International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Assessment of Practical Work in Ireland: A critique

P. S. C. Matthews & P. J. McKenna

To cite this article: P. S. C. Matthews & P. J. McKenna (2005) Assessment of Practical Work in Ireland: A critique, International Journal of Science Education, 27:10, 1211-1224, DOI: 10.1080/09500690500102573

To link to this article: http://dx.doi.org/10.1080/09500690500102573

International Journal of Science Education
Vol. 27, No. 10, 19 August 2005, pp. 1211–1224

RESEARCH REPORT

Assessment of Practical Work in Ireland: A critique

P. S. C. Matthews and P. J. McKenna
Trinity College, Dublin, Ireland

This paper critically examines the model of practical assessment discussed by Bennett and Kennedy (2001), and considers it in the light of recent changes in the assessment of science courses in the Republic of Ireland. The model is discussed in detail and the empirical results are re-evaluated. The discussion has wider relevance for the methodology, and reporting, of practical assessment in general.

Introduction

Bennett and Kennedy (2001) point out there has been considerable debate in Ireland about introducing assessment of student practical work as part of the Leaving Certificate examinations in science subjects.[1] A Government committee that reviewed the state of science and technology education in Ireland recommended that practical assessment should count for as much as 40% of the total marks on the examinations (Oireachtas, 2000). Two documents that bear the identical title of *Feasibility Study on Practical Assessment for Leaving Certificate Physics and Leaving Certificate Chemistry* (Kennedy, 1998a; Report, 1997) describe aspects of a study carried out in 1997, the major aim of which was to assess the practical skills of a sample of physics and chemistry students. Kennedy (1998a) provides the raw data collected, but is mainly concerned with the results of the chemistry students. The document that we have called Report (1997) bears no author or publisher, but has the title 'Report prepared by the Steering Committee for the Department of Education and the National Council for Curriculum and Assessment'. This provides an overview of the assessment procedures and lists recommendations (but no analysis of the data). The feasibility study was the first major attempt in Ireland to assess students' practical skills at Leaving Certificate level, and could be of major significance in shaping the mode of practical assessment in Irish science education.

*Corresponding author. School of Education, 3087 Arts Building, Trinity College, Dublin 2, Ireland. Email: pmtthews@tcd.ie

ISSN 0950-0693 (print)/ISSN 1464-5289 (online)/05/0101211–14

DOI: 10.1080/09500690500102573

We argue that the model described in Bennett and Kennedy (2001), Report (1997), and Kennedy (1998a, 1998b), does not represent a sound procedure for assessing students' practical abilities. In discussing the model, a number of points will be raised that have general relevance for the treatment of empirical data related to practical assessment. (For convenience we shall refer to Bennett and Kennedy (2001) as 'B&K' and Report (1997) as just 'the Report'.)

Design of the Assessment Procedures

Neither B&K nor Kennedy (1998a) provide an account of the precise nature of the tasks performed by the students taking part in the study, the demands made of examiners or of the marking scheme that was used to evaluate the students' work. The Report does present a summary of such details, but no analysis of the results upon which its conclusions are based. We suggest that the procedures employed have a great bearing on the interpretation that should be placed on the results of the study. We concentrate on the part of the study that dealt with physics. Initially we present the major features of the assessment procedures with little comment. Then we discuss these features in the context of the results reported by B&K and the wider literature. Finally, we re-evaluate the data presented in Kennedy (1998a).

The Scale of the Study in Physics

Thirty schools took part in the physics component of the study, with 473 students being examined. However, only data for the 337 students taking the Higher Level Physics examination is presented in Kennedy (1998a).[2] An examiner visited each school and assessed students according to a scheme that had been initially drawn up by the study Steering Committee. Modifications were made to the original scheme following meetings between the study team and the teachers that had been selected to act as examiners. The practical assessment was split into three 'phases' (see Table 1). The examiner had to examine the work of students on an individual basis, with 15 minutes being allowed per student.

