# Effectief evalueren. Een empirische studie naar de doelmatigheid van aanwijzingen voor de evaluatiepraktijk

Berg, Gerard van den

SUMMARY

The present study is a report of two research projects whose explicit goal is supplying empirically based recommendations for improving the professional practice of evaluation researchers. Since evaluation research should be subjected to demands of both pratical relevance and scientific quality the study is concerned with both aspects. The first project concerned a meta-evaluation of fourteen curriculum evaluation projects, the second a summative curriculum evaluation.

The research question of the first of the two projects may be described as a reconstruction of the discussion about the relation between evaluation and decision making, and concentrates on: (1) the empirical formulation of criteria for the actual use of the results from curriculum evaluation research, (2) the way in which evaluation researchers do research, (3) the context of curriculum development.

The research questions of the second project may be summarized as follows: (1) which criteria should be used to determine the quality of evaluation research? (2) how can these quality criteria be realized in the practice of evaluation research? and (3) to what degree is the competition model a satisfactory heuristic for solving the quality problems?

The first project investigates the influence of nine factors on the use of evaluation results, viz.:

1. the evaluation approach;
2. the quality of the information reported;
3. the type of information reported;
4. the professional qualities of the evaluator;
5. the influence of decision makers on decisions concerning the research;
6. the quality demands, specified beforehand, on the information reported;
7. the nature of the information need;
8. the specificity of the evaluation questions;
9. the type of decision making.

Use is regarded as an important criterion for the practical relevance of evaluation research. Three main types of use are distinguished in the literature: instrumental, conceptual and persuasive use.

Instrumental use is defined as direct and immediate use of specific research information (data, results, implications) with a certain policy issue. Instrumental use implies there is application-oriented knowledge.

Conceptual use is defined as affecting the way of thinking of policy makers. A broad range of research results is used as well as concepts, ideas, notions, etc. from the social sciences. Here there is use of research information in a diffuse and indirect way, often on a longer term. Since this influence works on a longer term it is more difficult to distinguish from other influences.

Persuasive use is defined as the use made of research information for supporting one's own opinion in the discussion with others.

Since the research was limited to formative curriculum evaluation and thus conducted within the context of curriculum development--in which

developers need information about the relative success of the as yet experimental (or parts of it)--instrumental use is a valid criterion for practical relevance.

The nine factors which influence the use of evaluation results were chosen on the basis of an extensive study of the literature.
To describe the evaluation approach within the fourteen projects evaluation research is seen as an activity conducted by rules; a distinction between technical and conformative rules is made, with the following preferences for the collection and analysis of data as operational characteristics.
Within a more technical system of rules there is a preference for data collection through external, systematic and structured observation techniques, standardized and closed interviews, and questionnaires. Consequently, this preference also leads to a preference for quantitative processing and analysis of data. Within a more conformative system of rules more intensive methods of data collection are preferred, such as participatory observation and open interviews, and also the qualitative processing and analysis of data.
The quality of the information reported is operationalized as the degree to which attention was paid to the quality of the information in terms of reliability, validity, degree of recognition, and/or practical intersubjectivity, when results were reported. This operationalization is chosen because many researchers neglected to include in their reports sufficient information about the methods and techniques they had used. This made it virtually impossible to determine the quality of the information reported. Therefore it was only determined whether the reports did or did not pay attention to one or more quality criteria. Whether or not these criteria were satisfied was not checked.
The type of information reported is also operationalized as a dichotomy: only describtive/evaluative information that, or both this kind of information but also explanative/predictive information.
The professional quality of the evaluator has been operationalized as the degree to which the evaluator has had (1) a specific methodological training, and (2) a reasonable amount of research experience.
The influence of decision makers was determined by asking the project leaders what influence the decision makers had had on the research design, the development of instruments, and the collecting, processing and analysis of data.
Since the quality of the information reported is often used by decision makers as an argument for not using evaluative information the developers were asked which criteria for the quality of the information they had formulated beforehand.
The developers were asked for the information need(s) they had formulated beforehand. There can only be a responsible assessment of the information to be collected if there is a clear and concretely formulated decision problem. Only then a correct transformation of the information need into a concrete evaluation question is warranted.
In order to assess whether there is or is not a clear, concretely elaborated evaluation question various project documents were con-

