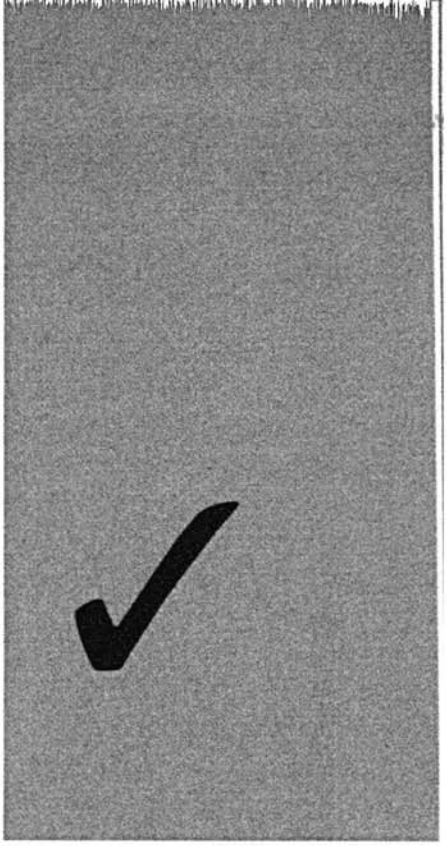
TIMSS puts England first on scientific enquiry, but does pride come before a fall?



Per Morten Kind

England's 13-year olds scored highly in the TIMSS survey of 1995 so why do they still display poor understanding of the nature of science?

The Third International Mathematics and Science Study (TIMSS) included a performance assessment study (TIMSS PA). This was a practical test based on problem-solving and investigative tasks in both mathematics and science. The overall international results for 13-year-olds show Singapore on top (71% of maximum score), but on the science tasks England is equal to Singapore and in certain areas they score far better than most other countries (see Harmon *et al.*, 1997 for a full report). For example, on the investigative task *Solution* (Investigate what effect different water temperatures have on the speed with which the tablet dissolves) England scored 68%, while the international average was 49%. Similar results were achieved for the tasks named *Pulse* (Find out how your pulse changes when you climb up and down on a step for 5 min) and *Rubber band* (Find out how the length of the rubber band changes as more and more rings are hung on it). For those who know of the tradition of practical work in English schools, with its emphasis on science enquiry, this is perhaps no surprise. England, more than most other countries, has been through several stages of development of 'process-oriented' and 'investigative' school science. TIMSS PA reveals some of this tradition, partly through the results as described above, but also through the test format that was used in the study. This article therefore will use the study to reflect on how practical investigative work is taught and assessed in school science.

ABSTRACT

England, more than most other countries, has been through several stages of development of 'process-oriented' and 'investigative' school science and has put considerable effort into developing this approach both in teaching and assessment. In the 1995 TIMSS performance assessment study English 13-year-olds received high scores for problem-solving and investigative skills in science. However, students in England still seem to have a poor understanding of the nature of science. This article examines to what extent this performance assessment tested real understanding of scientific process and reflects on how practical investigative work is taught and assessed in school science.

Process skills and performance assessment

'Doing' is sometimes contrasted with 'knowing'. We are likely to talk about 'doers' and 'thinkers'. In science education we sometimes find a similar contrast between working in the laboratory and reading and studying in the classroom. The latter is regarded as using knowledge and understanding while the former is seen as using skills. This contrast, however naive it sometimes might be, seems to have dominated certain rationales for practical work in science. In particular it was taken to the extreme in the pedagogical trend referred to as the 'process approach' (DeBoer, 1991). This trend was very much based on Gagné's analysis of cognitive processes used by scientists and his ideas for developing the same processes in science education (Gagné, 1965). Several school science projects used this as a basic philosophy and found ways of teaching students skills. The science processes were seen as important mainly because of their contribution to developing students' thinking and learning abilities, and were given a high status:

> *It could be said that the most valuable elements of a scientific education are those that remain after the facts have been forgotten.*
> (Screen, 1986, Introduction)

Several aspects of the process approach have been commented on and criticised (Millar and Driver, 1987; Hodson, 1993), and today we find more attempts to build bridges between understanding and doing in the laboratory (Millar, 1995).