The Nature of the Assessments made in the Three Phases

Phase 1

For the purposes of the Leaving Certificate course, every student is expected to keep a practical notebook in which he/she writes up an account of a series of

Table 1. Outline of the key features of the study

Phase	Assessment to be made	Time allocated	Maximum marks
1	Examination of practical notebook	5 minutes	21
2	Understanding of experimental procedure	5 minutes	18
3	Assessing generic practical skills	5 minutes	21

practicals listed in the syllabus document. The examiner was required to inspect every student's practical book to check that the reports contained an introduction, a procedure and a set of results. A single mark was to be awarded for every practical with the three features, up to a maximum of 21. Also, if the examiner found a report that did not have the three characteristics, he/she was instructed to ignore that practical and continue reviewing the reports until 21 marks had been awarded or no reports remained. Thus a student would only be given a mark less than 21 if the notebook contained less than 21 practicals with the three characteristics. Allowing an examiner time for carrying out merely mechanical processes such as opening books, turning pages and recording marks, it is probable that, at most, 10 seconds was available for the inspection of an individual practical report. It should be noted that the quality, including accuracy and analysis of results, was *not* assessed.

Phase 2

The stated aim of this phase was to examine students' 'ability to explain the practical work which they have performed' (Report 1997: 74). The method chosen was for the examiner to select two practicals from the student's practical notebook and ask the student questions about the use of apparatus and procedures in making measurements in the practicals, and about sources of error. Examples of questions that were suggested as suitable to ask students were (Report 1997: 74–75):

- What were you measuring and why?
- How was the apparatus set up?
- Why is the smallest measurement approximately 30 cm? (Set in relation to an optics experiment.)

In addition, examiners were instructed as follows:

> Where it is apparent that the candidate does not have a familiarity with the experiment being examined, move on to another one. Repeat this procedure until you are able to secure some response from the candidate. Account of this should be taken in marking the candidate. (Report, 1997, p. 74)

Marks were to be awarded according to the scheme presented in Table 2.

Subdivisions of the 6, 12, or 18 marks were not permitted. However, if a student refused to answer questions on a given practical, then 3 marks were subtracted from the total. Thus 3 was the lowest possible mark.

Table 2. Mark scheme for phase 2

Student readily able to answer all, or nearly all questions	18 marks
Student able to answer most questions	12 marks
Student able to answer some questions, prompting required	6 marks

Phase 3

This phase was given the heading ‘Assessing Generic Practical Skills’ and was concerned with assessing “candidates’ practical skills by requiring them to manipulate apparatus and make measurements”. Students were required to perform two different tasks ‘typical of the practical skills’ on the syllabus. An outline of the two types of task and marking scheme are shown in Tables 3 and 4. For task 1 the student was presented with the apparatus already set up, and for task 2 the student was given the minimum set of apparatus/equipment necessary to complete the task successfully. The decision of which pair of tasks to set the student was made by the examiner. No guidance was given about criteria for making the selection.

Commentary on the Assessments

Phase 1

There is no valid link between the superficial appearance of the write-up of a practical and a student’s ability to perform the practical. For example, the student may have copied the results from a partner; the teacher may have given the students instruction on how to write-up the experiment (and verified that this was done); the results may have been very inaccurate; the conclusions drawn may have been in error; the method and results may have been copied from a textbook. Thus the validity of this phase of the assessment is highly questionable (although the reliability may have been very high). The marks awarded may be a measure of a student’s diligence in writing up practicals, and be an indication of motivation; but even this is far from certain (some students will have competed the write-ups as a duty — in response to a teacher’s insistence).

