

Chapter 2



EVALUATION METHODOLOGIES

2.1 INTRODUCTION

This chapter provides an overview of evaluation theory and methods, and describes and justifies the methods and instruments used to conduct the research in this thesis. In Chapter 1, the term ‘activity’ was used to describe the science communication activities evaluated. Of course, evaluation is not only conducted in the field of science communication. When describing evaluation in the current chapter, the more general terms ‘intervention’ or ‘programme’ are used, with the understanding that a science communication activity is a specific type of intervention. In the remainder of the thesis, the terms ‘intervention’ and ‘activity’ are used interchangeably.

2.2 EVALUATION THEORY

Evaluation has been described as a ‘transdiscipline’ (Scriven, 1996), and as an ‘overarching metadiscipline’ (Picciotto, 1999). Evaluation is important in any area of activity where issues of effectiveness and impact are to be considered (Rossi and Freeman, 1989). Chen (1990) describes six domains, and associated theories, in programme evaluation, derived from normative and causative theory. These domains and associated theories are summarised in Figure 2.1, and described in more detail in Sections 2.1.1 and 2.1.2.

Figure 2.1 Six domains relating to programme evaluation

Domain		Definition
Normative	Treatment	Treatment is the action or element that produces the change within a programme
	Implementation environment	Describes the environment within which a treatment is implemented
	Outcome	The intended and unintended outcomes of a programme
Causative	Impact	Assesses the impact of the treatment on the outcome
	Intervening mechanism	Investigates the mechanisms relating implemented treatment with outcome
	Generalisation	Provides information on how evaluation results can be generalised to apply to future systems

Summarised from Chen (1990)

2.2.1 Normative evaluations

Normative evaluations deal with the ways in which a programme's outcomes and delivery can be optimised by comparing the normative goals, treatments and environments to the actual programme activities. *Normative outcome evaluations* aim to improve the linkage between goals and outcomes, and aim to assist stakeholders in identifying and prioritising programme goals, as well as ensuring the realistic setting of such goals. *Normative treatment evaluations* consider the link between programme implementation and delivery and outcomes. The term treatment is defined as the actions into which the programme goals are translated, that is, the basic element that may or may not produce the desired changes. *Normative implementation environment evaluations* explore the environment in which the programme is delivered, considering, for example, participants, implementers, partner organisations and mode of delivery.

2.2.2 Causative evaluations

Causative evaluations aim to explore causal relationships between an intervention and its outcomes and impact. Types of causative evaluation include *impact evaluation*, *intervening mechanism evaluation* and *generalisation evaluation*.

Impact evaluations are the best-studied type of evaluation. They can be rigorous and evidence-based, and are closest to the image many science communicators would associate with the term, as it is similar to the summative (Dunn, 1981) or outcome (Posavac and Carey, 1989) evaluation that would typically be employed to judge the extent to which a programme had succeeded. An impact evaluation aims to judge the effect of the treatment on the programme outcomes. There are two lines of thinking within the field of impact evaluation, relating to if or how the goals of a programme should be incorporated into its impact evaluation. The goal-oriented approach (Tyler, 1942; Weiss, 1972) places the emphasis on achieving objectives as the key measure of a programme's success. However, this model has limitations – it can ignore potentially important unintentional outcomes, and the goals themselves may be vague (Scriven, 1972; Chen and Rossi, 1980; Weiss, 1972). Scriven (1967) introduced the concept of 'goal-free' evaluation, where the evaluator explores all impacts of a programme with no prior knowledge of the stated goals. It is usual, however, to take the approach advocated by Verschuren and Zsolnai (1998), who concluded that:

'...the value of a program or a decision is determined not only by the achievement of its stated goals but also by its intrinsic ethical value and its performance for the stakeholders'