The process approach, however, seems to some extent to have been carried on in the area of performance assessment in science. Here we still find frameworks identifying performance, what you do, and knowledge, what you understand, as rather separate dimensions of science. No radical new framework seems to have been presented, at least for large-scale performance assessment, since the early attempts developed under the process skill rationale.

Among the first influential frameworks we find Klopfer's categorisation of processes of scientific inquiry (Klopfer, 1971). This was used as a base for developing a performance test in the IEA's (International Association for the Evaluation of Educational Achievement) Second International Science Study in 1984 (Tamir, Doran, and Chye, 1992). An even more influential framework was developed by the APU (Assessment of Performance Unit) Science Project (APU, 1984). This framework, in spite of an effort to combine process and concept, laid a foundation that led to many later performance-assessment projects being process-related. This was partly due to presenting their '*view of science*' as '*an experimental subject concerned fundamentally with problem-solving*' (APU, 1984, p. 4).

In the categorisation framework for APU science a distinction was made between three dimensions of science learning:

- the *science processes* involved in solving a task;
- the *conceptual understanding* required for its solution;
- the *context* in which the task is set.

The process dimension was regarded as '*the key facets of science*' and the idea was to assess the '*general skill*' detached from any particular content or context (APU, 1984).

To develop categories for the process dimension, skills related to each step in a science investigation were identified, as illustrated in Figure 1. The boxes on the left-hand side show actions or activities carried out by students from being given an investigative task to reaching a final conclusion. Each of these steps is related to specific skills or abilities (on the right-hand side). In the APU framework for scientific performance (APU, 1984, 1989) six such assessment categories were developed:

1 Use of graphical and symbolic representation
2 Use of apparatus and measuring instruments
3 Observation

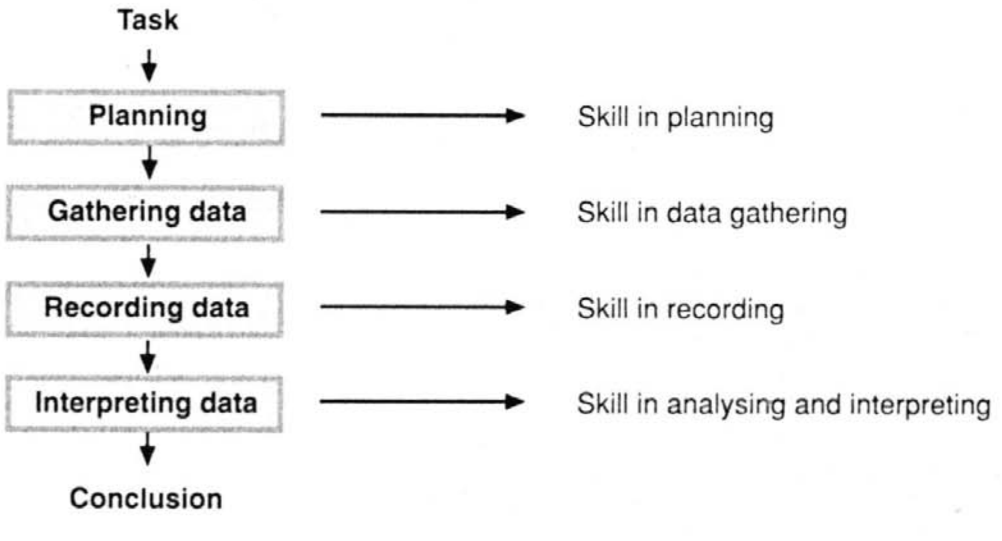


Figure 1 Developing assessment categories for performance assessment.

4 Interpretation and application

5 Planning an investigation

6 Performance of investigation

Inclusion of a separate category for carrying out a whole investigation (category 6) was justified with the argument that this encompasses more than the simple sum of the elements represented in the other categories (APU, 1989). Each of the main categories has been broken down further into subcategories describing more detailed skills.

As mentioned above, APU Science has been important for many later performance assessment projects, as well as for the practice of assessment of science in England in general. This is both because of the ease of use of the framework, and the large number of tasks that were developed and the test procedures based on these.