sulted. In all cases, the evaluation questions identified were judged by two experts on the degree of elaboration. In addition the evaluators were interviewed to get a better understanding of the precise research question.
The decision making process in the curriculum development projects is described as either more rational of more incremental. Three indicators have been used: (1) the degree to which a problem was diagnosed, (2) the degree to which the decision problem is clearly formulated, and (3) the degree to which the information need is clearly formulated. The first indicator refers to the decision making procedure; the other two refer to the product of the decision making. The decisive feature for distinguishing rational from incremental decision making is the degree to which the possible means/actions that can be committed are compared to each other. Since this feature is an essential element in the diagnosis of the problem it is not used as a separate, fourth, indicator for the type of decision making, in terms of rational versus incremental. To assess the type of decision making various project documents were analyzed. To supplement these data interviews were held with project leaders.

The main conclusions from the research are that a reasonable degree of documented, instrumental use of evaluation results can be predicted when evaluation research is conducted within a rational decision making context; when, in addition, a concretely elaborated evaluation question is the starting point, and when attention is paid to quality criteria of validity and reliability.
From the perspective of use the research results give little or no support to the notion that within an incremental decision making context a conformative evaluation approach should be chosen. Such a context gives great problems to the evaluation researcher if he is to conduct the research in such a way that it leads to a use of results as intended by both decision makers and evaluators. The best guarantee for such use seems to be a technical evaluation approach. The quality of evaluation research also benefits greatly from a technical approach.
However, some comments must be made about these conclusions. Firstly, the study has methodological limitations. One of them is the small sample, another the absence of well-developed theories (so that it was not possible to focus on a small number of crucial differences of opinion). Because of the absence of such theories it was not possible to use a hypothesis testing design, in which crucial variables could have been manipulated (at least ex post facto in the selection of objects of research). The research had to be conducted in a naturalistic setting, with all kinds of natural variations. However, an attempt was made to conduct the research as critically as possible by: (1) a most careful theoretical analysis of the factors, mentioned in the discussion, that influence use, (2) subsequently formulating these factors in the form of hypotheses, and (3) controlling for these factors in a careful elaboration analysis. Despite the attempt to avoid post hoc explanations the methodological limitations force us to be cautious in drawing far-reaching conclusions.
Secondly, there are contextual limitations. The research was limited

to evaluation projects in the area of curriculum development; projects finished between 1978 and 1982. However, there are hardly plausible arguments for supposing that an extension of the research with evaluation projects from other contexts or with projects finished before 1978 or after 1982 would have had consequences for the conclusions of the present study.

Although the use of evaluation results is an important effect criterion for good and relevant evaluation research the quality of the information reported is another important, albeit conditional, criterion. The quality of evaluation research depends on the degree to which solutions are found to (1) the criterion problem, (2) the standard problem, (3) the problem of choosing the most adequate research design, and (4) the problems concerning the control of interfering circumstances. These problems are the concern of the second research project. The main object of curriculum evaluation is that the attainment of a well-defined goal (usually a learning effect) should be determined adequately. This means that a measuring instrument should be sufficiently reliable and valid for measuring what a pupil has learnt of the subject matter dealt with in the curriculum.
An adequate formulation of the criterion is therefore the first necessary condition for good evaluation research. In this area curriculum developers are answerable, i.e. in theory they should supply evaluation researchers with the necessary criteria. However, the practice of curriculum evaluation is often different. The formulation of standards should also be part of the responsibility of curriculum developers, i.e. curriculum developers should indicate which values of a criterion variable do and which values do not show an intended effect. We should realize that any evaluation standard is dependent on goals, values and norms of people (in this case developers). Judgments will be different according to the standards applied. This means that after deciding on the effect criteria to be used the norms for the 'success' of the curriculum should be specified. Here important critical questions are: (1) when has a pupil sufficiently benefited from it? and (2) how many pupils have sufficiently benefited? In order to get a grip on this one needs a yard-stick, a standard.
Starting from the idea that in evaluation research the professional reputation of curriculum developers is at stake, Hofstee advocates designing evaluation research as a bet, i.e. research in which a difference of opinion is the starting point. This implies however that a difference of opinion can only be resolved in an acceptable way when there is the willingness to make a joint effort to settle this dispute. For this rules must be developed which all parties accept since parties who make assertions without being able to appeal to sources recognized by the other party and also unwilling to defend these assertions in another way do not contribute to a settlement. A settlement of the dispute which is acceptable to both parties can only be reached when both pro-arguments (attempts to justify one's own position) as well as counter-arguments (attempts to weaken the other party's position) may be given and when the parties have agreed upon a common test procedure.

