British Educational Research Journal, Vol. 30, No. 1
February 2004

Carfax Publishing
Taylor & Francis Group

Teachers' assessment of students' research skills

Karel Stokking*, Marieke van der Schaaf, Jos Jaspers & Gijsbert Erkens
University of Utrecht, The Netherlands

Received 15 October 2002; resubmitted 21 January 2003; conditionally accepted 28 February 2003; accepted 2 April 2003.

Today, teachers are expected to develop complex skills, such as research skills, in their students while implementing new views on learning and teaching and using authentic assessment strategies. About these new assessment strategies there is much debate and teachers are vulnerable in using them. We studied upper secondary education natural and social science teachers' practices using two surveys and two rounds of expert panel judgement on teacher-submitted assessment-related material and information. Our study shows that there are grounds for concern regarding the clarity of teachers' assessment criteria, the consistency between teachers' goals, assignments, and criteria, and the validity and acceptability of teachers' assessment practices. The extent to which it is justifiable to judge teachers' assessment practice by professional quality criteria is discussed, and suggestions are given as to the main quality criteria for formative and summative assessment and as to ways in which teachers could improve their assessment practices.

Introduction

In many countries, education is being directed towards new goals (skills), new views of learning and teaching (constructivist), and new assessment strategies (performance assessment, authentic assessment). Skills are increasingly being assessed using assignments that give students options in topic choice, that stimulate cooperation, and that allow occasional help from the teacher, a situation in which instruction and testing are no longer strictly separated. Such less structured and more authentic (real-life) assignments have no strict time limit, rarely have a single correct answer, and permit the use of a variety of aids and sources. They are supposed to test skills, to be more valid than other assessment means and to be suitable for classroom use (Linn, 1989; Linn *et al.*, 1991; Birenbaum, 1994; Keeves, 1994). However, the potential of this approach is not yet clear (Hambleton & Murphy, 1992; Linn, 1994; Messick, 1994). This is all the more important for examination assignments that count towards qualification (De Groot, 1970; Baker *et al.*, 1995).

This issue cannot be settled without investigating teachers' actual assessment practices and addressing the quality criteria against which those practices should be judged. It is desirable to start research into these questions in a context where the

*Department of Educational Sciences, University of Utrecht, PO Box 80140, 3508 TC Utrecht, The Netherlands. Email: k.m.stokking@fss.uu.nl

ISSN 0141-1926 (print)/ISSN 1469-3518 (online)/04/010093-24

DOI: 10/1080.0141192031000162998 3

teaching and assessment of skills using more authentic assignments is already common practice. Research skills in the natural and social science subjects in upper secondary schools fall into this category and are an obvious choice. However, the meaning of the term 'research' should be clarified. Teachers in a number of subject areas have long been familiar with 'practical work', variously specified in the literature as practicals, fieldwork and lab experiments (Gee & Clackson, 1992; Hodson, 1992; Kirschner, 1992; Duggan & Gott, 1995; Meester & Kirschner, 1995; Thair & Treagust, 1997). Meester and Kirschner (1995) differentiate between 'cookbook' experiments (conducting standard tests in a laboratory environment), practical work (more freely experiencing the activities of a researcher), and experimental seminars (learning from discussions to interpret experimental data correctly). The meaning of the term 'research' as used here is broader and is to be equated with what is denoted by Duggan and Gott (1995) as 'investigation': research in which the approach and results are not obvious (see also Wiggins, 1993).

We report on an empirical study designed to answer the following two questions: (1) how do upper secondary school teachers in the natural and social sciences assess the research skills of their students? and (2) what is the quality of their assessment practices? Our conclusion will be that teachers' choices and activities are not very consistent and that the quality of their assessment practices is not quite satisfactory. The extent to which it is justifiable to judge teachers' assessment practice by professional quality criteria is discussed, and suggestions are given as to the main quality criteria for formative and summative assessment respectively and as to ways in which teachers could improve their assessment practices.

Assessment of students' research skills

Assessment as a teacher's task

During the 1980s, it appeared that many teachers in the USA had difficulty choosing, developing and using assessment instruments and procedures (Stiggins & Bridgeford, 1985). The problems were partly organizational (time, peer cooperation) and partly concerned the correct determination and interpretation of results. Stiggins and Bridgeford concluded that the practice of many teachers carried great risks for the quality of their assessments. Five years later, standards for teacher competence in assessment were issued by the American Federation of Teachers *et al.* (1990). Almost another five years later, just as the educational changes referred to at the start of this article began to take effect, Seyfart *et al.* (1994) signalled that a number of students and parents had problems with 'soft' assessment of 'soft' skills. It appears that if no standardized tests are in force and students are permitted to work cooperatively, students and parents are easily capable of distrusting the fairness and objectivity of the assessment.