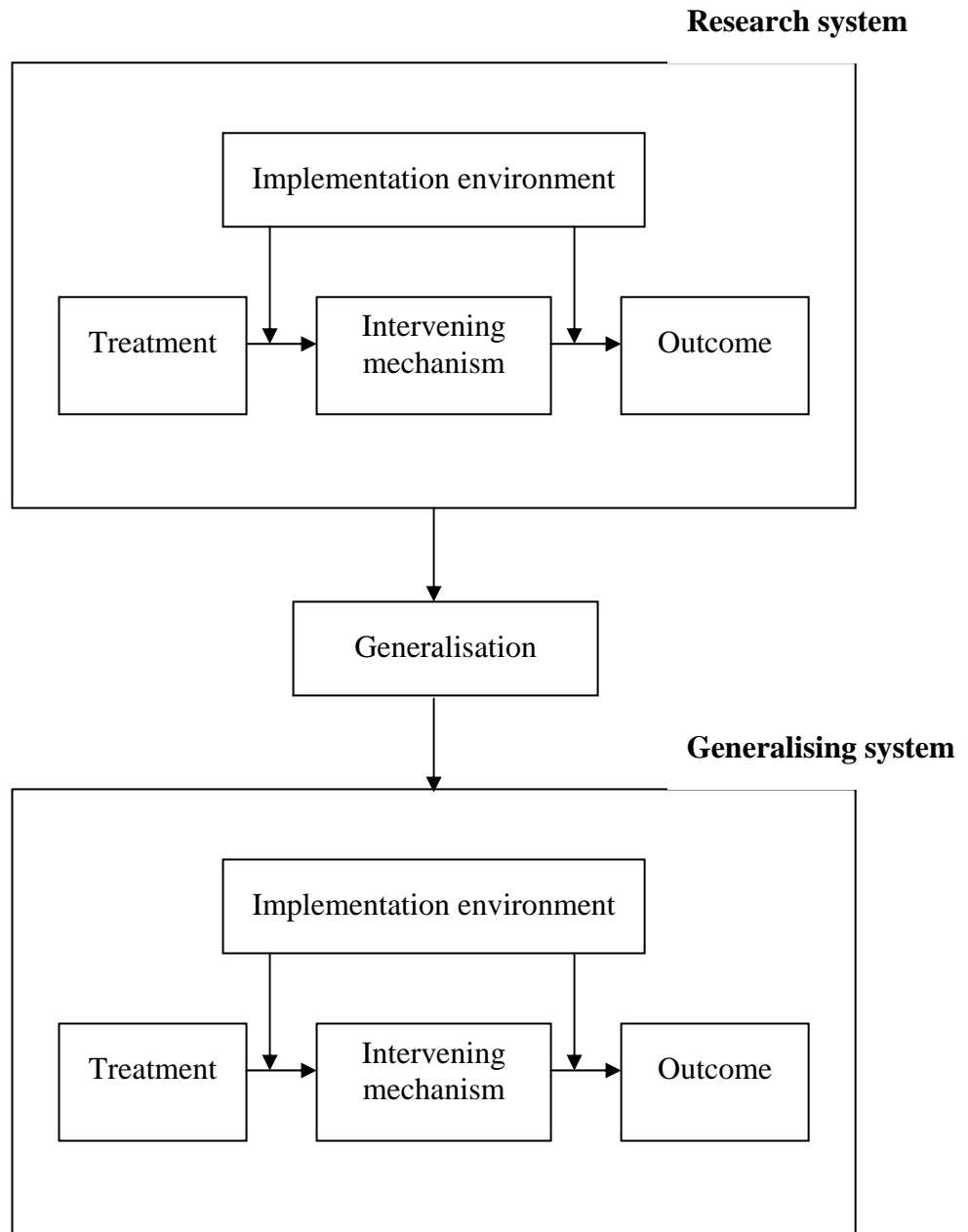
Impact evaluations often consider the stated goals of a programme while remaining open to the exploration of unintended outcomes.

Intervening mechanism evaluations aim to probe the relationship between the treatment and the outcome, with a view to identifying causal factors. In this way, the evaluation can provide information on the reasons for the success or failure of a programme – far more valuable than simply reporting whether or not it was successful. Another type of causative evaluations, *generalisation evaluations*, explore the ways in which the results from an evaluation can be applied to other situations of interest to the stakeholders. If the facility for generalisation is not included in the initial evaluation strategy, problems associated with subsequent under- and over-generalisation can occur (Chen, 1990).

