FORMATIVE EVALUATION

Professor Wynne Harlen

After graduating in physics at Oxford, Wynne Harlen taught in schools and colleges for a number of years. She is now Professor of Science Education at Liverpool University with particular interest in encouraging in schools the kind of learning where processes of science are used to develop conceptual understanding and scientific attitudes. Studying the ways children use ideas gained from out of school experience in more formal learning settings led to an interest in the sources of these ideas and hence in applying evaluation methods to interactive centres.

As the word 'formative' in its title indicates, this article is concerned with the evaluation which has a role in helping to create or improve an exhibit. That role can be contrasted with the role of *summative* evaluation which is to provide information about how a finished product is fulfilling its intended function, giving value for money, or how it compares on certain measures with other products. The formative/summative distinction is not clear cut, since the information from a summative evaluation can always be used in the development of further products. However, in the new field of interactive technology centres, our concern is chiefly with formative evaluation.

What is attempted here is a brief outline of the process of evaluation and a discussion of its components as applied to the evaluation of interactive technology exhibits. Some examples are picked out from the small amount of experience to date, but it has to be recognised that the current 'state of the art' is immature and it is hoped that developments now underway will soon provide case studies to provide a firmer base for theoretical discussions in this field.

Stages in evaluation

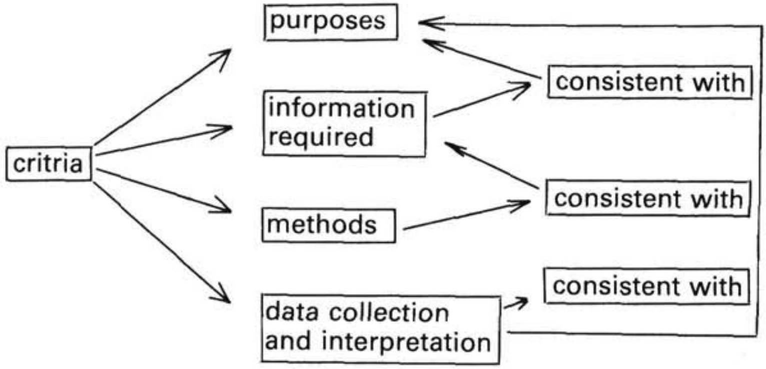
The nature of the evaluation process is that it leads to decisions or judgements based on:

- information from relevant enquiries
- criteria to be used in making judgements

Obtaining the information begins with deciding what information is required. The next steps are to find appropriate and efficient methods of obtaining the information and then to proceed to collecting and interpreting it. It is evident that the criteria are applied at the end of this series of steps, when judging how well the findings match up to them, but it is not just in this final step that the criteria have their effect. Without criteria in mind it would be impossible to decide what information is required and whether possible methods would yield it in a form suitable for a basis for decision. Thus evaluation criteria play a part from the start and their importance in the whole process cannot be overemphasised.

To illustrate this point, imagine two different sets of criteria applying to interactive exhibits – both are extremes and neither is to be defended in reality. One set states that the exhibits should result in the maximum learning for the maximum number of people regardless of how this happens; the other that the exhibits should engage visitors in physical activity of some sort, regardless of what effect this may have on their learning. An evaluation designed so that the first set could be applied would ensure that information was gained about the difference, before and after visiting the exhibition, in people's knowledge about certain things relating to the exhibits, and it might not be thought necessary to collect any information about interaction with the exhibits. On the other hand, the information required to apply the second set of criteria would be almost the reverse – it would have to concern interaction and would neglect learning. Moreover, the data obtained for such an enquiry would not allow those who wanted to, to make judgements on the basis of learning. Neither would a learning-focused evaluation allow decisions to be influenced by information about interaction. Although these examples are only for the sake of illustration, it is not too difficult to find very similar instances in the museum evaluations literature.The moral is that much attention shuld be given to the criteria before an evaluation is begun. It involves difficult thinking, since it hits right at the heart of why we set up interactive technology centres at all.

