Re-thinking evaluation: Notes from the Cambridge Conference

BARRY MACDONALD and
MALCOLM PARLETT

In December 1972 the authors convened a small working conference of people concerned with educational evaluation.[1] The meeting took place at Churchill College, Cambridge, and was financed by the Nuffield Foundation. Its aim was to explore 'non-traditional' modes of evaluation and to set out guidelines for future developments in this field. Participants were chosen for their known reserve about established evaluation practice, or because they had suggested or experimented with new approaches.[2] The conference arose out of preliminary talks between the authors and the officials of the Nuffield Foundation, the Department of Education and Science, and the Centre for Educational Research and Innovation in Paris. These talks had reflected, on the part of these agencies, a general concern that the rapid increase of evaluation activities was not being accompanied by an equivalent surge of new thinking about either evaluation methods or their usefulness for decision-making.

First, a word about the traditional style of evaluation whose continued predominance the conference set out to challenge. Five years ago, at the inaugural meeting of evaluators of Schools Council curriculum projects, the chairman opened the proceedings with the remark: 'Can I assume that we're all familiar with Bloom's Taxonomy of Educational Objectives?' Nobody laughed. The assumption was reasonable, the issue crucial to the matters in hand. Most of those present, novices in the field, shared the chairman's implied view that

Barry MacDonald is head of the evaluation unit at the Centre for Applied Research in Education, University of East Anglia.
Malcolm Parlett is at the Centre for Research in the Educational Sciences at the University of Edinburgh.

the task of the evaluator was to determine the congruence of pupil performance and project objectives, i.e. to assess the extent to which pupils exposed to a new curriculum achieved its intended learning outcomes. They were further agreed that the assessment procedures incorporated should be de-personalised and preferably psychometric in form. What could be more central to such a task than the Bloom/Krathwohl formulations of curriculum objectives in terms of measurable learned behaviour!

This model of evaluation stemmed from a long established and securely rooted tradition of educational measurement on both sides of the Atlantic. It had been a dominant influence in American education since its prototype was launched by Ralph Tyler in the early 1930s.[3] In the sixties, with the influx of massive federal finance for curriculum improvement, the model came into its own. Federal policy-makers demanded of educational innovators that they both pre-specified the intended performance gains and provided subsequent proof of 'pay-off'. Despite mounting criticisms of the engineering-type assumptions of such 'pre-ordinate' evaluation,[4] and the tentative emergence of alternative approaches,[5] the model was still serviceable enough to be exported to us, and to command the initial allegiance of the first wave of professional evaluators in Britain. It was endorsed by our doyens of educational research, and found consistent with the influential pre-occupations of prominent curriculum philosophers. But it didn't work. Most of these evaluators, having experienced the conceptual and practical inadequacies of the model, are still licking their wounds. They found, for instance, that not all the curriculum developers shared the model's assumptions about the essential need to clarify goals in advance. Even with defined objectives the problems of measurement were far from tractable—even less so in practice than they were in theory. They discovered, too, that the task of evaluation seemed to call for a greater variety of research skills and to raise more complicated issues, than those involved in educational measurement.[6]

In America, in the last few years, disillusion at the apparent failure of the curriculum reform movement has sharpened criticism of its assumptions and techniques, and those of evaluation; and has evoked a greater willingness to face the daunting complexity of educational realities. The optimistic rationalism which shaped and sustained the movement has been muted, its inherent assumption of value-consensus exposed as an hallucination. In a paper written for

the conference one of the American participants, Robert Stake, pinpointed some of the weaknesses of the evaluation mode which characterised the movement:

> 'It is likely to be underfunded, understaffed, and initiated too late. But even under optimum conditions it often will fail. A collection of specific objectives will understate educational purposes. Different people have different purposes. Side effects—good ones and bad—get ignored. Program background, conditions, transactions are likely to be poorly described. Standardised tests seldom match objectives, criterion referenced tests oversimplify and fail to measure transfer, and custom-built tests are poorly validated. And people cannot read many of the reports or do not find them useful'.