2.2.3 Relationships between domains

The ways in which these domains, theories and types of evaluation interact are summarised in Figure 2.2.

Figure 2.2 Relationships between domains



Adapted from Chen (1990)

2.2.4 Approaches used in the current research

The current research focuses on impact evaluation of several science communication activities. Intervening mechanisms are also considered where appropriate. By employing similar methodologies to a number of programmes, some generalisation is possible. The final chapter of this thesis proposes the basis for a framework which

would allow greater generalisation of science communication activities and their evaluation.

2.3 EVALUATION RESEARCH METHODS

2.3.1 Experimental study design

Before-and-after and after-only designs were used, depending on the nature of the intervention. The before-and-after design was used with the school groups in Chapters 3, 4 and 5, where it was possible to gain responses from the students who would be involved in the intervention at both stages. This was not possible at the science festival (Chapter 6) or generic venue (Chapter 7) as there was no way of identifying audiences prior to the interventions. Control groups were not used, as one of the biggest problems with comparable study designs, especially in social science contexts, is that it is impossible to ensure that the control group and treatment group are comparable in every sense other than the intervention (Kumar, 1996).

2.3.2 Methods of data collection

The main methods available for collecting data for evaluative research, and their key advantages and disadvantages, are summarised in Figure 2.3.

Figure 2.3 Possible methods of data collection

Method	Advantages	Disadvantages
Observation	<ul style="list-style-type: none"> • Suitable for collecting data related to behaviour • Works well when subjects are involved in an interaction and unable to provide objective opinions 	<ul style="list-style-type: none"> • Subjects may change their behaviour if they are aware they are being observed • Potential for observer bias or difference in interpretation between observers • Difficult to simultaneously observe and record
Interview	<ul style="list-style-type: none"> • Appropriate for complex situations • Allows collection of in-depth information • Responses can be probed further • Questions can be explained 	<ul style="list-style-type: none"> • Potential for interviewer bias • Requires skill on the part of the interviewer • Time-consuming and expensive
Focus group	<ul style="list-style-type: none"> • Very 'rich' source of data • Allows group interactions to be observed as well as opinions gathered 	<ul style="list-style-type: none"> • Time-consuming and expensive • Requires skill on the part of the interviewer as group dynamic is crucial to collecting useful data
Questionnaire	<ul style="list-style-type: none"> • Less expensive • Greater anonymity • Can be distributed in a number of ways 	<ul style="list-style-type: none"> • Appropriate questionnaire design is crucial to success • Inappropriate for use with some groups, e.g. young children, illiterate adults • Potentially low response rate • Self-selecting bias • Clarification of questions not possible
Secondary sources	<ul style="list-style-type: none"> • Generally inexpensive • Convenience 	<ul style="list-style-type: none"> • Validity and reliability problems • Data format may not match format required by researcher

2.3.3 Sampling

The manner in which a research sample is selected is important, as steps must be taken to eliminate or understand any bias present in the samples surveyed. A brief overview of the main sampling techniques available is given in Figure 2.4 below.

Figure 2.4 Possible sampling techniques

Type	Method	Description
N/A	Census	All members of the population to be studied are included
Random/ probability	Random	Sample members selected from the population randomly
	Stratified	Homogeneous strata within the population are identified. Random samples are then taken from each stratum
	Cluster	For larger populations, clusters are identified, potentially at a number of levels, until the stratified sampling technique can be used
Non-random/ probability	Quota	Sample members selected by means of a visible characteristic (e.g. gender) until quota is met
	Judgemental	Sample chosen based on researcher's judgement of who can provide the most valuable information
	Snowball	Sample selected using networks where each sample member is asked to recommend future sample members
Mixed	Systematic	Selection of the nth member of a population or stratum

2.3.4 Data collection methods and sampling techniques used in the current research

The current research used open- and closed-form questionnaires and structured interview schedules as the primary data collection instruments. These allowed information to be collected in a consistent manner between studies, enabling systematic analysis and comparison. In addition, data were collected for some

studies from observation, electronic voting and secondary sources. In order to reduce bias, census rather than random sampling was used where possible in the schools research. In some situations, the census would be among a cluster (for example, a group of school students involved in an intervention) rather than a population.

