Assessment of Sc1: alternatives to coursework

Ros Roberts and Richard Gott

Current methods of assessing investigative work are not working. How can we do it better?

ABSTRACT
This article considers the problems associated with reliable performance assessment of Sc1 investigations and explores the pros and cons of alternative forms of assessment of pupils' ability to investigate.

The assessment of investigative work in the National Curriculum for England and Wales, science attainment target 1 (Sc1) has been problematic since its introduction in 1989. Sc1 has undergone a number of changes of detail but remains a significant part of the assessment of pupils aged 11 to 16. Recently, assessment of practical work was identified as a concern in a report from the UK House of Commons Science and Technology Committee (2002), which made a number of trenchant criticisms of the assessment arrangements for GCSE (the end-of-compulsory-school examination system for 16-year-olds). The report described Sc1 coursework, which consists of whole and part investigations, as '*boring and pointless*' (p. 50) and also said that:

> *The way in which coursework is assessed for GCSE science has little educational value and has turned practical work into a tedious and dull activity for both students and teachers.* (p. 21)

In this article, which follows on from an earlier one on the role of different types of practical work (Roberts, 2004), we restrict the discussion to the assessment of investigations which involve variable-based tasks where the explicit focus of the activity is the search for a link, causal or otherwise, between two or more variables and for which there is no easily recalled solution. We also assume that teacher assessment of investigations is not an option for reasons of time, reliability and politics.

So, how are investigations to be assessed?

Assessing performance

Assessing how well pupils perform investigations is a complex business. If you imagine yourself doing an unfamiliar investigation you will find yourself trawling around in your head (or books or other information sources) for ideas of all sorts that might help. They might be ideas about melting point, or velocity. They might be about designing suitably fair tests, or how repeatable a measurement is likely to be. They might be mechanical skills to do with assembling apparatus. You might look for similarities with another task and try to lift a whole design and amend it as you go along.

How can it be possible for somebody else to get at what is going on in your head while you're doing this? That is, after all, what assessment is supposed to be about.

By observation and interview

Assessing this complex activity is a research task in itself. It would involve detailed checklists and interviews with pupils during or after their investigations (similar to the assessments done by the Assessment of Performance Unit (APU) – see Gott and Murphy, 1987). Whilst this would be a very valid assessment, it is simply not a practical proposition for a large-scale assessment system. Not only would it take up far too much time and resources, but there are also issues of reliability – of which more later.

By using written reports

If we cannot interview and observe, what can we do? The next obvious step is to ask pupils to record what they do, and assess that. Although Welford, Harlen and Schofield (1985) tentatively noted in their APU work that older pupils were fairly accurate in their reporting, Baxter *et al*. (1992) found that inexperienced pupils showed a low level of agreement between observation and the report. Training, however, resulted in a reasonable correspondence between actual performance and the pupils' reports.

So far so good. Using pupils' own written records for assessment seems a reasonable substitute for watching what they do. We are substituting assessment of the self-report of the pupils' performance for assessment of the performance itself. But there is still a fundamental problem which revolves around issues of reliability.

Reliability

The purpose of assessment is to 'measure' how much a pupil knows about a topic or subject, say physics. To do this, tests contain many questions that together provide a measure of what 'physics' is inside the pupil's head. Imagine a test consisting of just a single question about anything – electricity, radiation, force or whatever. That one question may, quite by chance, happen to be about something the pupil knows because they were taught it yesterday. On the other hand, it might be something which the pupil missed because they were away that day. One, or a very few questions, is not a good way of measuring a pupil's understanding of 'physics'. Therefore, lots of questions are used in tests to get a reliable overall picture of pupils' understanding of physics and to act as a predictor for future performance.

Let's take an example nearer to our topic: an investigation on forces is quite likely to be one that one of us (RG) could have a reasonable stab at. If it were the ecology of a stream, then it would be much more like starting from scratch – there wouldn't be as many ideas to call on. So which would be the better test of ability to do investigations? Neither really. Ideally, we need to do lots of investigations in different contexts and then average them out – which is how written tests (particularly multiple-choice ones) can be claimed to be a reliable predictor of performance.

Research suggests that the 'context effect' (including the subject matter of the investigation, the setting or context, for example lab or field) and the 'procedural complexity' (such as the variable structure and type of variables, the degree of interaction with the apparatus required and the openness of the task) are so great that you would need up to 10 assessed investigations to be reasonably sure that the result was a reliable predictor of future 'ability to do investigations'. This is unrealistic given the constraints of an overcrowded curriculum.

Reducing problems of unreliability

Various approaches have been taken to obtaining an appropriately reliable assessment of performance using a more manageable number of tasks, say, two or three.

