Review

Assessment of biology investigations

Ros Roberts and Richard Gott
School of Education, University of Durham, Durham, UK

The restricted range of biology investigations submitted for assessment in England and Wales, almost exclusively laboratory-based and with very little fieldwork, can be seen as a consequence of Sc1 being perceived as a skills-based performance model. An alternative approach to procedural understanding conceptualises investigations as the process of utilising a knowledge base, the Concepts of Evidence. Four biology investigations set in different contexts are analysed in this paper and the Concepts of Evidence used are shown to be very similar for each. However, the sequence in which ideas are used and actions occur differs between lab-based investigations and fieldwork. A case is made for the assessment of investigations against the Concepts of Evidence in written tests as being potentially a more reliable and valid way of assessing the ideas used in all types of biology investigations, thus reducing the distorting effect of assessment on the curriculum.
Key words: Assessment, Biology investigations, Concepts of evidence, Fieldwork, Procedural understanding.

Introduction

Since its inception in 1989, the National Curriculum (NC) for Science in England and Wales has included a strand, now known as Scientific Enquiry (Sc1), which specified what should be taught about scientific investigation. Early versions, while not specifically excluding 'field' investigations, were interpreted as being restricted to a laboratory context where control can be maintained by manipulating variables (Roberts and Gott, 1999). A range of investigations are used in biology: lab-based physiology, investigations into behaviour, ecological surveys, to name but a few. Later revisions of the NC (the latest being in 1999), explicitly broadened the context for Sc1 investigations, and specified that pupils aged 11-16 'should be taught…how evidence can be collected in contexts [for example fieldwork, surveys] in which the variables cannot readily be controlled.' (Science National Curriculum, 1999: KS4 Sc1 2d). Some of the problems perceived by biologists would seem to have been addressed by this change.

However, in our experience, the vast majority of teaching and assessment of Sc1 still takes place in the context of lab-based investigations, which emphasise manipulation of variables. Endless variations on the theme of enzyme activity and osmosis seem to predominate. Does this restricted view of biology investigations matter? We believe it does, for reasons that have been developed elsewhere and which we will not rehearse here other than to point to the obvious backwash effect on the curriculum (Donnelly, 1995; Bencze, 1996; Roberts and Gott, 1999; Watson *et al.*, 1999a; 1999b; Roberts, 2001).

It is easy to lay the blame for the current problems at the doors of the Examination Boards or a poorly specified National Curriculum. However, we will argue that it is the fundamental approach taken to Sc1 that lies behind the exclusion of many valid investigations from the curriculum and its assessment.

In this article we explore why Sc1 assessment has become dominated by the lab-based manipulation model of investigations and what might be done to move away from the current situation.

The underlying problem

Why is it that many biology investigations, while not being excluded explicitly from the latest version of the National Curriculum, have failed to feature in assessment? In what follows we will be referring to the de facto position as perceived by teachers during moderation procedures rather than the theoretical position adopted (or not) by examination boards. We would like to suggest that the narrowness of the tasks used for assessment results from two main, not unrelated but insufficiently considered, aspects of investigative work in schools:

- The predominant view of Sc1 is based around a skills and performance based approach to science rather than one based on an understanding of evidence.
- As a consequence, the criteria used for assessing Sc1 do not give credit for fieldwork where pupils have to handle large amounts of data and frequently have to dichotomise continuous variables (treating them as categoric in effect).

A typography based on variable structure

Before we can consider how this position might have arisen, however, we need to have some sort of structure, or typography, of investigative tasks against which we can make judgements. A number of such typographies have been developed before (e.g. Watson *et al.*,1999a) which encompass ideas such as measurement tasks, 'pattern seeking', 'design and build' activities, 'forensic' identifications in chemistry, and such like. However, we shall restrict ourselves in what follows to variable-based tasks where the explicit focus of the activity is the search for a link, causal or otherwise, between two or more variables. If we further restrict ourselves to just two variables it is possible to

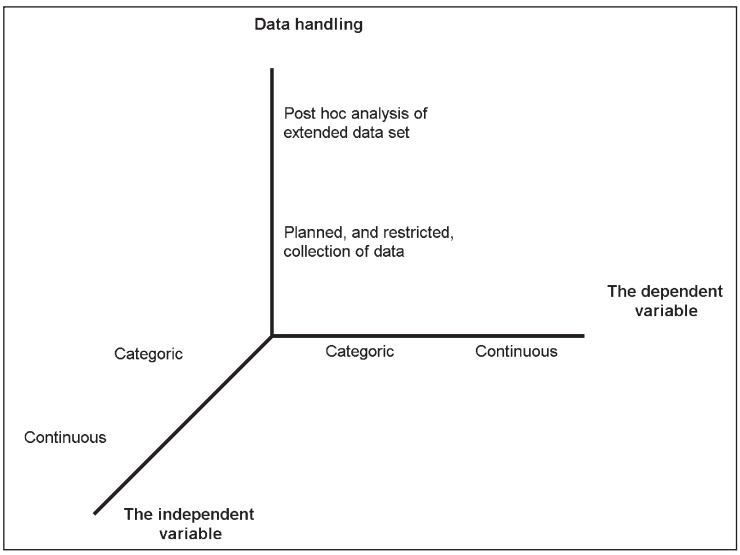

Figure 1 *Variable-based typography of investigations.*

make a broad definition of task types, based on the variable structures and how the data is handled. This is shown in Figure 1. (Please note that, for simplicity, this typography does not include the possible effects of interacting variables.)