Table 3. Pre-defined practicals for phase 3, tasks 1 and 2 (summarised from Report, 1997, p. 75)

Task 1
To measure the mass and temperature of water in a beaker
To measure the length and period of oscillation of a simple pendulum
To locate an image and measure the object and image distances
To measure the resistance of an electrical component. (Note: this was done using an Ohmmeter.)
To obtain resonance for a given tuning fork and measure the appropriate length
To measure the diameter of a wire. (Note: this was done using a micrometer.)
Task 2
To measure the velocity of a moving object
To assemble a circuit, given a circuit diagram, and measure the current, *I*, and potential difference, *V*, for a resistor
To measure the angles of incidence and refraction for light passing through a glass block
To set up a circuit using the given components, so that either switch lights a bulb or both switches light a bulb. (Outline circuit diagram provided.)

Table 4. Mark scheme for phase 3 (examples from Report 1997: 75, 81–82)

Task 1	
To test candidates' ability to use equipment to make measurements	9 marks
Example of how marks were apportioned:	
Experiment 1: to measure the mass and temperature of water in a beaker	
Temperature of water	3 marks
Mass of water and beaker	3 marks
Measure mass of beaker, calculate mass of water	3 marks
Task 2	
To test candidates' ability to set up apparatus to carry out a particular task (e.g. to make measurements or to perform a particular function)	12 marks
Example of how marks were apportioned:	
Experiment 7: to measure the velocity of a moving object	
Apparatus: partly set up	3 marks
Apparatus fully set up	3 marks
Measurement of appropriate time	3 marks
Measurement of appropriate length	3 marks

Why there was a decision to reward form over substance in this phase of the assessment is not clear, but we suggest that the Feasibility Study confused two issues: *rewarding* students for work done, and *assessing* the quality of their work. A sound method of assessment should serve both functions, but the two aspects are different. Students whose work is properly assessed should receive an appropriate reward for demonstrating a specified range of knowledge, skill, and so on. However, awarding students marks against ill-defined criteria is not assessment.[3] Similarly, it is not clear what aspect of a student's knowledge, skill and so on, the award of the minimum mark (3) represented.

Phase 2

The key issue at this stage is to ask whether the assessment procedure was a reliable and valid measure of 'understanding' of experimental procedure. Notice that the procedure allowed different students to display their understanding (or otherwise) of different practicals. We believe that different practicals differ in most, and probably the majority of, aspects. For example, the demands made upon a student to set up and complete an experiment to measure the acceleration due to gravity using a pendulum are not equivalent to those in using an air track to confirm/verify conservation of momentum. Students may find one more (or less) easy to recall, and to explain, than the other; and they learn different things from each one. Indeed, this is merely to emphasize the context dependence of learning, the importance of which has become increasingly recognized. For a general discussion, see Hennessy (1993), and in relation to practical work, Lock (1989, 1990) gives a considered account. On balance it remains the case that context independence is only achieved on some very general, personal, skills; for example, self-reliance (Lock, 1990, p. 44).

It seems to have been an underlying assumption of the Feasibility Study that it did not matter about which practicals students were questioned. We doubt that this assumption is valid; indeed, Wood and Ferguson (1975) claimed that seven or eight assessments on different experiments is desirable. However, without evidence in the specific context of the phase 2 assessments it impossible to be sure either way. It is an unfortunate aspect of the study that the relevant data were not collected (or if they were, that they have not been published).

The way marks were awarded in this phase is unusual in a number of respects. First, the great majority of students would not fail to score less than 6. Indeed, examiners were expressly forbidden to award a zero mark (Report, 1997, p. 74). Table 5 presents the distribution of marks for this phase.[4]

It may not be surprising that so many students should clear the first hurdle of answering ‘some questions’, especially ‘with prompting’; neither might it be thought remarkable that almost 60% of the students could ‘answer all questions, with little or no prompting’. The reason lies in the vagueness of the criteria: it is impossible to ensure reliability of examiners’ judgements about what constitutes the quantities ‘some questions’, or ‘little or no prompting’. In addition there is the possibility that given the wide gap in scoring 6, 12 or 18, examiners would tend to award the higher mark if they were in doubt. It is apparent from Table 5 that the results were strongly weighted towards the high end of the scale.