TIMSS performance assessment

TIMSS PA was an international study, involving 21 countries and containing both mathematics and science tasks. The selected age groups were 9-year-olds and 13-year-olds, but only ten of the countries tested the first of these. A minimum sample of 450 students was tested at each age level in each country. The students underwent a 90-minute test session visiting three workstations, each with either one 30-minute task or two 15-minute tasks. Nine students were tested in each session, rotating among a similar number of workstations. In total there were 12 tasks for each age level. The tasks were named according to their content.

The science tasks were all investigative tasks intended to measure a mixture of enquiry skills. The large-scale test format used a written response-sheet containing instructions and questions about the investigation to be answered by the students while working on the tasks. For instance, the task *Solution* (Investigate what effect different water temperatures have on the speed with which the tablet dissolves) had the following sub-items:

> *1. Write your plan here. Your plan should include*
>
> *– what you will measure*
>
> *– how many measurements you will make*
>
> *– how you will present your measurements in a table.*
>
> *2. Carry out the tests on the tablets. Make a table and record all your measurements.*
>
> *3. According to your investigation, what effect do different water temperatures have on the speed with which a tablet dissolves?*
>
> *4. Explain why you think different water temperatures have this effect.*
>
> *5. If you had to change your plan, describe the changes you would make and why you would make them. If you did not have to change your plan, write 'No change'.*

For some of the tasks students were also asked to hand in something they had made, for example boxes made in a mathematics task. Students' responses were scored and coded according to a coding system that allowed for the identification of common approaches and types of errors in student responses (Harmon *et al.*, 1997).

The links between TIMSS PA and APU Science are obvious even if the projects differ in purpose and style. TIMSS adopted tasks similar to those used by APU, and the test procedures and the frameworks tended to focus on much the same categories.

A secondary analysis of TIMSS PA results

To learn more about practical performance assessment, a secondary analysis of results from the Norwegian sample in TIMSS PA was conducted. The analysis aimed to find out how students responded to similar sub-items in different tasks. Following science process thinking each sub-item in the TIMSS PA tasks could be regarded as measuring specific scientific skills. For example, the task *Solution* presented above could be seen to measure the skill of planning (sub-item 1), the skills of data gathering and presentation (sub-item 2), and the skill of data analysis (sub-item 3), and so on. A common approach in performance assessment has been to aggregate scores for such skills across tasks. In this analysis the correlation between two sub-items that were supposed to measure the same skill was calculated. If these two sub-items really measure the same skill we should find a relatively high positive correlation (for an exact match the correlation would be +1). Since we have several tasks with sub-items measuring the same skill a mean was calculated for all the correlations (based on a correlation matrix for each skill).

Table 1 Mean correlation between measures of the same skill.

Skills	*Mean correlation between measures of the same skill* *Science*	*Maths*	*All tasks*
Using equipment	-0.05	0.24	0.07
Using routine procedures	0.12	0.24	0.13
Data collection; Measurements	0.13	0.33	0.11
Data organisation and representation	0.22	0.37	0.17
Data analysis	0.10	0.15	0.08
Data interpretation	0.03	0.19	0.06
Application; Problem solving	0.10	0.22	0.17
Designing experiments	0.12	0.26	0.18
Concept understanding; Use in explanation	0.10	0.17	0.11
Communication	0.12	0.17	0.13
Average across skills	**0.10**	**0.23**	**0.12**

Table 2 Summary of mean correlation between sub-items within tasks. The last column shows a reliability index (Cronbach's α) indicating the consistency of each task. This value ranges from 0 to 1, where 1 indicates maximum consistency.

Task	*No. of sub-items in the task*	*Mean correlation between sub-items*	*Reliability*
Science			
S1 Pulse	4	0.29	0.58
S3 Batteries	4	0.28	0.56
S4 Rubber band	7	0.13	0.47
S5 Solutions	7	0.16	0.62
SM1 Shadows	6	0.13	0.49
Average across science tasks		**0.20**	
Mathematics			
M1 Dice	6	0.21	0.59
M2 Calculator	7	0.22	0.67
M3 Folding and cutting	4	0.57	0.82
M4 Around the bend	8	0.19	0.65
M5 Packing	3	0.44	0.68
SM2 Plasticine	8	0.33	0.82
Average across maths tasks		**0.33**	

The results from this analysis are presented in Table 1. The skills are all taken from the process dimension of the TIMSS PA framework (Robitaille *et al.*, 1993), and separate values are calculated for mathematics and science tasks. It can be seen that, for example, the mean of all the correlations between the skill *Using equipment* is –0.05 for science and 0.24 for mathematics. The mean for both science and mathematics is 0.07. Table 2 presents similar correlation values for all sub-items *within* tasks.