While it is not surprising that such distrust has emerged in the American culture of standardized testing, it is by no means limited to that country. In the Netherlands, for example, a behavioural code has been developed for testing, assessment and decision-making in secondary education (Creemers-van Wees *et al.*, 1997).

Doubts about the quality of teachers' assessments are not in themselves strange. Teachers *per se* are not competent assessors, certainly not considering the many ways in which human judgement is subject to unintentional distortion, such as by halo effects, order effects and shifting of norms. In addition, the trend towards 'soft' assessment adds complications of its own: assignments that involve options for topic choice by students can diminish the equality of the assessment between students, while assignments with multiple correct answers can make it difficult to construct a scoring scheme.

The task of avoiding such technical problems is added to the many practical matters that teachers must address. Teachers must make a number of choices. The *criteria* of the assessment will depend on the research steps or the skills that may be emphasized, and on the teacher's goals. Regarding *scoring*, teachers must make choices about how much to take into account any assistance that may have been received by the student and how to keep sight of individual performance when assessing group products; and they must decide how to combine subscores. Teachers must also develop *norms* for the performance level that is to be regarded as satisfactory. Finally, teachers must *register* and *communicate* their approach and the results, and they must be accountable to students, parents, colleagues, the school administration and government inspectors.

The trend towards integrating education and assessment has the potential to improve the quality of assessment through greater transparency and acceptability. Communicating and discussing the assessment criteria in advance with students, assessing students periodically, and introducing forms of peer- and self-assessment by students can all contribute to this. However, such measures can also increase the assessment's vulnerability. It is important that teachers' assessments continue to meet certain quality criteria, a matter that we now consider.

Quality criteria for assessments

Some 20 criteria can be identified from the literature on quality criteria for assessments. These criteria fall into four main categories. Figure 1 provides a summary, principally from De Groot (1970) and Messick (1989, 1995).

The *reliability* of assessments depends on reproducibility and repeatability as conditions for consistency (between assignments and between assessors).

The *validity* covers a number of facets. The demand for compatibility with the teaching material, assignments and assistance (criterion 2a) means that assessment should cover the *knowledge content of the lessons* (Haertel, 1985). The assessment of *skills* should be in tune with the *execution* of the task (2b). If a teacher conceives research as consisting of separate steps, this should be reflected in the assessment and scoring instructions (2c). Criteria 2d and 2e concern the correct interpretation of the results (Messick, 1984, 1994, 1995; Cronbach, 1988; Kane, 1992; Moss, 1992; Wiggins, 1993; Crooks & Kane, 1996). Criteria 2f and 2g suggest that, based on the assessment of a certain assignment, a teacher should be able to predict reasonably well a student's performance in other assignments, on qualifying examinations, in subsequent study, and even in future employment.

RELIABLE AND VALID MEASUREMENT

1 **reliability**:
intra- and interassessor agreement (consistency): if the same teachers some weeks later assesses the products once again, or if a colleague assesses the products also, the scores should not differ too much;

2 **validity**:
the assessment should measure what the teacher wants to measure (and not something else); validity has a number of facets:

a content: sufficient coverage of the construct or domain one wants to measure; in educational settings also: fitting the lesson content, assignments and assistance;

b process: as much as possible based on a model of the task; showing the development of the relevant knowledge and skills;

c scoring: the scoring model and scoring rules should mirror the structure of the construct, domain or task;

d specificity and unambiguous interpretation: the assessment should distinguish between more and less skilful students (see also criterion 4b), and the scores should not be explainable by other factors than the knowledge and skills intended;

e convergence (a special case of 2d): measurements of the same construct, domain or task using different methods should correspond as much as possible;

f prediction: the results should be connected with external criteria;

g generalizability: the degree to which the results are valid for a broader range of tasks, conditions and populations;

ADEQUATE ASSESSMENT

3 **acceptability**:

a objectivity: precise scoring instructions, minimal need for interpretation, procedures and criteria are defensible to others;

b transparency: procedures (taking, scoring), criteria, norms and consequences should be clear and explainable to others (students, parents, colleagues, school director, inspector) and the students should know them beforehand;

c equality: students should be treated equally and should get equal chances (in view of the time available, means, assistance);

d lack of bias: items should not discriminate between personal characteristics which are not relevant;

4 **practical utility**:

a functionality: the results should be suitable for the functions intended, such as:
-give information on the students' progress and weak points;
-be suitable for grading;
-be suitable for communication with students and others (e.g. parents);

b appropriate difficulty level and discriminating power;

c workability: teachers should be able to realize the assessment in their situation (in view of their knowledge and skills, means, available time);

d efficiency.