The values on the data handling axis in Figure 1 can be distinguished by whether the investigator is able to change the values of the variables by manipulating them or not. Thus the 'planned and restricted collection of data' results from an investigation where the investigator is able to change the value of the independent variable (IV) (e.g. the type of fertiliser or the distance from a light source). Validity is ensured by manipulating the values of control variables (CVs) so that their possible effects do not affect the dependent variable (DV). Such data collection is referred to in this article as 'lab-based' contexts but can occur outside in some field experiments as well. 'Post hoc analysis of an extended data set' occurs when variables cannot be manipulated. In such circumstances, typified perhaps by the ecological investigations, the aim is to collect as much data as possible whilst in the field. Particular questions are then investigated using the data set by selection of suitable sets of measurements which make up a post hoc equivalent of a controlled lab-type experiment. Such a post hoc analysis may be to identify possible independent variables from a survey or may have a specific independent variable identified from the outset. We shall refer to such investigations as 'fieldwork'. Of course there are intermediate positions.

Investigative tasks can be seen as being located in one of the eight cells in this diagram. The simplest task, and the least powerful in terms of the explanatory power of the data, would be a planned one in which both variables were categoric; whether bean seeds can germinate in dry or wet conditions, for instance. The most complex would be one involving two continuous variables and set in the 'field' such that analysis is carried out post hoc on an extended data set that includes different values of variables whose affects have to be accounted for in the analysis. Such a task might be a stream survey, looking at how the abundance of mayfly larvae is affected by aquatic pollution. (There are, of course, any number of more complex tasks involving two or more independent variables and/ or dependent variables, which may or may not interact.)

These various possibilities in a biological context are detailed in Table 1 where we number them for convenience, but without implying any necessary hierarchy of difficulty.

Current assessment practice favours types 2 and 4 in this typography. This typography enables us to consider what are often thought to be very different approaches to investigations in a common framework. These approaches are seen to have a common variable structure but can be distinguished by whether the data collection is planned and the variables manipulated or whether larger amounts of data are collected for post hoc analysis. This typography structures the argument that follows.

How have the problems with assessment arisen?

To understand how the restrictive assessment position might have arisen and to develop an alternative perspective that enables fieldwork (requiring post hoc analysis) to be validly and

Table 1 Variable-based typography with examples.

Type	*Independent variable*	*Dependent variable*	*Data handling is:*	*Example of investigative task*	*Currently used for assessment purposes?*
1	Categoric	Categoric	Planned	How does the presence or absence of water affect the germination of a seed?	X
2	Categoric	Continuous	Planned	How does the presence or absence of 'pollution' affect the number of duckweed 'leaves'?	Lower levels of Sc1
3	Continuous	Categoric	Planned	How does the light intensity determine the presence or absence of starch?	X
4	Continuous	Continuous	Planned	How does the light intensity affect the number of bubbles (from pondweed)?	Higher levels of Sc1
5	Categoric	Categoric	Post hoc	How does mowing affect the growth form of species present?	X
6	Categoric	Continuous	Post hoc	Do north or south facing slopes give a better crop of apples ?	X
7	Continuous	Categoric	Post hoc	How does the pH of the soil affect the species present?	X
8	Continuous	Continuous	Post hoc	How does the total area of a dandelion's leaves affect its root length?	X

Note: Current assessment practice favours types 2 and 4 in this typography.

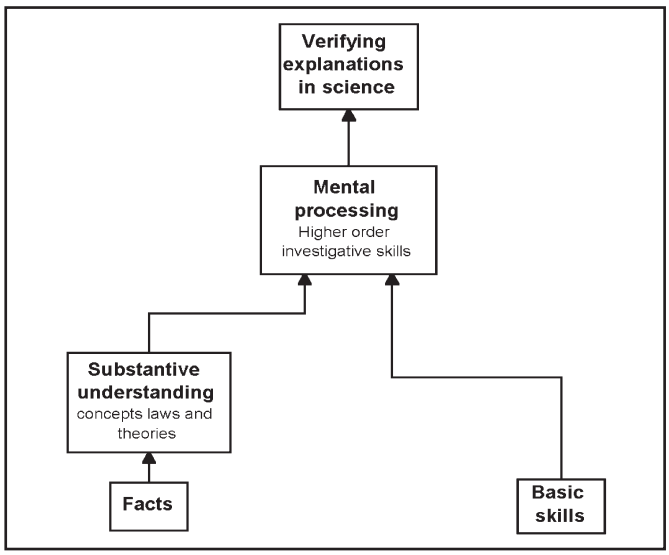

Figure 2 *A skills approach within an explanatory focus (from Gott and Duggan, 2002).*

reliably assessed, we need to apply ideas developed by Gott and Duggan (2002). They argue that the approach to both teaching and assessment of investigative work can be considered from two very different positions, which we will rehearse here for the purpose of analysis. Each position depends on how investigative work is defined in relation to 'skills' and the rest of science, i.e. the explanatory frameworks of biology, chemistry and physics.