2.4 DATA COLLECTION INSTRUMENTS

This section describes in more detail the data collection instruments used in the current research. A few questionnaire items differed slightly between interventions, although the main sections of the questionnaires were the same. Copies of all data collection instruments are given in appendices to the relevant chapters.

2.4.1 School groups questionnaires


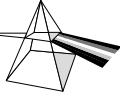

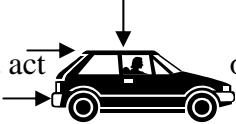

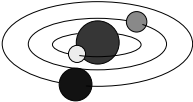
The before-and-after studies conducted in Chapters 3, 4 and 5 were designed to measure students' attitudes towards physics and their physics knowledge before the relevant intervention and then, by comparison with their views after the intervention, explore any changes in attitude and knowledge. Students were presented with an identical set of attitudinal statements and '*knowledge quiz*' questions at two stages, in the week before and the week after the relevant intervention. In addition, the second stage questionnaires included some items typical of '*after-only*' survey designs. These items aimed to collect, in a direct manner, students' opinions of the intervention, and asked students to rate their own self-perceived attitude and knowledge shifts. The structure of the questionnaires is described in more detail below.

Cover sheet

At both stages, the questionnaire cover sheet recorded each student's name, age, year group, gender and school. In addition, students were asked to record their names, so that each individual's responses before and after the intervention could be compared. This allowed a more detailed analysis than a population-based study alone. Students were asked to rate how they felt about physics on a five-point scale from '*really like physics*' through '*neither like nor dislike physics*' to '*really don't like physics*'. Students were also asked to rate their own ability in physics and to record how often they visited museums and/or science centres. The information regarding perceived ability and museum visits was recorded at the first stage only, as these variables were considered independent of the intervention. For Year 8 and Year 10 students, the cover sheet also included a definition of physics. This was recommended following research by Spall (2005), which found that many students were unable to clearly distinguish between chemistry, biology and physics. This is a likely consequence of the combined science course that most students take up to age 16. Often, lessons are simply labelled 'science' with no differentiation between the sciences. The description used bullet points and simple images to describe some of the main topics included in physics. A copy of the description is given in Figure 2.5.

Figure 2.5 Description of physics included on questionnaire cover sheet

Physics is about all of the below:

- Electricity  and circuits
- The way light  and sound work and travel 
- Gravity and other forces that act  on things and how things move
- Magnets  and electric motors
- Space, stars  and planets

Affective impact

The attitudinal tracking statements used before and after the interventions were those developed by Spall (2005) for use with students from Key Stage 3 to undergraduate level. The questions are shown in Figure 2.6.

Figure 2.6 Attitudinal tracking questions

Physics is an **interesting** subject
You need to be **good at maths** to do physics
Physics is more of a **boys'** subject
Physics is a **boring** subject
Physics is more to do with **remembering facts** than understanding ideas
Things I learn in physics **relate to my everyday life**
People who really like physics **don't mix very well** with other people
Physics is more of a **girls'** subject
Physics uses **difficult, complicated** words
Physics is an **easy** subject
Physics uses **easy, everyday words** but with a **different meaning**

Adapted from Spall (2005)

Students were asked to rate whether they agreed or disagreed with the same set of statements on a 5-point Likert scale before and after the intervention.

Cognitive impact

For each of the school-based interventions, ten multiple-choice knowledge quiz questions were developed to test students' knowledge of facts relating to the activity content before and after the intervention. These were presented in the form of a 'Quick Quiz' in order that, as far as possible, this part of the questionnaire was not seen as a formal test. This method is somewhat crude, and measures only the improvement of factual knowledge; improvements in procedural or conceptual knowledge were not considered in the present research because students had limited time available to complete the questionnaires. The difficulty level of the questions varied, and four control questions were included, which tested knowledge that was not included in the activities. The remaining six questions related to information covered during the intervention. An example question from the evaluation conducted in Chapter 3 is shown in Figure 2.7 below:

Figure 2.7 Example question

Which planet is closest to the Sun?