Specifying the task more closely

Solano-Flores *et al.* (1999) attempted to reduce the variation due to subject matter and context in assessment tasks. They did this by creating 'shells' (which were effectively templates for designing the task) that took account of this 'procedural complexity' in the construction of assessment tasks (for example by defining how many variables were to be included, how much was provided in the way of hints). However, they found that there was still considerable variation in pupils' performance on tasks of apparently similar demand. The intention was to take out some of the variability in the task and leave behind a measurement of performance that could be shown to be reliable for a smaller number of tasks – a bit like short-response items which limit options and guide pupils into making certain sorts of answers. But it didn't help all that much: the tasks were so complex that too much variability remained.

Making the tasks routine

We have argued that custom and practice with Sc1 assessment has resulted in a similar reduction in variation between tasks. Assessment criteria have been exemplified in particular contexts so that, over the years, a repertoire of standard items has, necessarily, built up. This repertoire suggests acceptable procedures and contexts in which the criteria can be demonstrated and which are aligned with moderation examples. This goes beyond the 'shells' of Solano-Flores into almost a 'seen exam' situation. From this it could be argued that Sc1 assessment has become routine, with a limit on the number of cases assessed. In some instances, Sc1 coursework has become so formulaic that performance is more akin to the recall of a complex protocol than the creative solution of a problem.

A serious downside of these attempts to increase the reliability of accounts of investigations is the resulting narrow experience of investigative work and the consequent reduction in the validity of the tasks (Donnelly *et al*., 1994; Donnelly, 2000; Gott and Duggan, 2002; Watson, Goldsworthy and Wood-Robinson, 1999) and the effect this has on the curriculum.

Current coursework assessment, for very good assessment reasons, has become an assessment of the recall of complex protocols in a limited number of 'standard' investigations; not necessarily what was intended.

Use it to measure 'content' instead

Another response to the problem of the reliability of the assessment of Sc1 in the UK has been to resort to assessing the pupil's understanding of the underlying substantive concept. So, for example, in the assessment of the planning of the investigation or of the analysis of the results much credit is assigned to using detailed scientific knowledge and explanation (Gott and Duggan, 2002). It is not sensible to assess substantive ideas using this cumbersome method when a well-honed set of tests (GCSEs and SATs) is available (whatever their shortcomings). It also results, of course, in relatively less weight being attached to the assessment of procedural understanding.

Summary

The current assessment system seems not only to be failing to assess all of the elements required for problem-solving in science but is also in danger of distorting teaching so that the elements may not be taught either.

So, we are in a position where:

- valid assessment using observation, checklists and interviews is not possible;
- using self-reports is possible, but

  – too many assessments are required unless they become standardised, or

  – assessment focuses on the substantive ideas, both of which make the assessments more reliable but they are no longer measuring what we set out to do – the 'ability to do investigations'.

There seems to be no easy answer to the problems of performance assessment. So are there any other options?

Assessing planning rather than performing

One option would be to assess pupils' paper-and-pencil planning of whole investigations without their actual performance. The APU (Gott and Murphy, 1987) asked pupils to write a plan of how they would carry out an investigation presented to them either in prose or using pictures. Pupils responded better to the pictorial clues than the text but they found both of these more difficult than actually carrying out the investigation.

Other research (Gray and Sharp, 2001) into pupils performing practicals and doing just paper-and-pencil tasks seems to indicate that results are consistently better on the practical tasks. Not only do pupils seem to find written tasks harder, there is some evidence that they are actually assessing something different (Baxter *et al.*, 1992; Lawrenz, Huffman and Welch, 2001) – practicals seem to bring out something in pupils that other tests don't!

Writing about whole investigations without performing them seems to be hard, and therefore would not be a very good discriminator for assessment purposes. The validity of the written task must also be questioned, since conducting investigations is an iterative process; decisions are made in response to what is happening and many cannot be predicted without reference to the situation as it is being performed.

Assessing performance of simulated investigations

An alternative to assessing performance of a practical investigation would be to assess performance of simulated investigations based on CD-ROMs. Of course, there is no guarantee that pupils who could use ideas in a practical situation would be able to cope with them on a CD-ROM, or vice versa. Gott and Duggan (2003) have developed a CD-ROM that allows pupils to carry out investigations, making the same decisions as they would if they were actually doing the practical themselves. They have to decide which variables to control, which measuring instrument to use, how many repeats are required, which is the best way to present the data, and so on. Other, similar, developments are under way through commercial publications and exam boards. They have the advantage of being quick and may enable more tasks to be assessed, thus addressing some of the problems of reliability mentioned above.