Figure. 1. Quality criteria for assessments

The *acceptability* primarily relates to aspects such as fairness and lack of ambiguity. It also embraces the need for teachers to avoid dealing with illegitimate criteria (such as their 'overall impression' of a student) (criterion 3a), and to effectively communicate to students the intentions, organization and implications of the assessment (3b) (Millman & Greene, 1989).

The *utility* refers both to the feasibility for the teacher (4c and 4d) and to the clarity of the meaning assigned to the assessments and the assessment results (4a).

It is partly their utility that allows assignments and assessment results to play an effective role in the learning process, since assignments are designed to motivate students, to facilitate the teachers' ability to achieve insight into learning difficulties and to give students meaningful feedback (Crooks, 1988; Snow & Lohman, 1989; Madaus & Kellaghan, 1992; Cowie & Bell, 1999).

It should be pointed out that Messick (1989, 1994, 1995) has put forward a broader meaning of the term 'validity'. This subsumes all quality criteria, not only those relating to measurement, evidence and interpretation, but also those concerning the assessments' use and consequences. We prefer, however, to preserve the traditional, more restricted meaning of validity and to maintain the differentiation between the four main categories of criteria.

The literature on new forms of assessment is full of positive conclusions and expectations (Linn, 1989; Wiggins, 1989; Wolf *et al.*, 1991; Rowe & Hill, 1996). Many authors are optimistic: with these instruments we would be able to measure higher order skills, and do so better than traditional tests possibly could. However, in our view the potential of this approach is not yet clear. We believe that more authentic forms of assessment must be subjected to the same critical review as other claims to validity (Linn *et al.*, 1991; Linn, 1994; Glaser & Silver, 1994; Messick, 1994, 1995; Haertel, 1999; Stokking & Voeten, 2000). First, new forms of assessment emphasize less the correctness of interpretations and more the validity of consequences. However, the former (the traditional validity) remains significant in view of the importance of consistency between judgements and generalizability of tasks, especially in classroom assessment with only one judge (the teacher) and a small number of tasks. Secondly, the design and development of good assessments must be attuned to the goals and content of the lessons and must be suitable for teachers to use. Thirdly, the assessments have to be general enough to accommodate the broad variation in what students choose and produce. Satisfying such diverse requirements demands a great deal of development work in order to create good scoring rules (Arter, 1993; Baker *et al.*, 1995; Novak *et al.*, 1996). In practice, it is often difficult to distinguish between students in a sufficiently reliable and valid manner. Because of this, in many cases the acceptability and the utility of the assessment will be in danger of being compromised.

Considering the fact that teachers carry a heavy assessment task (see before), the quality criteria discussed above, when considered together, are probably overly demanding. However, it does appear to be legitimate to expect teachers, at least, to make their assessment criteria transparent, and to maintain a certain consistency between their goals, assignments, instruction, assistance and assessment criteria, so that their assessments will be valid, acceptable and useful. Linn *et al.* (1991) and Linn (1994) state that the extent to which assessments should meet traditional technical standards depends on the particular purposes of the assessment and the uses of the results. However, they do not systematically develop this idea further. At the end of this article we give some suggestions how to differentiate between formative and summative assessments.

It would be of great assistance if teachers could be supported by knowledge about the development and application of the skills they must assess. The state of the art regarding the development of skills tests even includes the use of an explicit model for this (Messick, 1984, 1995; Willett, 1988). We now turn to this issue.

Is there a theoretical model for the assessment of complex cognitive skills?

Fitts and Posner (1967) proposed what has become a well-known three-phase model for skill development. Anderson (1982) adapted this model to the development of cognitive skills. The first phase addresses the development of declarative knowledge. Knowledge compilation converts this into procedural knowledge (a more or less fixed order of mental tasks). Finally, exercise (and reflection on any problems) leads to the proficiency of an expert. However, this model has a particular bearing on specific knowledge-intensive problem-solving skills but less so on broad, complex skills such as research skills. The reason for this is as follows.