So the criteria, the bases for making evaluative judgements or decisions, influence all stages of an evaluation, as summarised in the following figure.



As we have already discussed criteria, the other constituents of the process will now be considered.

Purposes

Within the general focus of formative evaluation there are two main sub-sets of concerns.

General: to arrive at general statements which have application in designing and adapting a whole range of exhibits. For example, about how to include a strong experimental element, about providing a point to the exhibit which is clear to visitors, about the type and placing of labels.

Specific: to provide information which will help to improve the features of a particular exhibit. For example, about how to improve the safety of a particular exhibit. about the effect of position of an exhibit on the kind of interaction it invites.

These are not totally separate; usually an evaluation designed to provide specific data will yield some generally applicable statements from the combination of studies of several specific exhibits. However, they can be incompatible when it comes to deciding how to allocate evaluation resources. Intense study of one exhibit is time-consuming and cannot usually be repeated in the same depth for 20 or 30 exhibits. A compromise is necessary if the intention is to pursue both general and specific goals to some degree.

A useful approach to achieving a compromise is to identify certain issues which apply to exhibits which impinge on these issues. One such issue is the extent to which an exhibit has to be made to represent a 'real' object, event or situation. A dimension from 'real' to 'wholly representational' can be defined, as follows (the examples are from the Liverpool Technology Testbed).

Real	Actual wheels and axle of a car, with steering wheel so that the effect of turning the steering wheel can be seen
↓	Pulley systems with ropes, to lift loads otherwise too heavy to lift
Partly real	Stress pattern in a sheet of translucent material observed through polaroid film
	Model water pump
↓	Loads pulled up inclined planes by strings passed over pulleys and with small weight carriers attached
Wholly representational	Optical illusions

This dimension is an important one, for there is evidence that the more real an exhibit is the more sure its point is grasped and the greater the motivation to engage with it and try to understand what is happening. Having chosen such a dimension, specific information could be gathered about certain exhibits, contributing to the improvement not only of those exhibits but also to others which fall at the same point along the dimension.

For example, there is a group of exhibits which are models on a small scale of a real thing – the popular Arch Bridge is a good example. It may or may not be that these exhibits can be improved if the link between the real thing and the model is more clearly made. In-depth study of this issue in one case could well suggest modifications in other model-type exhibits which would be worth trying.

Information required

This must be determined so that the various questions which are being asked about the exhibits can be answered.

Some of these are embodied in the formal criteria for the evaluation such as 'the extent to which visitors find the experience enjoyable, stimulating and educational'. (Criteria for success offered by the Liverpool centre developers).

There are many other questions of a more detailed kind which have to be answered on the way to meeting these overall criteria. For example:

- what is the nature and extent of the interaction between visitor and exhibit?
- what do visitors understand to be the point of an exhibit or of the exhibition as a whole?
- what exhibits are most and least liked by visitors and for what reasons?
- what positive and negative features of an exhibit do visitors perceive and how do these influence their interaction?
- in what ways do they feel visiting the exhibition is time well spent?
- . . . and so on?

Each of these can be broken down into more specific questions about the value of written information provided to accompany an exhibit, the role of demonstrators/helpers in encouraging interaction, etc.

Methods of data collection

In theory, the step of deciding methods of data collection is simple once the questions to be asked have been identified; it is just a matter of choosing what is most appropriate for the information required. In practice this is not an easy matter, particularly in view of the inevitable constraints of time and resources within which evaluation has to take place. But there are problems which even unlimited resources could not solve. These are problems of finding valid and reliable methods of obtaining data. Any one of the questions above can be seen to pose a considerable problem in these terms. How, for instance, does one *really* find out what visitors *understand* to be the point of an exhibit? How can one capture the *nature* of their interaction with it? The first of these requires one to get inside the head of the visitor and the second to experience the interaction exactly as the visitor does (video-taping would only record the external features of this interaction). Given the impossibility of doing these things, it is inadequate to some extent. A safeguard that can be adopted is to attack the problem of collecting a certain kind of data from several angles. So, to video-tape *and* discuss their interaction with a visitor is better that to do either alone; one can ask what people found to be their favourite exhibit *and* also watch to see how much and for how long they interact with it.