The overall picture from the results shows no strong link between what purport to be measures of the same skill (see Table 1). For example, there is little evidence that the items on data collection in the tasks *Pulse* and *Solution* measure the same skill. One explanation of this might be that the two tasks measure different aspects of a relatively broad skill that we call data collection. However, there seems to be a somewhat stronger link between sub-items measuring *different* skills within each task (see Table 2). Even these correlations are rather low and, with some few exceptions, the strengths of correlations between similar sub-items are at the same level as that between sub-items selected randomly (0.12).

There are some differences between mathematics and science, with somewhat higher correlation values for mathematics. A closer study, however, indicated that correlations contributing to a high mean value were mostly found *within* tasks. The mathematics tasks often consisted of several sub-items asking for the same type of procedure, e.g. making several measurements on a set of objects. Such a procedure focuses on a fairly narrow skill such that a candidate is likely to perform much the same in different tasks. Even under these circumstances the correlations are not particularly high, which seems to point to the effect of context and content on performance in the tasks.

Mismatch between rationales for science enquiry

What can be learned from the results presented above? At least it tells us that we should be careful about interpreting what the tasks really measure. We cannot assume the tasks directly measure specific skills. To many teachers this will be rather obvious. Solving a practical task is a complex affair and what students actually do depends on many details, such as equipment, hints given in the text, subject matter knowledge, and so on. We could, therefore, argue for more focus to be placed on each task as an 'entity' and for holistic assessment (Woolnough, 1989).

There is, however, another and more serious matter related to this mismatch between theory and results that goes beyond the practical complexity of performance assessment. Do students use investigative skills at all as we intend? Scientific skills only have meaning within a certain rationale: a problem or phenomenon is studied systematically with structured planning and systematic data gathering *because* scientists want to find evidence for whatever conclusion they reach. This is also apparent when scientists present their research to others, in that they give a detailed presentation of the method used, include data to support their conclusion, and discuss their research critically. We sometimes find the scientific approach presented as a form of 'argumentation' (Driver, Newton, and Osborne, 2000), where scientists use data to 'argue' for their conclusion. This logic is also tacit in scientific investigative work in the school laboratory, and in science performance assessment, but is it part of the students' rationale for doing science? The data presented above do not directly answer this question, even if they tend to reveal a lack of internal validity in students' responses. Another analysis (Kind, 1996), however, has revealed more clearly that many students draw conclusions that only partly match their own data presentations. This analysis was based on students' actual responses from a TIMSS PA pre-test in six different countries, including England. It sometimes seemed that students were working on two different problems. First, they were trying to answer the problem given in the task, e.g. find the effect of temperature on a dissolving tablet, measure and describe your pulse, find which battery is uncharged, etc. This may be seen as 'common-sense' problem solving. Second, they were trying to make the task *scientific*. What students see as 'scientific' is not necessarily the same as we intend. Some students seemed not to have reflected much on this issue at all; for instance, it was found that some students omitted almost entirely to present any data, but still scored maximum marks on drawing conclusions. Another common trend was to present many measurements, but without referring to them in the conclusion. Very often clear tendencies in the data were not commented on (for instance the relationship between solution time and temperature not being linear). Students seemed rather to have been thinking of making measurements as an aim in itself, and maybe as their way of making the investigation more scientific. For these, it is not meaningful to talk about their skills of designing and conducting an investigation. They rather show rote learning of certain procedures without the expected and necessary understanding of why these should be included.