It is often assumed that development takes place in phases or levels that always follow one another in the same order, and that each phase or level can be separately assessed (Reed, 1968; Goldstein, 1979; Fischer, 1980; Frederiksen *et al.*, 1990; Keeves, 1994). However, development can also take place spasmodically (Reed, 1968), and the route along which skills develop is subject to individual differences (Fischer, 1980). Furthermore, the order in which complex skills develop is not necessarily the order in which the component skills are to be applied during the execution of a task (Nitko, 1989), and this order can also differ individually depending on the stage of skill development (Colley & Beech, 1989) and the personal approach (Snow & Lohman, 1989). Performance differences can stem from differences in how the task or problem is represented, in the performer's *a priori* knowledge, in the approach taken, and in the monitoring (Messick, 1984; Millman & Greene, 1989; Alexander *et al.*, 1991; Bereiter & Scardamalia, 1993). Finally, Colley and Beech (1989) have indicated that there are many types of models for skill development, of which the phases or stages model is but one. Development can also be modelled as increasing differentiation, as a cyclical process, and as a gradual attention shift from implementation to preparation.

It should be clear that we do not yet have a conclusive theoretical model for the processes that play a role in the development and application of more complex cognitive skills. Consequently, the assessment of research skills cannot be based on such a model. We now examine whether teachers could at least have something to hold on to in considering examination requirements and the learning goals that they want their students to achieve.

The image of research in examination requirements

In the 1990s the examination requirements in the academic track of Dutch secondary education were revised to lay greater emphasis on general skills, especially research skills. The Netherlands employs a two-part examination system: the 'school examination' and the national 'central examination'. The school examination consists of a dossier with assessments of prepared assignments for each subject.

Teachers may create the assignments themselves or make use of third-party material. Assignments are prepared individually or in groups; the teacher can assist when needed. The assignments are graded, and the grades are included in the school examination. To qualify for the central examination, a pass mark is required on the last, largest assignment.

Nomenclature and classification vary, and we have mapped the examination requirements concerning research skills for physics, biology, history, geography, economics, Dutch and mathematics. The requirements can be represented in 10 consecutive steps:

1. identify and formulate a problem using subject-specific concepts;
2. formulate the research question(s), hypotheses and expectations (if any);
3. make and monitor the research plan: research design and time schedule;
4. gather and select information/data;
5. assess the value and utility of the data;
6. analyze the data;
7. draw conclusions;
8. evaluate the research;
9. develop and substantiate a personal point of view;
10. report (describe) and present (communicate) the research.

The 10-step classification is blind to differences between subject areas. The subject areas vary as to the degree to which problems are well or poorly defined and to the degree to which research is designed to produce a singular (good) result or can have multiple outcomes. Subject areas also vary in their primary research types. Nevertheless, the 10 steps appear to represent a workable arrangement across subject areas.

With our earlier conclusion in mind, that there appears to be no conclusive model for the development of complex cognitive skills, it is interesting to consider whether these 10 steps could represent the basic research skills. However, if research skills are conceived as representing the cognitive capabilities necessary to carry out research, then these steps cannot be considered to be the skills themselves. After all, different steps can presuppose the same cognitive activities, and for any given step, more than one cognitive activity may be necessary (for example, see Lock, 1989, 1990; Brown *et al.*, 1996).

The 10 steps can also be put into perspective in another way. Hodson (1992) argued that science is holistic and can only be taught, learned and assessed as such. The idea of basing assessment on a step model of research is, according to Hodson, not without risks; the steps are not fully differentiated, and they are dependent on both the subject matter and the context in which the research is carried out. Kirschner (1992) also warned against limiting research to practical execution, and advocated reflecting on the interdependency of the relevant conceptual and procedural knowledge and on the overall quality of the research. Similarly, Gott and Duggan (1996) focused on research in terms of concepts of evidence, and White and Frederikson (1998) focused on overall criteria such as systematic work, meticulous reasoning and clear communication.

The question whether the 10 steps accurately represent research skills can be addressed in an even more fundamental way. The examination requirements define research in terms not of skills but of concrete, sequentially ordered activities. Conceiving the 10 steps as tasks to be carried out (rather than as skills) fits well into a constructivist approach. In such an approach, tasks must be meaningful and realistic, and precisely these criteria can be met with less structured assignments: students can make their own choices and work cooperatively, and the result is not completely prescribed. In a constructivist approach, the cognitive-psychometric view of skill development is abandoned, and little attention is paid to theoretical modelling of the process.