This brief historical overview describes some of the background dissatisfaction that prompted the conference. The question uppermost was where do we go from here? One way ahead was indicated in 1969 by Tom Hastings in his presidential remarks to the American National Council for Measurement in Education, when he made a plea for psychometrists to join forces with historians, economists, sociologists, and anthropologists in a concerted attack on the problems of educational evaluation.[7] In a way the Cambridge conference was an attempt to open up evaluation thinking in this direction, seeing what could be learned from other research standpoints, and assessing the implications, difficulties, and potentialities of new evaluation styles. As convenors of the conference, we were particularly concerned to address what we judged to be one crucial problem for new evaluation designs—namely, the fact that educational decision-makers are forced by the complexity of required decisions to review a much wider range of evidence than has usually been gathered in evaluation research. In Ernest House's phrase, the decision-maker's 'vocabulary of action' is not matched by the current lexicons of evaluators.[8]

The second day of the conference was devoted to intensive review of three evaluation reports produced by participants.[9] In each study the evaluator had rejected the 'agricultural-botany' assumptions of the classical model in favour of an approach rooted more in sociology and social anthropology. The three reports—though far from uniform in style and aims—nevertheless had major features in common. For instance, each evaluation (i) featured naturalistic, process-oriented

field studies of educational experiments which attempted to portray the innovation in the context of a recognisable social reality; (ii) documented a full range of phenomena, perspectives, and judgements that might have relevance for different audiences and diverse interest groups; (iii) utilised observational and interview techniques extensively, and gave less than usual prominance to measurement procedures; (iv) followed a flexible rather than pre-determined research plan, thus enabling the investigation to be responsive to unpredictable and atypical phenomena and to sudden changes of direction in the form of the experiment, as well as to the planned and to the typical.

There was general agreement that such studies—more extensive, naturalistic, and adaptable than evaluation designs in the past—represented the best and possibly only way of bridging the gap between evaluation research and the more practical and down-to-earth concerns of practitioners and local and government officials. Against this background of consensus, however, were many points of discussion and dispute. Clearly it is impossible to summarise all of these here. But three recurrent themes of discussion deserve mention.

The first theme centred on methodology. For those in educational research who revere the canons of tradition, the 'anthropological' style of evaluation raises the twin spectres of subjectivism and a fatal lack of discipline. Conference participants pointed out, first, that the intrusion of the researcher's own values is equally a problem in traditional-type evaluations, though often effectively concealed amid numerical analyses; second, that to permit methodological considerations to determine what is significant and relevant research to do, is rather like confining a search to the area under the lamp post because that's where the light is; and, third, that other disciplines (e.g. history and anthropology) have standards of what constitutes valid and reliable evidence, that are generally considered reasonable given the nature of their subject-matter, and which might well be appropriate to adopt in the evaluation field. Over other methodological issues there was clearly less agreement: for instance about how soon and how much should an evaluation researcher focus his enquiries; and about whether objective testing, as a mode of inquiry, is compatible with the maintenance of the good personal relations necessary for informal observation and interviewing.

A second group of issues related not so much to detailed tactics as to broader research strategy. For example, in setting up a study,

should the evaluator have a 'contract' with his sponsors, or not? In favour was the argument that establishing the scope and basic form of the study in advance, lessened the possibility of subsequent misunderstandings. The view against the contract was that 'it stagnates, it ritualises, it forces things, just like pre-ordinate designs and experimental method'. Another extremely tricky question was whether an evaluator should intervene if he sees a programme going 'wildly off the rails'. Morally, he perhaps should—even if he is summative 'rather than formative'—but if he does so, an opportunity may be lost for seeing how the innovation or scheme stands up to particular types of acute strain. There were questions, too, of how evaluation studies should be reported. Should they constitute a 'display', in 'raw' form, of the range of different opinions, results from questionnaires, and so on? Or, alternatively, should the report distil, summarise, organise, and interpret the data for its different audiences? The argument for the former approach was that readers should be given the opportunity to judge the scheme reported for themselves; and that structuring the report could inevitably influence the reader in one way or another. It is the reader's task to 'evaluate', in the literal sense of the concept, and the evaluator's task to provide the reader with information which he may wish to take into account in forming his judgement. The arguments against this, and in favour of a more interpretive treatment, were that 'straight narrative reporting' is probably a misnomer anyway; that without distillation the report is likely to be long or indigestible for readers; and that a focussed, more analytic treatment is necessary in order to contribute to a deeper understanding of the phenomena reported.