Jupiter	<input type="checkbox"/>	Earth	<input type="checkbox"/>
Mercury	<input type="checkbox"/>	Venus	<input type="checkbox"/>

Evaluation questions

The evaluation questions were designed to assess students' opinions on the intervention, as well as asking them to rate their own learning and attitudinal shifts. Students were also given the opportunity to comment in open questions in this section of the questionnaire, in order to explore some of the intervening mechanisms contributing to the impact of each intervention.

Survey of teachers

Teaching staff were asked to complete a short open questionnaire after the schools activities. The questionnaire had two aims: firstly to collect teachers' opinions of the intervention, and secondly to use teachers as a means of gauging the impact of the lecture on their students and exploring the intervening mechanisms between the intervention and its impact on students. Staff were also asked to record which subject and year group they taught, and their gender.

2.4.2 Questionnaire administration

Schools were recruited into each study by the researcher before the intervention, with the assistance of the activity organisers if appropriate. The first stage questionnaires were distributed by post, with a covering letter explaining the purpose of the research and the conditions under which the questionnaires should be completed.

For the benefit of future researchers, it is worth describing certain particulars of the questionnaire administration that allowed good response rates. For two of the interventions, (the '*Science is Cool*' lecture evaluated in Chapter 4 and the '*Great Balls of Fire*' lecture evaluated in Chapter 5) the researcher was able to offer the intervention to a school for no charge, in exchange for participation in the study. This was a favourable situation because administering the questionnaires before and after the activity involved considerable effort on the part of teachers, and providing an incentive such as the free lectures allowed a good response rate to be achieved for the studies. In addition, contact with teachers was prolonged while the date, time and arrangements for the lecture were confirmed (this was co-ordinated by the researcher). This allowed teachers to be reminded about the study, and gave them the opportunity to ask any further questions of the researcher. Where possible, the researcher visited the school on the day of the activity, to meet the teacher co-ordinating the questionnaire distribution, collect the first stage questionnaires and deliver the second stage forms. Again this worked well because it was possible to thank the staff involved in the data collection in person, and explain the purpose of the research. These factors contributed to the good response rates achieved for the studies involving the lectures that visited schools.

When the incentive of a free lecture and the close contact between researcher and schools was not possible, response rates suffered. This was particularly acute for the study in Chapter 3, where few schools included in the sample successfully completed questionnaires both before and after the visit to the National Space Centre, meaning that much of the data collected had to be discarded. Due to the location of the schools involved in the intervention, it was impossible for the researcher to visit the

schools, and the study was conducted by post. Similar issues arose for the visits to Culham Science Centre, evaluated in Chapter 5, although the postal questionnaire issue was compounded by the fact that few visits take place, and often fewer than 10 students take part in each visit.

2.4.3 Structured interviews used for public audiences

Where audiences for an activity could not be identified prior to the intervention, an ‘after-only’ design was used. The primary data collection instruments used in both Chapters 6 and 7 (evaluations of Cheltenham Festival of Science and Science in the Fast Lane) are structured interviews, although data from other instruments such as questionnaires and electronic voting are included. The structured interview methodology was chosen because it can partly eliminate the self-selecting bias encountered with questionnaires because respondents are actively approached. In addition, structured interviews can allow inclusion of items that were consistent with items included in the questionnaires used for school groups. Also, in Chapter 6, data were collected by several interviewers, so it was important to maintain consistency among the responses. The structured interviews included the item ‘*before you came to [intervention], how did you feel about science?*’ as an attempt to probe respondents’ pre-existing attitudes. The interviews also included the items from the questionnaires asking respondents to rate their self-perceived attitude and knowledge shifts, as well as exploring opinions of the intervention.