B&K report that a system of moderation was used, with a moderator visiting the examiners with the aim of developing consistency of the application of the assessment criteria. This is, of course, sound practice; but moderation of the application of unreliable assessment criteria cannot produce reliable results. For example, a moderator cannot ensure that examiners reliably apply the criteria of a student answering ‘most’ questions if ‘most’ is itself unquantified. (The interpretation of ‘most’ will, in any case, vary with the number of questions the examiner asks — a quantity that was not controlled in the assessment procedures.)

In brief, there are two reasons to doubt the reliability and validity of the assessments made in phase 2:

(i) the restriction in the range of marks that could be awarded leading to a potential bias on the part of the examiners to award marks in the higher ranges; and

Table 5. The distribution of marks awarded in phase 2

Mark	Count	Percentage
0	2	0.6
6	27	8.0
12	106	31.5
15	6	1.8
18	196	58.2
	$n = 337$	

(ii) imprecision in the instructions given to the examiners left too much to individual interpretation, both in assessing which of the three marks to award and which practicals they asked students about.

It is also relevant to note that the mark scheme imposed three discrete points on what should be a continuous scale. In effect, the range of student abilities were constrained to fit just three broad categories, and thereby be unlikely to provide a meaningful description of the full range of student abilities.

In passing it is worth comparing the type of question asked in this phase with some of the questions asked on the practical questions on the 1997 Leaving Certificate Physics examination. For example, Q5 was about pendulum motion, and parts 2, 3 and 4 were:

> How might the student have ensured that the point about which the pendulum was swinging remained fixed?
>
> Explain how the number of oscillations affects the accuracy of this experiment.
>
> While counting the number of oscillations for each value of the length the student noticed that the amplitude decreased slightly. Did this affect the accuracy of the final result? Explain.

It is apparent that the thrust of the questions asked on the examination paper and those used in phase 2 were similar. We shall return to this point later.

Phase 3

There are two issues upon which we wish to focus:

(i) to what extent were the practicals 1–6 (Table 3) equivalent to each other, and to what extent were the practicals 7–10 (Table 3) equivalent to each other; and
(ii) did the assessment procedure measure 'generic' practical skills?

We argue that the practicals in each group were far from equivalent: they used different pieces of equipment; they involved different types of physical manipulations; and they required different levels of knowledge and understanding to interpret the methodology used and to analyse the meaning of the results. Just one or two comparisons serve to illustrate the points at issue. Consider measuring the temperature of water, determining the location of an optical image and measuring the diameter of a wire using a micrometer. Is it the case that a student who performs well on one of these tasks will *necessarily* perform well on the other? In fact there is no such connection — the tasks require specific knowledge of the individual pieces of apparatus. For example, it is easier to measure temperature using a linear scale on a thermometer than it is to use the Vernier scale on a micrometer; and these are different from moving a lens or mirror to adjust the position of an image. This being the case, there is the strong likelihood that the marks awarded in this phase were highly dependent on the *examiner's* choice of which practical the student was asked to do. It would have been most useful if one (or more) of B&K, the Report, or Kennedy (1998a) had presented data that would allow this hypothesis to be tested. We again refer to the work of Lock (1989, 1990) and, for example, to Schoster and von Aufschnaiter (1999) in relation to context dependency of student performance.

The results of this phase are, in any event, of dubious reliability because no criteria were set about the degree of accuracy that students had to achieve (or, if such criteria were set, there is no statement of them in the documentation). For example, it is possible that one examiner considered a temperature measured to ±1°C to be acceptable, and another chose ±0.5°C. A similar criticism can be made about all the measurements made in phase 3.