This issue related to the third area of discussion, namely the role of the evaluator *vis-à-vis* the decision-maker. Should the evaluator be 'subservient' or 'dominant'? Should he be merely providing information along lines requested, or should he be an independent, challenging, and critical figure, who introduces new notions from his expert position with which to confront decision-makers? There was a range of opinion, from those who argued that the evaluator should 'contain his arrogance', reflect 'society's values', and fit his studies closely to the needs of policy-makers; to those who thought there was an obligation to 'educate decision-makers', if not to a particular point of view, at least into a more sophisticated way of examining educational issues.

These represent a handful of the topics discussed during the four-

day conference, which also included—as an exercise—devising evaluation designs for an impending British curriculum project; and review of possible future plans for developing non-traditional evaluation in Britain.

At the end of the conference, the participants decided that it might be useful to make available an agreed summary of their conclusions, drawing attention to significant issues which still divided them. The statement, or manifesto, which is still in a draft form, reads as follows:

'On December 20, 1972 at Churchill College, Cambridge, the following conference participants[10] concluded a discussion of the aims and procedures of evaluating educational practices and agreed

I. That past efforts to evaluate these practices have, on the whole, not adequately served the needs of those who require evidence of the effects of such practices, because of:
   (*a*) an under-attention to educational processes including those of the learning milieu;
   (*b*) an over-attention to psychometrically measurable changes in student behaviour (that to an extent represent the outcomes of the practice, but which are a misleading oversimplification of the complex changes that occur in students); and
   (*c*) the existence of an educational research climate that rewards accuracy of measurement and generality of theory but overlooks both mismatch between school problems and research issues and tolerates ineffective communication between researchers and those outside the research community.

II. They also agreed that future efforts to evaluate these practices be designed so as to be:
   (*a*) responsive to the needs and perspectives of differing audiences;
   (*b*) illuminative of the complex organisational, teaching and learning processes at issue;
   (*c*) relevant to public and professional decisions forthcoming; and
   (*d*) reported in language which is accessible to their audiences.

III. More specifically they recommended that, increasingly,

(*a*) observational data, carefully validated, be used (sometimes in substitute for data from questioning and testing);

(*b*) the evaluation be designed so as to be flexible enough to allow for response to unanticipated events (progressive focussing rather than pre-ordinate design); and that

(*c*) the value positions of the evaluator, whether highlighted or constrained by the design, be made evident to the sponsors and audiences of the evaluation.

IV. Though without consenus on the issues themselves, it was agreed that considered attention by those who design evaluation studies should be given to such issues as the following:

(*a*) the sometimes conflicting roles of the same evaluator as expert, scientist, guide, and teacher of decision-makers on the one hand, and as technical specialist, employee, and servant of decision-makers on the other;

(*b*) the degree to which the evaluator, his sponsors, and his subjects, should specify in advance the limits of enquiry, the circulation of findings, and such matters as may become controversial later;

(*c*) the advantages and disadvantages of intervening in educational practices for the purpose of gathering data or of controlling the variability of certain features in order to increase the generalisability of the findings;

(*d*) the complexity of educational decisions which, as a matter of rule, have political, social and economic implications; and the responsibility that the evaluator may or may not have for exploring these implications;

(*e*) the degree to which the evaluator should interpret his observations rather than leave them for different audiences to interpret

It was acknowledged that different evaluation designs will serve different purposes and that even for a single educational programme many different designs could be used'.