According to Van Tilburg and Verloop (2000) and Van Rens and Dekkers (2000), working in terms of research steps is accepted by the teachers, but the same teachers have little knowledge of research, little experience of conducting research themselves and little experience in constructing and assessing research assignments for students. What are the goals, then, that teachers themselves want their students to achieve?

Teachers' goals for giving their students research assignments

Students can be engaged in research for various purposes, and an evaluation of teachers' assessments of students' research skills should take this into account. Seven different goals can be identified from the literature (Hodson, 1992; Kirschner, 1992; Duggan & Gott, 1995; Meester & Kirschner, 1995; Thair & Treagust, 1997):

1. develop their knowledge about the concepts and content of the subject;
2. motivate them to work actively on the subject and facilitate exercise in independent learning;
3. develop their knowledge about the essentials of research, the language and argumentation and the research cycle: the commutation between theory, research question, design, data and conclusions;
4. let them gain experience with the research steps or component skills;
5. introduce them to the tool box of the researcher;
6. introduce them to the 'ethos' of doing research;
7. develop the insight that knowledge is developed by people and is continuously being developed further.

The Dutch examination requirements primarily fit Goals 3 and 4. The more structured variants of practical work primarily match Goal 5. The experimental seminars of Meester and Kirschner (1995) combine Goals 1, 3 and 5. Duggan and Gott's (1995) difference between conceptual and procedural understanding lies between Goals 1 and 3.

A teacher's emphasis will be partly dependent on his or her conception of the discipline (that is, the importance of research within the field, the extent to which disciplinary knowledge is still developing), on his or her own affinity for research, and on his or her expectations of what students are able to learn. For example, the following questions could be asked in relation to the last aspect: is it expected that students develop subject knowledge by research (Goal 1), is it feasible that they

Table 1. Numbers of responding teachers and of sets of materials judged

	Physics	Biology	History	Economy	Geography	Total
Response survey 1998	10	15	14	10		49
Response survey 1999	37	41	23	28	36	165
Sets judged, first round	13	5	5	3		26
Sets judged, second round	6	6	6	6		24

achieve insight into the nature of research (Goal 3), or is a limitation to component skills (Goal 4) the most realistic option, or even a limitation to conducting research as an exercise in independent learning (Goal 2)?

Method

Our study consisted of a preparatory round of interviews followed by two questionnaire surveys of teachers, and two rounds with judgements by a panel of teachers' assessment practices on the basis of information and materials submitted. Table 1 contains an overview of the exact numbers of participating teachers and materials.

Surveys

Sampling and responses. At 122 schools for upper secondary education (a quarter of the national number of these schools), we approached the department heads of physics, biology, history and economics. We indicated that we sought teachers with students in the last three years before their final examinations and with some experience in assessing research assignments. We gave the teachers the choice of taking an interview or a questionnaire and we requested information about their own experience and their arrangements with colleagues. The response from history teachers was poor, and to make up the shortfall we subsequently approached the heads of the history departments at a random sample of 200 other schools.

In total, 119 teachers from 61 schools agreed to participate. These teachers had considerable experience in the assessment of research assignments, but had not often made arrangements with colleagues. In May 1998, 20 interviews were conducted at 11 schools. These interviews were used to develop the questionnaire. The remaining 99 teachers received the questionnaire in October 1998; 49 teachers with an average of 20 years of teaching experience responded from 40 schools. (The others had agreed to participate but did not respond.)

In November 1999, a second survey was sent to a random sample of 300 schools. The department heads of physics, biology, history, economics and (for additional comparison) geography were asked to complete a questionnaire. In total 165 teachers from 123 schools responded, again with an average of 20 years of experience.

Instruments. The two questionnaires (1998 and 1999) overlapped in content. Some areas were sufficiently addressed by the first survey and were therefore not included in the second one; subsequent study of the literature and discussions with the panel (see below) led to the inclusion of additional questions in the 1999 questionnaire. The two surveys consisted of both single questions and Likert-type scales containing a number of items, and covered four areas: the teacher's views on research and research skills; assignments; instruction and assistance; and assessment.

Analyses. First, scale analyses were carried out (criteria: all item-test correlations at least .20, preferably .35 or higher; Cronbach's α at least .60, preferably .70 or higher). Next, descriptive analyses were executed. Finally, differences between subject areas were analysed using variance-analyses (using a significance level of 5%).