A skills approach to Sc1

The position taken by this approach is typified by a *performance* model, which defines Sc1 as a practical activity. Pupils practice and perform 'skills', which would include not only handling apparatus and organisation skills but also 'higher order investigative skills' such as planning, measurement, observation, data presentation and so on, in the context of 'action' in a practical context. At the extreme, the assumption is that practice increases familiarity with these skills and that no specific understanding is required. It is the *doing* of science that matters and, what is more, the doing is there primarily to help in the understanding of substantive ideas and explanations stemming from them. This view is summarised in Figure 2, which we will contrast with an alternative model in a later section.

What would Sc1 assessment look like from the perspective of a skills approach?

Investigative work, seen from this perspective, would have an emphasis, in both the curriculum and its assessment procedures, on *the way* skills are put together to work like a scientist – the planning and routines of data collection, the associated safety assessments and the ways of presenting and analysing data. Curricula would therefore be couched in terms of *the way* pupils should do investigations, i.e. 'carry out preliminary work' or ' make observations and measurements to an appropriate degree of precision'. And assessment would also emphasise *performance* with behavioural criteria such as 'decide a suitable extent and range of evidence to be collected' or 'collect sufficient systematic and accurate evidence and repeat or check when appropriate'. Such is the case with the National Curriculum and Exam Board criteria respectively.

Let us consider one such example to illustrate the point: 'Use a procedure with precision and skill' (AQA mark descriptor O.8a) is expressed in terms of *the way of doing* an investigation. In some investigations it might be entirely appropriate to simply measure the length of a leaf with a 10 cm ruler or count the number of beetles in a pitfall trap. Another investigation into, say, the identification of photosynthetic pigments in different leaves would require very complex procedures, which would also meet the same criterion. The context of the investigation controls the level of difficulty. So essentially, assessment based on criteria that are specified as actions, or as procedures made up from a series of actions, runs into problems when the actions are disparately difficult in different contexts. The only way to make such an assessment system work is to exemplify the criterion in a particular context, so that everyone knows what it means. This we would contend has happened with assessment of Sc1. Over the years a 'case law' of standard items has, necessarily, built up, which suggests acceptable procedures and contexts in which the criteria can be demonstrated and that are aligned with moderation examples. The 'routinisation' of Sc1 assessment could be argued to follow from this. Custom and practice as well as the demands for reliable assessment (Gott and Duggan, 2002) have limited the number of cases with a consequent limiting effect on the investigative elements of the curriculum. To avoid this we would need to increase the number of investigations so that more 'cases' could be accommodated. This could be done by reverting to Teacher Assessment, but even then there will be a considerable encroachment into the time allowed for Key Stage (KS) 4 work, a time quite nonsensically short for such an overcrowded curriculum, to say nothing of the bureaucracy involved in documenting and moderating all the possible cases.

To consider how the assessment of investigations might differ from this we need to consider how else Sc1 could be viewed.

A knowledge base approach to Sc1

If, by contrast, we consider investigations dependent on a body of knowledge that has a status in its own right, then we can free ourselves from the straightjacket of a set of 'accepted cases' although we do find ourselves in a more complex situation. To be able to investigate not only requires basic skills, such as being able to read instruments and handle apparatus, but also requires a procedural understanding of *the ideas or concepts* that underpin evidence, a position summarised in Figure 3.

The fundamental difference between this approach and the skills approach is that problem solving in science is seen to require an understanding of *two* sets of specific ideas or concepts: a substantive understanding *and* a procedural understanding. Thus, the 'mental processing' that is required when solving problems in a biology context involves thinking about *both* the substantive ideas of biology and specific ideas required to collect valid and reliable evidence, such as knowing how to judge when sufficient repeated readings have been made or whether a sample is large enough for the investigation to be valid. This approach recognises that, in addition to the substantive ideas, there is something, per se, to think about. The 'mental processing' is akin to 'higher order thinking skills'. This approach acknowledges that thinking has to be about something and recognises a knowledge base to procedural understanding.