B&K do not explain the meaning they give to the term 'generic', but we assume it is used to imply that there are practical skills that can be applied across a wide range (perhaps all) practicals, and that exist independently of the context in which they are applied. Perhaps such skills (assuming they exist) might better be described as process skills such as observing, measuring, recording, interpreting data, and so on. However, we regard it as highly unlikely that a reliable measure of process skills can be made on just one occasion, and within the space of less than 5 minutes. It is not uncommon for teachers in the United Kingdom, and elsewhere, to assess students and award marks for process skills; but such marks are awarded as part of a system of coursework assessment. In this way, a student's ability to measure is judged on a variety of different tasks. There remains the problem of what an average mark for (say) measurement using different instruments means; but the system is more defensible in that the student is observed in a variety of different contexts and, to some extent, undue reliance is not put on any one assessment occasion. Brown et al. (1996, p. 380) succinctly state one of the points at issue:

> In order to determine generalisability of processes it is necessary to test the same students on the same tasks, all of which incorporate the same range of processes.

This matter has also been discussed more recently by Gott and Duggan (2002).

Before turning to further analysis of the results cited in Kennedy (1998a), it is pertinent to note the distribution of marks in phase 3 (presented in Table 6). This indicates that here too the examiners have shown a strong tendency to award marks at the top end of the range. Nearly 40% of students were awarded the maximum mark of 21, 23.5% were awarded 18 (this was exactly the median mark), and 20.5% were awarded 15. The skewed nature of the distribution of marks for all three phases has a bearing on the inferences that should be drawn from the results of the study.

Statistical Methodology

B&K report the data shown in the second column of Table 7 for the physics assessment. The third column contains data that we have calculated from the table of results in Kennedy (1998a). We shall consider the results in three sections.

Table 6. The distribution of marks awarded in phase 3

Mark	Count	Percent
0	3	0.9
3	3	0.9
6	6	1.8
9	16	4.7
12	26	7.7
15	69	20.5
18	79	23.4
21	135	40.1
	n = 337	

Overall Mark Distribution

The mean on the practical assessment is reported in B&K as 80% — a high result when compared with the mean of the results on the Leaving Certificate examination (59.25%). However, the mean of the set of results is not a very informative measure without some other indication of the distribution of marks. Figure 1 shows the distribution of the total marks for the practical assessment. It is negatively skewed, and the median mark equates to 83.3%. Not surprisingly, the distribution fails a Kolmogorov–Smirnov test for normality ($p < 0.001$). Among other things, the existence of such a strongly skewed distribution can be an indication that the measuring instrument is biased. Without strong reasons to the contrary, one would expect that 'practical ability' would show a reasonably normal distribution among a population of over 300 physics students, even if the distribution were somewhat skewed. Given the form of the distribution, one should investigate whether the measuring instrument(s) was faithfully measuring the quantity (or quantities) it claims to identify, or whether the way the marks were awarded does not permit the underlying distribution of abilities to make themselves manifest. We return to this matter later.

Table 7. Data for correlations involving physics marks

		Values of r^2
Average mark on practical assessment	80%	
Correlation coefficient of average percentage mark for practical questions versus mark in Feasibility Study	0.251	0.063 (6.3%)
Correlation coefficient for overall percentage mark on written paper versus mark on practical assessment in Feasibility Study	0.363	0.132 (13.2%)
Correlation coefficient of average percentage mark in practical questions versus overall mark on Leaving Certificate examination	0.782	0.611 (61.1%) (for corrected value, see text)

