated. Less independence means a poorer situation for an investigator and probably less expertise, but also more involvement with the evaluation process. While categorizing often is difficult, all five types of situations can be identified in the reports analyzed.

The study of these reports began with the pessimistic hypothesis that not much is being done with evaluation results to change management. In fact, very little explicit use of results can be traced. Short term use seems to be exceptional. Later use is hard to trace, but is unlikely. Evaluation reports often are carefully stored and forgotten. Evaluation seems to be performed as a kind of side line, separate from organization management. Results can be satisfactory or unsatisfactory, interesting or uninteresting . . . managers note the results and return to business as usual. The "enlightenment function" is mentioned as a worthwhile use of evaluation research. From the analysis reported here, this kind of use could be neither clearly established nor excluded. The general conclusion is that evaluation research may potentially be a useful management instrument, but in a wide variety of situations it is poorly used as such.

#### PROGRAM EVALUATION ON MICRO AND MACRO SCALES

Program evaluation seeks to investigate a highly varied set of problems. One important variable is the size of the problem under investigation. The problem varies from micro to macro problems (e.g., from an individual trying to solve an alcohol problem to a government department stimulating a campaign to curb excessive alcohol consumption). In The Netherlands, as in other countries, micro problems and interventions have been evaluated primarily by psychologists and related behavioural scientists; macro problems and interventions by sociologists and other social scientists.

In an intervention directed at a micro-scale problem, the unit of analysis is usually the individual. Here, the aim of the program evaluation is to generalize. The assumption underlying generalization is that the only variability among units is random. However, variability systematically related to other variables distinguishing the units of analysis produces generalization. Generalizations of several kinds can be made. For example, finding out if an alcoholic solved his or her problem, as part of a project of evaluation, derives its importance from generalization to: the addiction program he or she participated in, similar types of programs, addiction programs in general, programs in different countries, cultures, or periods. While such generalizations are not without problems, the underlying assumption is that the human constancies or continuities are such that generalizing is possible within acceptable ranges of error.

The unit of analysis for macro-problems and interventions may be the individual, but more often it will be a group. Consider, for example, a

campaign to curb excessive drinking. Not only individual drinkers but also their families, the members of their social networks, their employers and the police force are part of the target population. Informal social groups and organizations as well as the individual are at stake. This introduces relatively long chains of events. Such chains are characterized by relatively high "system noise" as soon as social research is introduced. Generalization is relatively difficult in such a research setting. As a result, cooperation between behavioural and social scientists has become an issue of concern in program evaluation. This issue has been handled in a number of ways, as other contributions to this volume illustrate.

### AN OVERVIEW OF THE VOLUME

In this section a brief overview of the papers appearing in this volume is given.

In 1986, the Dutch Minister of Education and Science established a Committee on Program Evaluation to advise him as well as to stimulate program evaluation (PE). Hofstee and Laros sketch the contours of the Committee's activities in the Dutch setting.

A biotechnology innovation program is used by Nederhoff to describe a longitudinal approach that utilized five data collection methods in the evaluation of a government program. Five other Western industrial nations are used as a comparison group .

De Brey field tested the standards of the (U.S.) Evaluation Research Society (ERS) in The Netherlands by having workers in the mental health setting judge them. He concludes that there are several problems of application, all related to the lack of internal consistency of the standards and their strong emphasis on a quantitative approach.

The admission to, path through, and outflow from the Jellinek Center (10 clinics of different types dealing with alcohol and drug problems) are analyzed by Nabitz and De Gelder. They conclude that the clinics function too independently. As a result, the center functions more as a revolving door than as an integrated network of facilities.

Evaluation Research (ER) focused on intervention is described by Huisman and Koch in an experiment on a social skills training. They show that the instrument they developed was useful to clients, giving them more insight into training objectives.

An analysis of the Dutch Educational Support System is reported

by Koster, Tillema and Van Batenburg. They examined the utilization of materials, activities, and projects offered by several support agencies. Perceived usefulness and transfer were found to be the variables that explained most variance. Need for the activities and variations between schools did not seem to influence the use of programs and materials provided by the agencies.

Visser, Breemhaar and Kleijnen investigated social desirability bias in responses to patient questionnaires. They show that social desirability is stronger when patients are questioned orally during their stay in the hospital. Social desirability is found to be related to demographic variables as well as to patient satisfaction.

Dekker and Leeuw argue that the gap between the several models of Program Evaluation and Effectiveness Auditing (EA) should be narrowed. They point out the similarities as well as the differences between various models. While both PE and EA deal with the process investigated, EA is more closely coupled to the goals and requirements of the national government.

Faludi makes a distinction between project plans and strategic plans: strategic plans guide project planning and related decisionmaking. Using two evaluations he discusses the consequences of this distinction for evaluation, with the goal of making better *ex ante* assessments of strategic plans.

The consequences of response-shift bias for program evaluation is discussed by Sprangers. She shows that in communication skills training program not only produced the desired effect, but also a response shift due to initial overestimation of students' functional level.

Weenig and Midden evaluated the effects of vertical diffusion networks as contrasted to horizontal social networks in information programs on energy conservation. Top-down and bottom-up analyses were used to deliver additional information in the applied quasi-experimental design with pre and post measures.

The "multi-level perspective" in PE is described by Van Eeden. Here "within-group" outcomes of the analysis of individual processes were introduced into the "between-group" analysis. Van Eeden demonstrates this procedure in an analysis of pupils and school types. Schools are found to differ to the extent they transfer a pupil's intelligence into his or her career score.

Taken as a whole, these papers reflect the present state of program evaluation research in The Netherlands.

## References

- Becker, H.A., 1986. "Towards Program Evaluation" in H.A. Becker and A.H. Porter (eds.) *Program Evaluation Today*. Utrecht, Van Arkel.
  - Cronbach, K.J., et al, 1980. "Toward Reform of Program Evaluation." San Francisco, Josse-Bass.
  - Driel, J.M. van 1988. Evaluatie in het Nederlandse Parlement (Evaluation in the Parliament of The Netherlands). In Dutch. Odijk: Bureau Odijk.
  - Leun, A.A.J. van der 1988. Evaluatie als managementinstrument (Evaluation as a Management Instrument). In Dutch. Odijk: Bureau Odijk.
  - Raizen, S.A. and P. Rossi (eds.) 1981. *Program Evaluation in Education*. Washington, National Academy Press.
  - Schnabel, P. 1984. Programma evaluatie: een nadere kennismaking bij het scheiden van de markt (Program Evaluation, a late introduction). In: G.M. Schippers, N.Sijben, C. Schaap & J. Mastboom (red.), *Programma-evaluatie in de geestelijke gezondheidszorg: een conferentieverlag*. In Dutch. Lisse: Swets en Zeitlinger, 218-230.
  - Vught, F.A. van 1982. *Experimentele beleidsplanning (Experimental policy planning)*. In Dutch 's-Gravenhage, VUGA. □
-