The ideas that have to be known and understood have been specified as the Concepts of Evidence and include such concepts as how the reliability of measurements might be affected

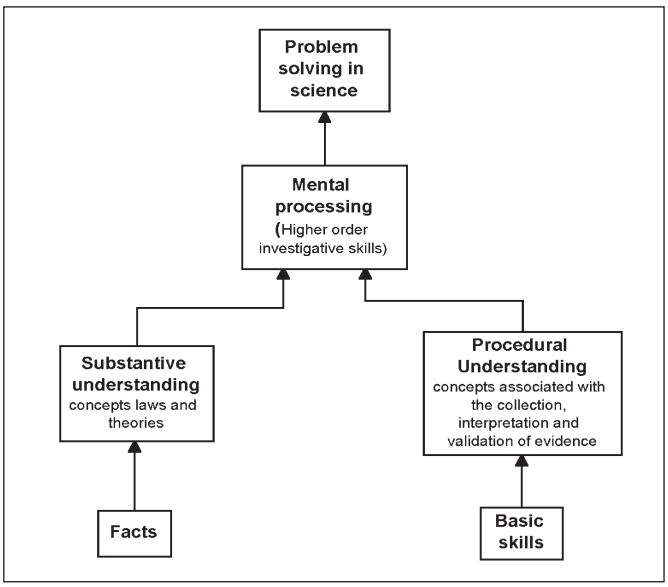

Figure 3 *A knowledge-based model of Sc1 (from Gott and Mashiter, 1991).*

by different measuring procedures and the concepts involved in understanding how variation can be reduced by different control strategies to increase validity. These ideas form a specific knowledge base which, set alongside basic skills, constitute procedural understanding. A complete list can be found at **www.dur.ac.uk/~ded0rg/Evidence/cofev.htm.** A summary of the sections into which the Concepts of Evidence fall is summarised in Table 2.

Each section contains closely related ideas. For instance, and to give some examples of the sort of ideas we are talking about, Section 9 *Design: Validity, 'fair tests' and controls* contains related Concepts of Evidence (see Table 3).

What would the assessment of biology investigations look like from the perspective of a knowledge-based approach?

Any investigation, biology or otherwise, can now be considered as a strategy for putting together substantive ideas (about photosynthesis or whatever), procedural ideas (concepts of evidence) about sampling or measurement, and the skills needed to make those measurements without breaking everything or reading from the wrong end of a scale. So, taking this argument a step further, we can see that the different sorts of biology investigation, such as lab-based physiology experiments and post hoc field surveys, might now be represented by the different strategies for putting together this complex of ideas and skills rather than as a series of routine actions. Investigations can now be seen as actions *based on understandings*.

Assessment from this perspective has a different emphasis. Instead of assessing actions, the focus of assessment can be on assessment of *understanding* — we have abstracted the 'understandings' from the particular set of actions, which characterise different contexts of investigations. We will consider *how* this might be done later, but firstly we need to consider whether the ideas required for understanding are different in different investigation tasks. It is this issue that we shall consider in the next section.

Table 2 Sections of the Concepts of Evidence. (see www.dur.ac.uk/~ded0rg/Evidence/cofev.htm)

Section	*Concept*
1	Instruments: underlying relationships
2	Instruments: calibration and error
3	Reliability and validity of a single measurement
4	The choice of an instrument for measuring a datum
5	Sampling a datum
6	Statistical treatment of measurements of a datum
7	Reliability and validity of a datum
8	Design: Variable structure
9	Design: Validity, 'fair tests' and controls
10	Design: Choosing values
11	Design: Accuracy and precision
12	Design: tables
13	Reliability and validity of the design
14	Data presentation
15	Statistical treatment of measurements of data
16	Patterns and relationships in data
17	Reliability and validity of the data in the whole investigation

Table 3 Concepts of Evidence associated with Design validity of investigations. (see www.dur.ac.uk/~ded0rg/Evidence/cofev.htm)

Topic	*Understanding that:*	*Example*
Fair test	...a fair test is one in which only the independent variable has been allowed to affect the dependent variable.	A laboratory experiment about the effect of temperature on dissolving time, where only the temperature is changed. Everything else is kept exactly the same.
Control variables in the laboratory	...other variables can affect the results of an investigation unless their effects are controlled by keeping them constant.	In the above experiment, the mass of the chemical, the volume of liquid, the stirring technique and the room temperature are some of the variables that should be controlled.
Control variables field studies	...some variables cannot be kept constant and all that can be done is to make sure that they change in the same way.	In a field study on the effect of different fertilisers on germination, the weather conditions are not held constant but each experimental plot is subjected to the same weather conditions.
Control variables surveys	…the potential affect on validity of uncontrolled variables can be reduced by selecting data from conditions that are similar with respect to other variables.	In a field study to determine whether light intensity affects the colour of dog's mercury leaves, other variables are recorded, such as soil nutrients, pH and water content. Correlations are then sought by selecting plants growing where the value of these variables is similar.
Control group experiments	...control groups are used to ensure that any effects observed are due to the independent variable(s) and not some other unidentified variable.	In a drug trial, patients with the same illness are divided into an experimental group who are given the drug and a control group who are given a placebo or no drug.