Further work would be required to see how these ICT tasks could be used as an assessment tool. Tentative results have shown that performance is affected significantly by the computer interface. Gott and Duggan (2002) found that pupils collected far more data than they could handle, in contrast to practical tasks where time limited the amount of data that could be collected. Such simulations of whole investigations may offer an additional assessment tool for the future. They at least involve pupils in making similar decisions to those needed in a practical investigation. Much work needs to be done if such tasks are to be used in high-stakes assessment such as SATs and GCSEs.

Assessing parts of an investigation

All that we have described so far are attempts, in various ways, to assess performance through observation, self-report, or intended performance through a plan. We have shown that these ways of assessing whole investigations are not without their problems. We next turn our attention to the possibility of assessing *parts* of that performance to see if that could offer a solution.

As we exemplified in our previous article (Roberts, 2004), problem-solving consists of a synthesis of different elements, including skills, observations, design and measurement. Instead of assessing the whole investigation, these smaller parts can be the focus of assessment. Each of these elements can be assessed, as described above, by practical performance, a written task or using ICT. They have the advantage, assuming they are run as discrete assessment tasks, perhaps in a circus, of being cheaper and quicker than assessment of whole investigations, and more reliable because more instances can be assessed. Thus a pupil might attempt several different observations, skill practicals and measurement tasks, covering a range of contexts, with increased reliability of the final score. This would also address some of the issues of validity, although the synthesis of ideas required in a whole investigation would obviously not be possible. (Of course, if pupils still carry out whole investigations but only small sections of them are used for assessment, the advantage of being able to do many different tasks has been lost.)

Such short tasks might include specific practical skills and protocols, observation tasks, design tasks or measurement activities. This has been tried in one form or another in the past – for instance, by NEAB Biology GCSE and APU (Welford *et al*., 1985). These shorter tasks may vary in their level of difficulty. Some may require only recall of skills, others understanding and application of ideas, and some the synthesis of ideas, albeit in a much reduced task compared with a whole investigation.

Assessment of parts of an investigation has certain advantages over whole investigations:

- They have been tried in one form or another and are therefore realistic possibilities.
- They reduce the procedural complexity of the task: the pupils don't have to keep as many ideas at a time in their head, which is difficult (and may be another factor driving the simplification of whole investigations to ritual recall of protocol).
- It may result in a focus on the procedural ideas rather than resorting to routine investigations.
- Each task would involve less time than whole investigations, allowing more to be done.

But there is still the issue of the organisational logistics for more assessed practicals, the associated time required and the hassle of marking and moderation. Would there be sufficient gains, in assessment terms, for this cost? This has led us to consider a rather different approach to assessment.

Assessing the 'thinking behind the doing'

What we have considered so far has largely been concerned with attempts to assess ideas by looking for evidence of their use within the context of investigations, whether whole investigations or parts. A more radical approach is to detach the ideas from the investigation and make the 'thinking behind the doing' the focus of the assessment. What would this look like? And does it have potential advantages?

Underpinning the assessment of the 'thinking behind the doing' is the premise that problem solving in science requires not only an understanding of the substantive ideas of science (ideas such as force, niche and chemical change) but also a procedural understanding that involves a knowledge-base of ideas to do with evidence. Whole, and to some extent part, investigations involve the synthesis of all these ideas. Assessment of these, whether practically, or using written tasks or ICT, is relatively time-consuming and problematic as we've shown above. An approach we have taken recently is to write a short test that

specifically focuses on the ideas of evidence. These ideas are a necessary part of doing investigations as well as evaluating other people's evidence. The test items do not assess a synthesis of all the ideas in a problem-solving context but target the understanding and application of the knowledge-base that forms the procedural understanding. Box 1 gives an example of a question from the test to show what we mean. The test was 50 minutes long and was given to pupils half-way through their GCSE course, with questions targeted at procedural ideas such as variable structure, the appropriateness of measuring instruments, and why repeated readings were necessary. We also asked them to make judgements about data, as well as asking questions in contexts that are not normally acceptable for Sc1 coursework.

While the written evidence test would need further development to ensure its reliability before it could be used on a wider scale, we found that a test to target procedural ideas was feasible, that the pupils were engaged with the ideas and, as one teacher put it, '*it made them think*'.