It is quite a serious matter if the established tradition for performance assessment, as we find it in TIMSS PA, is not able to pick up this problem. The tasks and the test format seem to invite students to go along with their alternative conceptions of science enquiry and do not reveal any mismatch if this disagrees with a scientific rationale. Nor is this mismatch brought to light when their responses are coded and analysed. Another serious matter is that some teaching of scientific enquiry seems to reinforce this situation. Many teachers teach scientific investigations through examples and tell the students *what* to do, but not *why*. A tradition seems to have been established in which teaching and assessment support each other in this respect. In the teaching students pick up certain procedures but do not get to understand the main ideas about scientific investigations. Next, assessment is carried out by having students repeat the same procedures in recipe-like tasks. One likely reason for this happening is the clear pattern established in tasks used for teaching and assessment. The task *Solution*, for example, demonstrates this. First, the task presents a rather simple problem of finding the effect of one variable on another. Next, this relationship (which many students probably already know) may be found quite straightforwardly by making a series of measurements. Tasks like this are common and give a 'model' for what is scientific investigation. The more or less hidden message is that many measurements are important for scientific problem solving and earn high marks in assessment. By solving several such tasks students see a pattern and learn a rule about what is 'scientific'. On performance assessment tasks they can repeat this pattern.

England on top of what?

The title of this paper raises question about how we should interpret the situation for practical enquiry work in English school science based on the data from TIMSS PA. It is beyond doubt that England has been a pioneer in developing school science practical work in general and investigative work in particular. The effort that has been placed on describing science enquiry in the curriculum guidelines and the effect this has had on classroom practice is unique internationally. Integrating assessment has been an important and maybe necessary factor in this development. When England scores high on the performance assessment test we see the fruits of this comprehensive effort.

There are, however, reasons to ask questions about this development if science investigation is reduced to a set of exemplary tasks copied by teachers, done by rote by students, and then assessed with some standardised tasks. It is difficult to tell how serious and extensive this problem really is. With TIMSS PA we may think the test format itself is part of the tradition and also the problem. Taking into consideration the effort that has been put into including investigative work, however, it is remarkable that students in England are still found to have a rather poor understanding of the nature of science. Several studies have pointed out students' lack of understanding of scientific experiments (Driver *et al.*, 1996; Solomon, Duveen, and Scott, 1994). Driver *et al.* (1996) categorise students' understanding of an investigation into three levels, understanding the relationship between theory and experiment being the most sophisticated. They found that only a minority of the students had developed this understanding, even at age 16. Most students regarded investigations in a simple 'relating cause and effect' frame, which is likely to have been developed through using standard tasks involving controlling variables. A similar tendency was found when students were asked to give warrants for scientific statements. Less than half of the students at age 16 could base their warrants on evidence in a satisfying scientific way. Studies like this give reason to think that students carry out their investigative tasks in science with a very limited perspective, without a wider understanding of the role of evidence.

Looking at the overall results for TIMSS PA, we still find England to be in a special position for enquiry science. The study tended to reveal very different traditions among countries when it comes to the place and role of practical enquiry in school science. In some countries students are obviously used to working with science in a practical setting, but do not necessarily carry out investigations. In other countries it is clear that the students could not handle the practical setting of using equipment and making measurements. The weakest aspect in all countries seemed to be students' ability to explain their conclusions by using subject matter knowledge. Students very often repeated observations rather than relating them to scientific models. Again this may be related to students' understanding (or lack of understanding) of science enquiry, and more specifically of the role of 'theory'. Several such aspects of the results remind us how complicated and difficult it is to foster enquiry in school science. In spite of all the effort and discussion, we have not found a way of handling the issue. This of course affects assessment and makes it difficult in the end to tell what scale or criteria really should be used.

A way forward

The main point made in this article is that much practical investigative work in school science seems to have developed around specific types of task, followed by a theoretical framework with a focus on scientific process skills. This has included assessment, as seen in TIMSS PA, that supports this teaching approach. Warnings against this approach have been apparent for a long time, and we have started to see a change.

Since 1995, when TIMSS PA was conducted, England has adopted a revised version of the science National Curriculum (QCA, 1999). This uses many of the ideas that have been developed in the period following the criticism of a process-skill based approach to science enquiry. Two crucial points are dominant. First there is an emphasis on what the students should be taught *about* scientific enquiry. Here we see a clear attempt to establish an understanding of the role of evidence:

> *Students should be taught about the interplay between empirical questions, evidence and scientific explanations.* (QCA, 1999, p. 28)

For this teaching, historical and modern examples are used but with a clear focus on some general ideas.