Sharing the same knowledge base

The knowledge base could provide the basis for assessment for all the types in the typography provided they all share the same knowledge base.

The same ideas? What distinguishes different investigations?

The question we must now turn to is this: do different investigative tasks require a different selection of 'understandings' (concepts of evidence) and skills? If they do, then all we have managed to do is to add detail to the types. If they do not, then our argument holds — types are no more than different strategies for applying the *same* understandings and skills, in different contexts. In Table 5 we analyse four investigations, all with continuous independent and dependent variables, against the concepts of evidence list referred to above.

The tasks we have selected are:

- **Investigation A:** The effect of temperature on changes during osmosis in potatoes; a lab-based investigation involving finding the percentage change in mass of potato 'chips' in water at different temperatures and typical of investigations used for assessment purposes. (Type 4 in our terminology)
- **Investigation B:** The effect of pollution on the growth of duckweed; a long investigation growing duckweed on the windowsill in pots containing different concentrations of detergent. (Type 4 also but taking longer than the standard 'lesson')
- **Investigation C:** The effect of trampling on species diversity; counting the numbers of plant species found at selected sites across a trampled path. (Type 8 in our terminology)
- **Investigation D:** What affects the distribution of freshwater shrimp in a stream?; a survey of a stream recording data about several variables and the number of freshwater shrimp at many sites. (Type 8 in our terminology)

We have restricted ourselves, therefore, to looking at the same types in terms of variable structure, but taking the 'planned' vs. 'post hoc' dimension as our issue of concern.

The procedure we adopted was to create an 'ideal' solution – what could be done if the investigation was being conducted 'ideally' at school level. So, for instance, it would be necessary under such conditions to sample potatoes of one variety, or even a stratified sample of different varieties at a particular time of year. While recognising that not all of these actions are necessarily overt or explicit at school level, we would argue that they constitute an idea that could be considered. The most probable sequence of likely actions for each investigation was recorded. For each action specified, the Concepts of Evidence that were required to make the decision were listed. The first part of the sequence for Investigation C is shown in Table 4 as an example.

Whilst acknowledging the likely variation in the sequence and number of actions according to exactly how the investigation was conducted, a comparison of the number of actions for each investigation that fell within each section of the Concepts of Evidence was made to give an indication of the range and number of concepts which could be used in each investigation (see Table 5).

We see from this analysis that, at this idealised level of school-based task, not all sections of the Concepts of Evidence are relevant. School level investigations use a sub-set of the Concepts of Evidence. However, there is no startling difference between the ideas used in the investigations (A – D) for the idealised solution: all use ideas related to variable structure and choosing values for instance. When we look at the detailed analysis (considering exactly which Concepts of Evidence were used, rather than just which section was referred to) however, a few differences do emerge. The main differences are:

- **Manipulation and selection of the Independent Variable:** In the lab, the values of the independent variable are changed by the investigator (Investigations A and B), whereas in contexts where the values change naturally, the investigator deliberately selects where values differ (Investigation C) or samples to include a range of values (Investigation D).
- **Control variables:** In the lab it is often possible to isolate all potential control variables and keep their value constant during the investigation (Investigation A). Some control variables' values cannot be kept constant but are allowed to change naturally and control is maintained by ensuring that this affects everything in the same way (Investigation B). Sites can be selected where the effect of significant control variables' values are similar (Investigation C). In surveys some selection like this may take place, but validity is also ensured by 'filtering' the data to some extent after it has been collected, so correlations are sought between variables when the value of significant control variables is controlled after data collection (Investigation D). Large data sets are also required to ensure patterns can be seen in the data.
- **Variation in the sample:** In many lab-based tasks, variation in the sample is assumed to be so insignificant that it is often ignored (Investigation A). As variation in the sample increases, ideas of sample size and representativeness become more important (Investigations B – D).

In summary, the differences identified above illustrate that there are a few ideas that are specific to certain contexts but that the understanding required in different contexts is very similar. So is this the only difference?

The sequence of Concepts of Evidence in the four tasks

Analysis of the four investigations also highlights slight, but important, differences in the sequence in which the concepts of evidence are used, and the consequent difference in the volume of data being handled at any one time in the investigation. These differ particularly in Investigation D.

The main differences are the consequence of exactly *when* in the sequence the validity of the investigation is considered by application of control variables: before data collection in Investigation A and afterwards in Investigation D. The amount of data that has to be recorded consequently increases in Investigation D. The data from a survey can be presented on a scatter graph plot — the scatter of points is indicative of uncertainty in the relationship. Some of the variation is unavoidable and is one reason for the large sample. However, some uncertainty can be reduced by controlling key uncontrolled variables, i.e. by only comparing the independent variable and dependent variable where the values of key variables are similar, such as controlling for flow rate, substrate or depth. It is this application of control by filtering *after* the data has been collected that affects the sequence of the Concepts of Evidence used. In Investigation A, the amount of data is reduced by only collecting the data required.