Results from the test were interesting in that some pupils performed very differently on the evidence test to the coursework assessment. We found that there was a link between pupils' Sc1 coursework grades, allocated for the carrying out and writing up of investigations, and both their verbal reasoning scores and an attribute that we called 'behaviour/conformity'. For instance, the teachers described a pupil who did well on the coursework as '*Sometimes anxious and unconfident. Very neat. Prepares well. In top set through hard work. Probably learns by rote.*' Coursework assessment was not linked to other attributes including non-verbal reasoning and 'quickness of thought' which we found worrying. On the other hand, the evidence test results were found to be linked to 'quickness of thought'. A pupil who had this attribute was described as '*Very able. Can think on his feet. Has an anti-school attitude. Doesn't get coursework done.*' The evidence test seems to identify strengths in some pupils that were not rewarded in the current coursework assessment.

Box 1 Example of an item from a written evidence test

You are a French wine grower and you are getting complaints that the taste of some of your best-selling wines is varying. You decide to try and find out what is causing the taste to be different. You start by considering the type (or species) of grape and where the vines are gowing (north, south or east facing fields) in your vineyard.

White grapes **South-facing field** **Field A**	**Green grapes** **South-facing field** **Field B**	**White grapes** **East-facing field** **Field C**	**Pale green grapes** **South-facing field** **Field D**	**White grapes** **North-facing field** **Field E**

A If you want to find out if it's the **type of grape** that makes a difference to the taste, which fields would you use for the comparison? Explain why.

B If you want to find out if it's the **direction of the slope** that makes a difference to the taste, which fields would you use for the comparison? Explain why.

The advantage of such a test is that it is quick to administer and many of the procedural ideas in different contexts can be targeted. While it is obviously not the same thing as actually doing a practical investigation it has the advantage of freeing up the teaching time that is currently spent doing routine coursework assessment, allowing teachers to decide how best to teach about problem-solving in science.

Conclusion

As Gott and Duggan (2002) concluded, '*there is no easy solution to the assessment problem*' (p. 197). Different assessments seem to assess different things with differing degrees of validity and reliability. We need to decide which provides the best 'measure' at the least cost (in terms of distortion of the curriculum, teaching time, bureaucracy etc.).

If current practical assessment is not working (and we assume here that the reintroduction of teacher assessment is still not on the political agenda) then it may be that a written evidence test is the best we can do, and it does have advantages. One of these may be that we have a way of identifying pupils who might currently be seen as under-performing.

Of course, not having any assessment of practical work may not be ideal. But the current Sc1 assessment is so problematic, taking up inordinate amounts of class time which could be spent teaching, that the alternative might be better. The current situation is even more galling when the validity of the Sc1 tasks is so suspect. If only more time were available, teachers could select the most appropriate ways of teaching the pupils, which might include a far more varied set of practical work: space could be created for stimulating investigations in different contexts.

Would this mean practical work would not be used in school? In our view, this would be most unlikely. Removal of Sc1 investigations as assessment rituals may well enrich the pupils' experience and the control of how to teach would be returned to the teachers where it belongs.

References

Baxter, G. P., Shavelson, R. J., Goldman, S. R. and Pine, J. (1992) Evaluation of procedure based scoring for hands-on assessment. *Journal of Educational Measurement*, **29**(1), 1–17.

Donnelly, J. (2000) Secondary science teaching under the National Curriculum. *School Science Review*, **81**(296), 27–35.

Donnelly, J., Buchanan, A., Jenkins, E. and Welford, A. (1994) *Investigations in science education policy: Sc1 in the National Curriculum for England and Wales*. Leeds: University of Leeds.

Gott, R. and Duggan, S. (2002) Problems with the assessment of performance in practical science: which way now? *Cambridge Journal of Education*, **32**(2), 183–201.

Gott, R. and Duggan, S (2003) *Building success in science 1*. London: Folens.

Gott, R. and Murphy, P. (1987) *Assessing investigations at 13 and 15*. Science report for teachers, 9. London: Department of Education and Science/HMSO.

Gray, D. and Sharp, B. (2001) Mode of assessment and its effect on children's performance in science. *Evaluation and Research in Education*, **15**(2), 55–68.

House of Commons, Science and Technology Committee (2002) *Science education from 14 to 19*. Third report of session 2001–2, **1**. London: The Stationery Office.

Lawrenz, F., Huffman, D. and Welch, W. (2001) The science achievement of various subgroups on alternative assessment formats. *Science Education*, **85**(3), 279–290.

Roberts, R. (2004) Using different types of practical within a problem-solving model of science. *School Science Review*, **85**(312), 113–119.

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**(3), 293–315.

Watson, R., Goldsworthy, A. and Wood-Robinson, V. (1999) What is not fair with investigations? *School Science Review*, **80**(292), 101–106.

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

Ros Roberts is a lecturer in science education and **Richard Gott** is professor of science education at the School of Education, University of Durham, Leazes Road, Durham DH1 1TA.
E-mail: Rosalyn.Roberts@durham.ac.uk