What are the methods likely to be of use in a formative evaluation?

Briefly the possibilities are:

- direct observation recorded in notes
- direct observation recorded using a checklist
- direct observation recorded on video tape.
- direct observation using still photography to aid notes
- interview/discussion during visit
- self-reporting via questionnaire
- self-reporting using a micro computer
- self-reporting using a tape recorder

It is assumed that the nature of these methods is well known and this is not the place to go into the pros and cons of each one. But it is relevant to make a few points, given the almost exclusive dependence on end-of-visit interviews and questionnaires in many evaluations. The limitations of these methods must be evident in the present context. Faced with the kinds of questions listed in the last section, which are bound to be asked about an interactive exhibition, it is arguable that some kind of direct observation of visitors has to take place. This is indeed essential if there is an emphasis on specific concerns rather than general ones. Consideration then has to be given to the possibilities and problems of using video recording, note taking or checklists, or a combination of these. Many factors will come into this consideration, but high on the list should be what is best for the particular questions being asked. Some form of grid of information required against possible means of obtaining it is helpful in making the decision about methods.

Information	Possible methods					
	Observation using			Interview/discussion		
required	notes	ch list	video	during	at end	etc
Biographical						
Understanding of exhibits						
Liking of exhibits						
Duration of interaction						
Type of intactn: reading						
touching						
etc						

Data collection and interpretation

The timing and extent of the data collection have to be chosen to fit the timing of the evaluation and the nature of the information required. So, for instance, for information of the more generalised kind it would be necessary to choose the timing of the data collection and the size of the sample so that it would include all exhibits and all age groups. If a sample is to be broken down into sub-groups, then it would have to be in the region of at least 200 visitors to give about 30 in each sub-group. On the other hand, if such sub-division is not envisaged, a smaller sample would suffice. Thus what is to be done with the data at the interpretation stage influences the scale of interpretations before deciding the required scale of the operation.

A small pilot exercise serves the same purpose and should enable all parts of the evaluation to be given a 'dry run' before full-scale use. This is a counsel of perfection which is rarely possible to follow in a rigorous way, but procedures do have to be tried out and it is well worth extending this through to the data interpretation stage, even if only on a very small scale. There are many ways in which the later parts of the evaluation process can be made easier by forethought at earlier stages and these won't always become apparent unless there is some try out. Where computer processing is envisaged, it is worth examining the design and lay out of checklists, interview schedules and any questionnaires to save time-consuming, hand coding later.

The conduct of the evaluation

The question may arise as to whether, in order to carry out evaluation, one needs 'an evaluator' who is separate from others in the development team. There have been many debates conducted in evaluation literature about the value of having a so-called 'independent' evaluator. For formative evaluation, present thinking seems to have moved away from insistence on the separation of evaluation and development roles. There is no doubt that having separate pairs of hands to deal with the evaluation activities leaves more time for the development activities, but it is less justifiable to suggest that *thinking* should be separated into evaluation and development categories. As long as developers can maintain a critical stance to their products and regard them as problematic, then they are capable of engaging in on-going evaluation.

On the other side, we should acknowledge that having an evaluator does not bring a guarantee of objectivity. All kinds of value judgements are made at all stages in evaluation. For example, in completing the grid above, the relevance that discussion and interviewing are seen to have may well be determined by the extent to which one views such things as learning and interest either as a product of the exhibit or as created by the visitor's interaction with it. This is a reminder that evaluation is not an objective, value-free process, but indeed one that itself involves value-based judgement which will affect the outcome. These value judgements cannot be avoided but, as far as possible, the basis of decisions within the evaluation should be made evident.