We might note here, in passing, that the differing sequence of the actions taken has implications for assessment that is based

Table 4 Part of the sequence for Investigation C.

Action	*Procedure*	*Example*	*Concept of Evidence required*	*Concept of Evidence section*
1	Identify research question	How does the amount of trampling affect the number of species growing there?	Identify IV (independent variable)	8
2	Define IV	Transect across trampled path across field, since this has continuum of very trampled to less	Validity of design	13
3	Define DV	Number of species present as measure of species diversity	Identify DV (dependent variable)	8
4	Identify other variables that could affect DV	Light, substrate, pH, moisture content, mowing regime, temperature, aspect, slope, disease and pests, grazing	Identify CVs (control variables)	9
5	Determine how to measure DV	Count number of different species touching a metre rule at 10cm intervals across transect (11 readings at each point on transect)	Validity of DV	13
6	Justify measurement method	Counting species will be very accurate. Identification is not necessary, just distinguishing leaves of different types and recording count	Human error	2
7	Determine range of IV	Across path on field. Max = from centre of path (most trampled) to 1 metre beyond visibly trampled area where bare patches (approx 3m in total)	Range	10
8	Determine interval of IV	Intervals = 25cm intervals initially, perhaps smaller at points of interest along the transect	Intervals	10
9	Selection of appropriate instruments	Soil pH kit, measuring tape, metre rule, compass: are their 'scales' going to allow measurement of significant differences?	Resolution	2
10	Selection of appropriate instruments	Soil pH kit, measuring tape, metre rule, compass: how easy to use well?	Human error	2
11	Selection of appropriate instruments	Soil pH kit, measuring tape, metre rule, compass: has their reliability been checked?	Reliability of measuring instrument	3
12	Sort out method	Determined by available apparatus		
13	Set values for CVs:	All class taking readings from sites with similar characteristics. Select section of path not part shaded by trees, fences etc. Soil moisture and pH = assume the same but take samples and record values at each point on transect. Mowing = select site where all transect mown. Slope = select flat site Aspect of transect = same if flat site Substrate = select site with all transect on soil	Keeping values of CVs constant	9
14	Set values for CVs:	Light = will vary daily but will affect all plants equally. Temperature = as above Pests and disease and grazing = unable to avoid this variation. See sample. Compaction might vary but this will be consequence of trampling	Keeping effects of CVs the same	9
15	Sample size	Decide to pool class data to increase representativeness	Sample size to get representative sample	10
16	Sample selection	Strategy must be unbiased: from points of contact on ruler	Representative sample	5
17	Sample size	Decide to pool class data to increase sample size	Sample size	5

on the performance model. The National Curriculum and the Exam Boards allocate marks for planning, obtaining evidence, analysis and evaluation (POAE) but this does not match the sequence of actions for the post hoc analysis with the result that some investigations (Investigation D, again, in particular) cannot be accommodated simply within the POAE framework.

Our analysis was based on two types of investigations (Types 4 and 8) between continuous independent and dependent variables. A similar analysis of investigations using categoric variables shows that they require an understanding of a subset of

Table 5 The number of concepts of evidence 'hit' – by section. (The number and range of Concepts of Evidence in the four 'idealised' investigations)

Section in Concept Evidence	*Description*	*Inv. A: Osmosis*	*Inv. B: Pollution*	*Inv. C: Trampling*	*Inv. D: Shrimp*	*Number of actions*
1	Relationships in instruments	0	0	0	0	0
2	Calibration and error of instruments	5	3	3	4	15
3	Reliability and validity of measurement	2	2	1	2	7
4	Choice of instrument	0	0	0	0	0
5	Sampling	2	3	3	1	9
6	Statistical treatment of datum	4	4	4	0	12
7	Reliability and validity of datum	0	0	0	0	0
8	Variable structure	5	5	5	7	22
9	Validity and controls	2	4	3	2	11
10	Choosing values	3	3	3	2	11
11	Accuracy and precision to determine patterns	1	2	1	0	4
12	Tables as organisers	1	1	1	2	5
13	Reliability and validity of design	6	4	5	2	17
14	Data presentation	1	1	1	2	5
15	Statistical treatment of differences	0	0	0	0	0
16	Relationships	1	1	0	1	3
17	Comparison with other data	3	3	3	3	12
Number of actions		**36**	**36**	**33**	**28**	**133**

the concepts of evidence required for Types 4 and 8, since some ideas are not relevant, e.g. range and interval of values of the independent variable.

The implications of this analysis are that, since all eight of the investigation types operate on the same underlying understandings, then an assessment based on those understandings, rather than the particular series of actions needed to carry them out, will be far less restrictive.