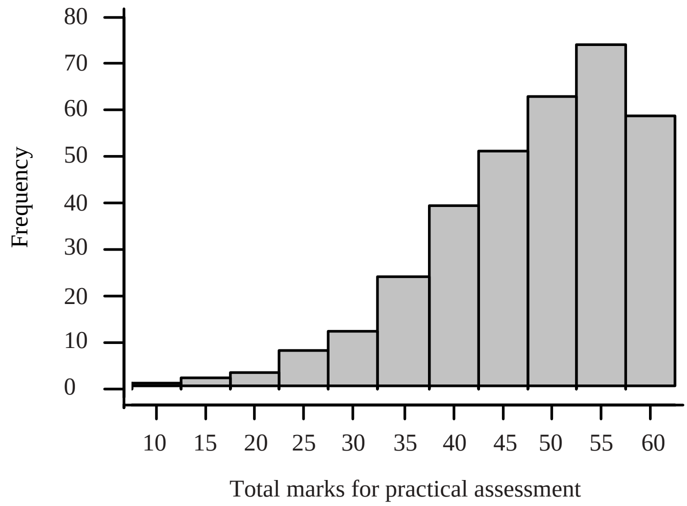

Figure 1. Distribution of total marks for the practical assessment

Correlation

A regression line and correlation coefficient, *r*, can be computed for any two sets of numerical data (scores). The mathematical procedures assume nothing about the origin of the data, their meaning (or lack thereof), or the distribution of the data. The degree to which the variation in one set of data can be associated with the second set is measured by the coefficient of determination, r^2, and not *r* itself. For example, a value $r = 0.6$ would often be regarded as a high degree of correlation between the datasets; but only $(0.6)^2$, i.e. 0.36, or 36% of the variation in one set should be associated with a variation in the other set.

With this point in mind, consider the results for physics in Table 7. This table should be read in conjunction with Table 8, which contains correlation coefficients we have calculated between other data tabulated in Kennedy (1998a) but not analysed in B&K.

The following points are especially relevant to the interpretation of the data:

(i) The cross-correlation of marks awarded in phase 1, phase 2 and phase 3 show low values. Values of r^2 between marks on the three phases never rise above the equivalent of 17%. The values suggest, for example, that: (a) phase 1 has little or no relation to the other two phases; (b) phase 2 bears a little more relation to phase 3 than to phase 1, but even so less than 20% of the variation of one is associated with variation in the other; and (c) phase 2 bears little relationship (ca. 6%) to the marks for the 'practical' questions on the examination paper, even though both are concerned with similar area of student understanding, and neither deal with the explicit performance of practical tasks.

(ii) The value that B&K report for the correlation of marks on the 'practical' questions with the total score on the Leaving Certificate written paper is misleading

Table 8. Correlation coefficients, *r*, not reported in B&K[a]

	Phase 1	Phase 2	Phase 3	Total of phases	Average of examination practical questions	Examination theory questions
Phase 2	0.225					
Phase 3	0.254	0.408				
Total of phases	0.745	0.697	0.733			
Average of examination practical questions	0.113	0.254	0.205	0.251		
Examination theory questions	0.224	0.348	0.237	0.363	0.677	
Total of theory and practical questions	0.213	0.347	0.243	0.359	0.782	0.988

[a]All correlations have $p \leq 0.001$.

because marks for the former were included in the latter. In other words, an element of the correlation that they report is for the 'practical' questions being correlated with themselves. A more meaningful measure is to correlate the 'practical' questions with the residue of the total marks after the 'practical' marks are subtracted. When this is done, we find $r = 0.677$ ($r^2 = 0.458$, or 45.8%).

(iii) The results indicate that the variation of the total mark for the three phases can be related to 13.2% ($r = 0.363$) of the variation of results on the theory only ('non-practical') questions on the written examination, and with 6.3% ($r = 0.251$) of the variation on the 'practical' questions. In other words, the marks of the practical assessment are only marginally related to the marks on the written paper questions (of either category); but of this, the relation with the 'non-practical' marks is at twice the level of the variation with the 'practical' questions. This is an inversion of what one would, and perhaps should, expect.