2.5 DATA ANALYSIS

Data are analysed in two ways in this thesis. Firstly, each chapter evaluating a particular activity (Chapters 3, 4, 5, 6 and 7) includes an analysis of the data collected for the relevant intervention. This includes a descriptive analysis and an exploration of associations and differences within the data sets. Secondly, Chapter 8 presents a meta-analysis of the data collected in each of the individual studies, allowing some comparisons to be made.

2.5.1 Descriptive analyses

The types of analysis that can be performed on a data set depend on the way in which the variables are measured. The main types of measurement scale are summarised in Figure 2.8.

Figure 2.8 Types of measurement scale

Type	Description	Example
Nominal	Classifies data points into groups that have no inherent order	Gender, favourite newspaper
Ordinal	Classifies responses into subgroups that have an inherent order or ranking	Attitudes measured on Likert scale
Interval	Classifies responses on a scale that has its own units	Height, age

The data collected using the attitudinal tracking statements (described in Section 2.4.1) is ordinal in nature. The data collected for the knowledge quiz questions are measured on a nominal scale – responses were classified as ‘correct’ or ‘incorrect’. The evaluation questions asked after the activities incorporate a mixture of ordinal and nominal data, and some of the demographic information collected was interval in nature. The statistical tests described below were used to analyse the data collected from the before-and-after studies in Chapters 3, 4 and 5. Statistical analysis was not

performed on the data in Chapters 6 and 7, these data were analysed using descriptive methods.

Tests for association

Associations between pre-existing attitudes towards physics, self-perceived ability and frequency of museum visits were explored using a nonparametric ranking test. Ranking tests are typically used with ordinal data (Kinnear & Gray, 2004). They compare pairs of responses, and look at the polarity of the difference between them. There are two types of ranking test that can be applied to measure associations between variables measured on ordinal scales. Spearman's rho is equivalent to the Pearson correlation used to measure linear relationships between interval data. Although the formula is different, the coefficient obtained is equivalent. Kendall's tau statistics consider pairs of ranks, and the number of reversals of pairs required to transform one set into another. Because the variables all used 5-point ranking scales, Kendall's tau-b test was deemed the most appropriate to use.

Tests for differences

Statistically significant differences between responses to the attitudinal tracking statements before and after the interventions were explored. Nonparametric tests were used, because unlike their parametric counterparts these tests do not require that the data are normally distributed. The Wilcoxon test, a nonparametric equivalent to the related-samples t-test, was used to explore differences between attitudes measured on the ordinal Likert scales before and after the interventions. Whether ordinal data can be meaningfully analysed using parametric tests is a grey area (Kinnear and Gray, 2004). However where the data are in the form of a set of ranks

such as the points on a Likert scale a nonparametric test is most appropriate. The Wilcoxon test pairs the two scores for each variable, and ranks the differences between scores in order to compare the medians of the two samples. McNemar's test is the nonparametric equivalent for nominal data. It was applied to measure differences in the responses to the knowledge quiz questions.

2.5.2 Meta-analysis

The meta-analysis, reported in Chapter 8, presents the data collected in each chapter in a collated manner. Firstly, the students' pre-existing attitudes towards physics are presented and described. These were explored with the attitudinal tracking statements in the first stage questionnaires for the three school studies. Secondly, the attitudinal and cognitive impacts of each of the schools interventions are compared using the before-and-after data, and differences in impacts on males and females is discussed. Thirdly, the impacts of all interventions are compared using the after-only data which asked audiences to rate their own self-perceived learning and attitude shifts.

The meta-analysis was designed to compare the impacts of different interventions, using a consistent set of indicators. Through this process, the robustness of the indicators themselves has also been tested. The meta-analysis compares the impacts measured directly (by comparing responses before and after interventions) with those measured indirectly (responses to questions regarding perceived impacts), and allows a statistical comparison of responses. The statistics used are described in more detail in Chapter 8. In this way, the trustworthiness of the after-only data can be explored. This is important because in practical evaluation of science communication activities

it is rarely possible to conduct a detailed before-and-after research study into the impact of an activity.