Summary and discussion

In summary, the analysis and argument leads us to the following position:

1. As a consequence of the necessary requirements for reliability in a performance-based approach, the range of acceptable biology investigations has become unnecessarily restricted to particular tasks in our eight-fold typography.
2. This can be traced to an assumption of a skills approach to Sc1.
3. The identification of a knowledge base (concepts of evidence) allows more types to be analysed (and therefore assessed – see below for more) against this knowledge base rather than against a series of actions in a performance based approach.
4. Different types of task utilise very similar concepts of evidence, thereby allowing the concepts of evidence to be seen as assessment criteria, but with some differences of emphasis and in differing sequence.

Let us now think of the curriculum issues that underpin the problem. We will take the position here that:

1. Investigations, of all types, should be an important element of teaching and learning in science.
2. That assessment should have as little negative backwash as possible.
3. Assessment should be as reliable as possible and not impose unnecessary bureaucratic burdens on teachers, burdens that militate against the very use of investigations that the NC is seeking to promote.
4. But, above all, we must have a system, which assesses the underlying understandings about evidence — its collection, interpretation and evaluation.

It would still be possible to meet these points with a skills approach to Sc1. The assessment criteria could be refined to take into account all the possible actions in different contexts but it would become inevitably unwieldy if criteria in every context were specified. By comparison, assessment against the Concepts of Evidence that are (relatively) few in number is a potentially more robust and easily understood assessment system.

If we look back to our diagram (see Figure 3) we see that a move to an understanding approach brings with it a number of possibilities for assessment — we repeat the diagram here (see Figure 4) with some annotations to indicate the sort of things we have in mind.

Understanding-based approaches

The identification of a knowledge base frees us from the shackles of a performance-based, routinised assessment. We can now think of the carrying out of an investigation, not as the thing to be assessed, but as one mode through which concepts of evidence can be assessed. Immediately we can see that there are several such modes:

- Complete practical investigations
- Parts of investigations
- Computer based simulations
- Written assessments (exams)

Which mode would satisfy our curriculum criteria? There is a general acceptance that the current method of practical assessment is causing difficulties (Keiler and Woolnough, 2002). The report of the House of Commons Select Committee on Science and Technology (2002) is just the latest to comment on the 'stultifying assessment arrangements' (p. 5) and that currently 'coursework is boring and pointless' (p. 5) and 'has little educational value and has turned practical work into a tedious and dull activity for both students and teachers' (p. 57). This points, as we have argued elsewhere, to the need to think seriously about teaching and assessing ideas about evidence.

There is an extensive literature discussing different modes of assessment of investigations (e.g. Baxter *et al.*, 1992; Welford *et*

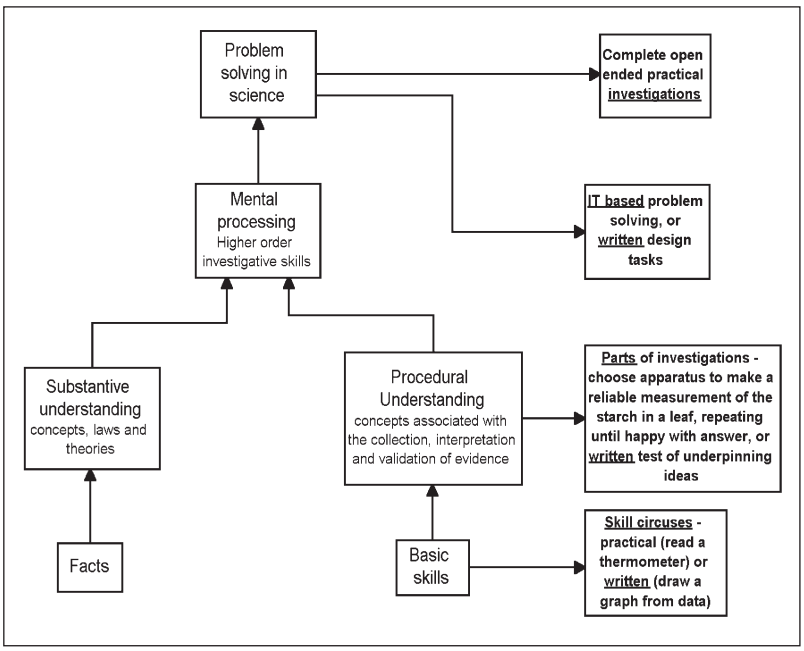

Figure 4 *Possible assessment tasks.*

al., 1985; Gott and Duggan, 2002; Solano Flores *et al.*, 1999; Ruiz-Primo and Shavelson, 1995). We shall discuss just one possible mode of assessment, the use of written 'exams' assessing pupils' understanding of the ideas of evidence, because this is perhaps the approach which could be seen as most problematic by teachers and assessors and is the mode being introduced into KS3 SATs in 2003 (QCA, 2002).