B&K make the following comment about the low value of *r* for the correlation of the examination marks with the practical assessments:[5]

> … [this] may be due to the fact that students are being assessed in different areas on the written papers than in the Feasibility Study i.e. students of good academic standard who achieve high marks on the written examination do not necessarily score high marks in the area of practical work as assessed in the Feasibility Study. (Bennett & Kennedy, 2001, p. 106)

This merely restates the result and has no explanatory power. However, it does fit with the discussion of construct validity on practical and written tests by Brown et al. (1992, p. 27). In this paper the authors make the point that, ideally (i.e. for valid tests of the same practical skill construct against tests of theory), there should be low correlations. However, it would be unwise to believe that the converse is true; that is, that because low correlations are found then the constructs being examined are valid. For example, in the mark scheme for practical 7 in Table 4, there is no indication of the accuracy to which distance and time should be measured, and no marks are awarded for the calculation of velocity. Indeed, owing to the nature of the mark schemes, either (i) there is no guarantee that the same mark awarded by different examiners reflects the same level of performance by the students, and/or (ii) there is no necessary link between the level of performance and the students' ability in the domain of physics.

In addition there is the possibility that the marks awarded in the three phases were only marginally, if at all, valid measures of student practical abilities. We suggest that this is highly likely, because the marks represented a combination of bias on the part of examiners and a scheme of assessment that did not specify in detail a range of specific student abilities. We have already indicated that marks for phase 1 are not valid as a measure of practical ability, that examiners tended to award high marks as a result of marking to a three-point scale that crudely partitioned a continuous range of underlying abilities, and that students were assessed on tasks that lacked comparability.

Conclusions

B&K say that the Steering Committee of the Feasibility study:

> … was satisfied that the assessment model adopted in the Feasibility Study provided a reliable and valid means of assessing practical abilities … (Bennett and Kennedy 2001: 107)

The Report does make these claims (Report, 1997, p. iv–v), but it provides no statistically valid evidence to support them. For example, the study states no null or alternative research hypotheses that are subject to statistical tests. Thus it is unclear how the Committee came to the stated conclusion.

We indicated earlier that in 2000 a committee of the Oireachtas (the Irish Parliament) published a report on the state of science education in Ireland. It made the recommendation that:

> The Steering Group's model for practical assessment should be immediately introduced for 15% of marks. Over the next three years … a refined system should be developed which would attract 40% of marks thereafter. (Oireachtas, 2000, p. 25–26)

We suggest that our analysis of the Feasibility Study in relation to physics shows the opinion of the Steering Committee to be unwarranted and, at least in respect of the first piece of advice of the Oireachtas Committee, unwise.

The matter of assessing practical work remains a key issue in Irish science education. That such assessment should be carried out has been recommended in a number of publications in Ireland (for example, Task Force, 2002). Indeed, a recent revision of the Junior Certificate Science Syllabus (NCCA, 2003) does, for the first time, establish procedures for such an assessment. However, before practical assessment is established at Leaving Certificate level, we suggest that at least two points need to be considered in more detail. First, in the context of Irish education there has been little attempt to identify in detail the knowledge and skills associated with practical work that are (i) desirable ones to assess, and (ii) capable of being assessed rigorously. It should be emphasized that in the Irish educational system there has never been assessment of practical work by science teachers at Leaving Certificate level, or any other system of practical examination implemented.

A key problem is that some aspects of practical work can, in principle, be assessed very easily — an example is the ability to use an electronic balance to 'weigh' a beaker of water (cf. Table 4) — but this begs the question of whether such a skill is so trivial as to be worthy of the time (and money) spent on a formalized system of assessment. There is already a body of evidence to indicate that performing practical work is very much a holistic activity; and the more one attempts to break it down into a scheme of discrete skills, the more one is likely to miss the essential nature of the activity — one where many different skills interact with one another, and with the student's knowledge base. Thus, in relation to their investigations of task performance assessment, Erickson and Meyer (2003, p. 861) say:

> … our data suggest that it is not possible to separate out, for example, for the purposes of assessment, unique skills (such as observing, listening and inferring) from the specific content demands of the task context.

However, if a scheme of practical assessment by examination is to be introduced into the Irish education system, we suggest that, following a careful analysis of the intended aims and objectives, a set of research hypotheses are established, and appropriate data collected that allow the hypotheses to be thoroughly tested.