Although the betting model was a source of inspiration for the design of the second research project the model actually used is more complicated. The betting model is developed for situations in which only the 'one-shot experiment' can be done. The assumption underlying the betting model is that there are differences of opinion about the effectiveness of one treatment. In the second project there were several 'treatments'. These 'treatments' may be seen as the various answers to existing differences of opinion about the way education should be structured. The research is designed as a comparative product evaluation. To determine the relative effectiveness of the curricula adequately the betting model is combined with the classical model statistical hypothesis testing. Hofstee called this model the competition model. In the research project there was hardly any problem in achieving commitment with regards to the research design and interfering circumstances. Solving the criterion and standard problems takes up much time and makes specific demands upon the evaluator, demands on his social skills as well as his professional expertise. Despite the limitations in time and money the experiences with the model used may be considered positive. The competition model is a negotiation model and, as the research established, suitable for determining the relative effectiveness of the curricula. The model emphasizes the realization of first and foremost commitment with regards to the criteria to be used. An important condition for achieving commitment is the parties' willingness to make a serious effort to settle their dispute. Naturally commitment cannot be forced by just using a procedure. There is more to it than that. If there is no real willingness then the competition model will not solve the problem. To determine the relative effectiveness of the curricula the analysis of variance is a suitable technique. The level of statistical significance ($p < .01$) is an arguable relative standard. In order to be able to draw conclusions about the relative effectiveness of the curricula, the research was planned to take interfering variables into account; otherwise, the absence of data for these variables would have resulted in unavoidable problems of interpretation when the question was asked which conditions are for the most favourable effectiveness of the curricula. There is a number of interfering variables which curriculum evaluation research cannot afford to ignore. These variables are: (1) the pupils' initial condition, (2) characteristics of curriculum implementation, and (3) the actual amount of time the pupils spend on learning.
The importance of controlling for the effects of the interfering variables mentioned is corroborated by this study. Although the differences in student achievements between the curricula were statistically significant these differences could not be ascribed to the differences between the curricula. The benefit of predictions in comparative curriculum evaluation is limited. These predictions are only suitable for determining the pretensions of the curriculum developers.



LITERATUUR

Albinski, M.: Onderzo
tenschappelijk on
1978.
Alkin, M.C.: Evaluati
Sanders: Educati
1973, p. 100-156.
Alkin, M.C., R. Daill
tion make a diffe
Becker, H.A.: Twee p
ciaal Beleidsonde
Berg, G. van den: Onc
onderzoeksmodel.
Berg, G. van den en M
nieuwingen: tuss
Schoolorganisatie
Berg, G. van den, P.
luatiegegevens in
In: R. Halkes en
1982, 93-106.
Berg, G. van den, P.
luatie-onderzoek:
in curriculumontw
en J. Vastenhouw,
wijsleerprocessen
Berg, G. van den en W
search and the u
tional Evaluation
Berg, G. van den: Cu
studie naar de
Haren, RION, 1983
Berg, G. van den en E
MAVO vergeleken (
Groningen, RION,
Berg, G. van den, F.B
onderwijsleerpakk
project 4213). Gr
Blumer, H: Methodolo
Denzin, Sociologi
Bulmer, M.: Applied s
'englightment' mo
209.
Caplan, N., Morrisson
knowledge in poli
Circourel, A.: Method
Clark, P.A.: Actie-or
Rijn/Brussel, 197
Cohen, M.A., March, J
nizational choice
1972, p. 1-25.