1 The conference participants were as follows:
Mike Atkins, University of Illinois
John Banks, Department of Education and Science
Tony Becher, Nuffield Foundation
Alan Gibson, Department of Education and Science
David Hamilton, University of Glasgow
David Jenkins, Open University
Barry MacDonald, University of East Anglia
Tim McMullen, Centre for Research and Innovation in Education, OECD
Malcolm Parlett, University of Edinburgh
Louis Smith, Washington University of St Louis
Robert Stake, University of Illinois
David Tawney, University of Keele
Kim Taylor, Centre for Research and Innovation in Education, OECD
Erik Wallin, Pedagogiska Institutionen, Goteborg

2 See, for instance,
MacDonald, B., 'Briefing Decision Makers: an evaluation of the Humanities Curriculum Project', *Schools Council, Bulletin of Evaluation Studies* (in press).
Parlett, M., and Hamilton, D., 'Evaluation as Illumination: a new approach to the study of Innovatory Programs', *Occasional Paper* 9, *Centre for Research in the Educational Sciences*, University of Edinburgh, 1972.
Smith, L. M., and Geoffrey, W., *The Complexities of an Urban Classroom*, Holt, Rinehart and Winston, 1968.
Stake, R., 'The Countenance of Educational Evaluation', *Teachers College Record, Vol.* 68, 1967, pp. 523–540.
Taylor, L. C., *Resources for Learning*, Penguin Books, 1971.

3 Tyler, R. W., *Constructing Achievement Tests*, Ohio State University, 1934

4 See, particularly,
Atkin, J. Myron, 'Some Evaluation Problems in a Course Content Improvement Project', *Journal of Research in Science Teaching, Vol. I*, pp. 129–132, 1963.
Eisner, Elliot, W., 'Educational Objectives: Help or Hindrance', *The School Review, LXXV* (Autumn 1967), pp. 250–260.

5 Stake, R., op. cit.
Stufflebeam, Daniel L., '*Evaluation as Enlightenment for Decision Making*', an address delivered at the Working Conference on Assessment Theory, sponsored by ASCD Sarasota, Fla., January 1968.

6 Experience was mixed. For an overview of the work of Schools Council evaluators see Bulletin of Evaluation Studies, Schools Council, 1973. (In press). This includes an account of a successful operationalisation of the objectives model by the evaluator of the Science 5–13 Project, Win Harlen.

7 Hastings, J. Thomas, 'The Kith and Kin of Educational Measurers', *Journal of Educational Measurement,* Vol. 6, No. 3, Autumn 1969.

8 House, Ernest R., '*The Conscieuce of Educational Evaluation*', paper originally presented to the Ninth Annual Conference of the California Association for the Gifted. February 1972, Col. 73, No. 3.

9 Stake, R. and Gjerde, C., '*An Evaluation of TCITY: the Twin City Institute for Talented Youth*', CIRCE, University of Illinois at Urbana-Champaign, 1971.
MacDonald, B., A draft section from forthcoming report on the Schools Council Humanities Curriculum Project.
Parlett, M. R., '*A study of two experimental undergraduate programs at the Massachusetts Institute of Technology*' (Unpublished Report), 1971.

10 It was intended that the statement, after further drafting, should be signed by all participants (except those representing the Department of Education and Science).

Public Examinations and Pupils' Rights

WILLIAM GUY AND PETER CHAMBERS

Traditional approaches to examinations in the British schools system follow the widespread belief that it is a privilege for pupils to take public examinations. It is widely held that the possession of a General Certificate of Education is a passport to improved life and career opportunities and that, therefore, the more pupils who follow courses leading to such qualifications, the greater the spread of privilege. This belief seems capable of surviving even the strongest recent attacks on examinations as being elitist and socially divisive and the psychological evidence of their inefficiency.[1] The use of examinations as a motivating factor has been questioned, the way they can dominate the curriculum of schools deplored, and the restraints they impose on educational innovation lamented. Nevertheless, the received wisdom of education still places examinations high on its value system and encourages approaches that are designed to extend the opportunities for more pupils to sit public examinations.

Yet there is another case. When teachers enter pupils for public examinations it is arguable they are disregarding a certain right that should be accorded to pupils. Certainly some teachers have misgivings about their part in this process. In this paper an attempt will be made to formulate this right and to describe those features of examinations which lead to the feelings of unease experienced by some teachers.

The right, to put it briefly, is the right to decline to sit certain examinations without social or economic disadvantages ensuing. To make this right more explicit, reference will be made to the distinction between criterion-referenced and norm-referenced tests. This

William Guy teaches at Alderman Jacobs School, Whittlesey, Cambridgeshire and Peter Chambers is Head of the Education Department at West Midlands College of Education