The second point is that the investigative skills to be taught and learned are presented in the context of understanding these ideas. Investigative skills, therefore, are not presented solely as something you do, but as something you understand in relation to the purpose of science inquiry:

> *Pupils should be taught to make sufficient relevant observations and measurements to reduce error and obtain reliable evidence.*
> (QCA, 1999, p. 28)

It would be interesting to know what effects this guideline has had on teaching in England. Maybe students today do investigations in a more meaningful way than was indicated by the TIMSS PA results, and learn what it really means to be 'scientific'. In order to find this out new assessment tools are needed. The QCA has highlighted Sc1 as an area to be strengthened for the national assessment of science from the year 2003, with an increase in the number of questions assessing scientific enquiry (www.qca.org.uk/). This change is based on a two-year consultation and trial period. We therefore look forward to seeing what emerges in England about 'ideas and evidence' for science enquiry.

References

APU (1984) *Science report for teachers: 2 The assessment framework age 13 & 15*. Department of Education and Science, Welsh Office and Department of Education for Northern Ireland.

APU (1989) *National assessment: the APU science approach*. London: HMSO.

DeBoer, G. E. (1991) *A history of ideas in science education*. New York: Teachers College Press.

Driver, R., Leach, J., Millar, R. and Scott, P. (1996) *Young people's images of science*. Buckingham: Open University Press.

Driver, R., Newton, P. and Osborne, J. (2000) Establishing the norms of scientific argumentation in classrooms. *Science Education*, **84**(3), 287–312.

Gagné, R. M. (1965) *The psychological basis for science – a process approach*. AAAS miscellaneous publication. Washington: AAAS.

Harmon, M., Smith T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. N. S., Gonzalez, E. J. and Orpwood, G. (1997) *Performance assessment in IEA's Third International Mathematics and Science Study*. Boston: TIMSS International Study Center.

Hodson, D. (1993) Re-thinking old ways: towards a more critical approach to practical work in school science. *Studies in Science Education*, **22**, 85–142.

Kind, P. M. (1996) Exploring performance assessment in science. Doctoral dissertation. Faculty of Mathematics and Science, University of Oslo.

Klopfer, L. E. (1971) Evaluation of learning in science. In *Handbook on formative and summative evaluation of student learning*, ed. Bloom, B. S., Hastings, J. T. and Madaus G. F. pp. 559–642. New York: McGraw Hill.

Millar, R. and Driver, R. (1987) Beyond processes. In *Studies in Science Education*, 14, 33–60.

Millar, R. (1995) Knowledge and action: students' understanding of the procedures of science enquiry. Paper prepared for the European Conference on Research in Science Education, University of Leeds, April 1995.

QCA (1999) *The National Curriculum for England: Science*. London: Department for Education and Employment.

Robitaille, D. F., McKnight, C., Schmidt, W. H., Britton, E., Raizen, S. A. and Nicole, C. (1993) *Curriculum frameworks for mathematics and science. TIMSS Monograph No.1.* Vancouver: Pacific Educational Press.

Screen, P. (1986) *Warwick Process Science*. Southampton: Ashford Press Publishing.

Solomon, J., Duveen, J. and Scott, L. (1994) Pupils' images of scientific epistemology. *International Journal of Science Education*, **16**(3), 361–373

Tamir, P., Doran, R. L. and Chye, Y. O. (1992) Practical skills testing in science. *Studies in Educational Evaluation*, **19**(3), 263–276.

Woolnough, B. E. (1989) Towards a more holistic view of processes in science education. In *Skills and processes in science education*, ed. Wellington, J. London: Routledge.

Per Morten Kind is Associate Professor at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. E-mail: per.kind@phys.ntnu.no

It must be true – its in the papers!

The following advertising blurb was delivered in the same plastic envelope as the *New Scientist* magazine.

Hand-charge your mobile phone!

If your phone runs out of battery power away from home, now you can generate more power without batteries or electricity. This emergency charger needs no power – just attach the lead to your phone and wind the handle – and it's much cheaper than anything similar on the market.