Notes

1. A note on terminology: in Ireland, the Leaving Certificate is the name given to the examinations sat by students at the end of their second-level education (typically at about age 18 years). The Junior Certificate examinations are sat by students at the end of the compulsory period of second-level education (typically at about age 15 years).
2. Students can enter for the Leaving Certificate examinations at two levels: Higher and Ordinary. We restrict our discussion to that part of the project that dealt with Higher Level students.
3. This matter is directly related to the business of establishing construct validity, the details of which would lead us too far astray here. See Brown and Njabili (1989) for useful comments.
4. The analyses presented are based on the tables of data given in Kennedy (1998a). Only results for Higher Level examination students were included. Statistical measures were made using SPSS version 11.
5. Kennedy (1998a) states that the value of r is 'statistically significant' but does not state the level. In fact, with a sample size of 337, even a value of $r < 0.15$ would be significant at the 5% level; but that does not mean that the interpretation of r will have any validity (or use).

References

Bennett, J., & Kennedy, D. (2001). Practical work at the upper high school level: The evaluation of a new model of assessment. *International Journal of Science Education, 23,* 97–110.

Brown, C. R., & Njabili, A. F. (1989). The construct validity of an examination designed to test practical ability in biology. *Research in Science and Technological Education, 7,* 61–71.

Brown, C. R., Pacini, D. J., & Taylor, D. J. (1992). Two different methods of assessing practical skills at an advanced level examination in biology: Demonstration of construct validity or the appraisal of non-events? *Research in Science and Technological Education, 10,* 23–35.

Brown, C. R., Moore, J. L., Silkstone, B. E., & Botton, C. (1996). The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations. *Assessment in Education, 3,* 377–391.

Erickson, G. L., & Meyer, K. (2003). Performance assessment tasks in science: What are they measuring. In B. J. Fraser, & K. G. Tobin (Eds.) *International Handbook of Science Education,* Part 2 (pp. 845–865). Dordrecht: Kluwer Academic Publishers.

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

Hennessy, S. (1993). Situated cognition and cognitive apprenticeship: Implications for classroom learning. *Studies in Science Education, 22,* 85–142.

Kennedy, D. (1998a). *Feasibility study on practical assessment for leaving certificate physics and chemistry*. Cork: University College Cork.

Kennedy, D. (1998b). The feasibility study on the assessment of practical work for leaving certificate physics and leaving certificate. In A. Hyland (Ed.) *Innovations in assessment in Irish education chemistry* (pp. 117–127). Cork: Education Department, University College Cork.

Lock, R. (1989). Assessment of practical skills. Part 1. The relationship between component skills. *Research in Science and Technological Education, 7,* 221–233.

Lock, R. (1990). Assessment of practical skills. Part 2. Context dependency and construct validity. *Research in Science and Technological Education, 8,* 35–52.

NCCA (2003). *National Council for Curriculum and Assessment, Revised Junior Certificate Science Syllabus.* Dublin: National Council for Curriculum and Assessment.

Oireachtas (2000) *Report of the Joint Committee on Education and Science on Science and Technology*. Dublin: Houses of the Oireachtas.

Report (1997) *Feasibility Study on Practical Assessment for Leaving Certificate Physics and Chemistry.* Report prepared by the Steering Committee for the Department of Education and the National Council for Curriculum and Assessment, Dublin.

Schoster, A., & von Aufschniter, S. (1999). The influence of learning environments on cognitive processes. In J. Leach & A. Pauson (Eds.) *Practical work in science education*. (pp. 297–311). Dordrecht, Kluwer Academic Publishers.

Task Force (2002). *Report of the Task Force on the Physical Sciences*. Dublin: Department of Education and Science.

Wood, R., & Ferguson, C. M. (1975). Teacher assessment of practical skills in A-level chemistry. *School Science Review, 56,* 605–608.