Written papers are likely to be more reliable than assessment of performance. The argument against them has been that they are not a valid measure of practical work. But if we shift our emphasis to an understating of Concepts of Evidence, then there is no reason why they should not be valid assessments of that knowledge base. Of course, we would not then be claiming to assess *performance* in this written paper. However, a case could be made for assessing the Concepts of Evidence in a written test as a 'least worst option', potentially avoiding the problems associated with current performance assessment. Such a radical change in the assessment system would need to be thoroughly researched. More data is required on the level of difficulty of different concepts of evidence as well as how questions can target different concepts. We have carried out research into the latter, which is the subject of a separate paper.

What about the problem of retaining investigations in the curriculum? We are definitely not advocating the demise of investigative work just because assessment is by written paper; in fact, quite the opposite. We agree with the House of Commons Select Committee on Science and Technology (2002) recommendation that 'fieldwork should be strongly recommended in all courses' (p. 57) and that 'practical work, including fieldwork, is a vital part of science education. It helps students to develop their understanding of science…[and] appreciate that science is based on evidence.' (p. 57). We can now see that investigative work is one (and possibly the best in many situations) way of *teaching* about evidence. It does not necessarily follow that it is the best way of *assessment*.

The current assessment system has affected teaching to the extent that many teachers feel in a straitjacket (Nott and Wellington, 1999; Donnelly, 2000). Assessment by 'case law type' has so distorted the teaching of Sc1 that in many schools the only practicals taught are those used for assessment (Keiler and Woolnough, 2002) and teaching, particularly biology teaching, has become distorted. If assessment was by written paper on work that pupils will have been expected to do in various contexts in (and out of!) school, this would return the 'how to teach' to its rightful place, in the hands of the teachers exerting their own professional autonomy in response to the circumstances in which they are teaching.

References

Baxter G P, Shavelson R J, Donnelly J, Johnson S and Welford G (1988) Evaluation of procedure-based scoring for hands-on assessment. *Journal of Educatonal Measurement*, **29,** 1 – 17.

Bencze J L (1996) Correlational studies in school science: breaking the science-experiment-certainty connection. *School Science Review,* **78,** 95 – 101.

DfEE/QCA (1999) *The Science National Curriculum*. London, UK: DfEE/QCA.

Donnelly J (1995) Curriculum development in science: the lessons of Sc1. *School Science Review,* **76,** 95 – 103.

Donnelly J (2000) Secondary science teaching under the National Curriculum. *School Science Review,* **81,** 27 – 35.

Gott R and Duggan S (2002) Problems with assessment of performance in practical science: which way now? *Cambridge Journal of Education*, **32,** 183 – 201.

Gott R and Mashiter J (1991) Practical work in science – a task-based approach? in *Practical Science.* Woolnough, BE (ed). Buckingham, UK: Open University Press.

House of Commons, Science and Technology Committee (2002) *Science education from 14 to 19*. London, UK: The Stationery Office.

Keiler L S and Woolnough B E (2002) Practical work in school science: the dominance of assessment. *School Science Review,* **83,** 83 – 88.

Nott M and Wellington J (1999) The state we're in: issues in Key Stage 3 & 4 science. *School Science Review,* **81,** 13 – 18.

QCA (2002) *National Curriculum Assessments from 2003*. Letter from QCA to all Heads of Science June 2002.

Ruiz-Primo M A and Shavelson R (1995) *Rhetoric and reality in science performance assessment: an update.* Paper presented at American Educational Research Association, San Fancisco, USA.

Roberts R (2001) Procedural understanding in biology: the thinking behind the doing. *Journal of Biological Education*, **35,** 113 – 117.

Roberts R and Gott R (1999) Procedural understanding: its place in the biology curriculum *School Science Review*, **81,** 19 – 25.

Roberts R and Gott R (2000) Procedural understanding in biology: how is it characterised in texts? *School Science Review*, **82,** 83 – 91.

Solano-Flores G, Jovanovic J, Shavelson R J and Bachman M (1999) On the development and evaluation of a shell for generating science performance assessments. International *Journal of Science Education*, **21,** 293 – 315.

Watson J R, Goldsworthy A and Wood-Robinson V (1999a) What is not fair with investigations. *School Science Review,* **80,** 101 – 106.

Watson J R, Wood-Robinson V and Goldsworthy A (1999b) Improving investigations. *Education in Science,* November 1999.

Watson J R, Goldsworthy A and Wood-Robinson V (2001) Sc1: beyond the fair test. In (Eds) Sears J and Sorensen P, *Issues in science teaching,* London, UK: Routledge Farmer.

Welford G, Harlen W and Schofield B (1985) *Practical testing at Ages 11, 13 and 15*. London, UK: Department of Education and Science.

Ros Roberts is a Lecturer in Science Education in the School of Education, University of Durham, Leazes Road, Durham DH1 1TA, UK. Tel: +44 (0) 191 334 8394; Email: Rosalyn.Roberts@durham.ac.uk. Richard Gott is Professor of Science Education also in the School of Education.