Panel Study

Materials. During the interview round and the first survey, we asked teachers to forward the following material and information. First, from one to three research assignments, using a broad definition of what qualifies as research and as an assignment. Assignments completed by both individual students and groups were acceptable. The sole restriction was that the final product had to be in written form. This is the most common form and was the most suited for this project. Second, for each assignment, we required a completed questionnaire covering the class, contact hours, course duration, course goals, execution and assessment; an example or description of the supplied instructions; an example or description of the assessment criteria; the assessments themselves; the finished assignments of three students or groups of students; and supplementary information on the assessment, such as the scoring and the weighting of components. The interviews and the first survey generated a total of 54 sets of materials from 35 teachers.

The assignments judged in the first round (described below) were spread over the last three school years before the final examination. The assignments were formulated by the teacher in about 70% of cases. According to the teachers, the students worked on the assignment for an average of 9 weeks, spending an average of 7.5 hours in total. The assignments judged in the second round were also spread over the last three school years and similarly were formulated by the teacher in about 70% of cases. According to the teachers, the students worked for an average of 9.5 weeks on the assignments, spending an average of 13 hours in total.

Compared to the results of the surveys (see below), those teachers who submitted materials equally often designed assignments themselves as the other respondents. The most complete sets of materials, which were selected for the second round, came from teachers whose answers indicated that they relatively systematically supervised and assessed assignments and relatively often informed students of the assessment criteria beforehand.

The panel: composition, beliefs and opinions. A panel of six experts was assembled, all active as researcher, teacher trainer, expert in pedagogical content knowledge, test developer and/or teacher, and all familiar with one or more of the four subject areas selected. After the first round, the composition of the panel changed (due to time availability): two members with expertise in test development were replaced by one with the same expertise, and the two teachers were replaced by two pedagogical content experts. The panel now comprised five members whose expertise continued to include the four subject areas.

As the assessment of research skills was still a new and complex area in secondary education, the panel study was not designed to reach a consensus, but deliberately constructed to be heterogeneous in order to generate a diversity of insights and arguments (Mitroff & Turoff, 1975). To control for this, in preparatory discussions we asked the panel about their conceptions of research and research skills, their opinions about the intentions and feasibility of the new examination requirements, and the quality criteria that could and should be used to judge teachers' assessment of students' research. In these conversations, which preceded the judging process, the panel showed that it did indeed represent a range of views and experiences. This was apparent, for example, in respect of the goals for student research that the panel found most important: these included the first five goals (see above). Most panellists considered some of the examination requirements attainable and others not attainable (this varied by subject area). The panel felt comfortable with the 10-step scheme (see above).

The panellists' remarks showed that they differed as to how far teachers can be expected to tailor their assessment process to meet all of the quality criteria (see Figure 1). One panellist stated that an assessment must meet all quality criteria to be acceptable. Other panellists felt that the complete list of quality criteria in its present form is not suitable for teachers. Those panellists who were also teachers themselves pointed out that in the classroom it is difficult to see what a student has learned and mastered. According to these panellists, assessment is routinely also based on other factors such as effort, grading does not necessarily give an accurate picture of knowledge and skills, and grades have other functions as well, such as motivating students.

Procedure. The panel ran through two rounds. In the first round, the panel made a primarily qualitative judgement of the first 26 sets of materials received, which were spread over the four subject areas. Each set was judged by two panellists knowledgeable in the subject area.

Based on the quality criteria (see Figure 1) and the initial conversations with the panel, we created a set of instructions to be used by the panel in the second round. Some examples are: criterion 1, reliability: 'There are two aspects the panellist can pay attention to: the extent to which the procedures and criteria the teacher is using are unambiguous and therefore suitable to copy (that is, another teacher would comprehend and execute these in the same manner), and the extent to which the procedures and criteria are vulnerable to faults and mistakes and therefore are less likely to be repeated by the same teacher'. Criterion 3, acceptability, sub c:

transparency: 'This aspect refers to many of the preceding points; the panel could, by way of operationalization, ask itself to what extent another teacher, a student, or a parent could be convinced of the quality, fairness, etc. of the assessment'.

In the second round, the panel judged in a structured, quantifying manner the most complete 24 sets of materials (of the total of 54), that is, six sets per subject area. About half of this second selection overlapped with the first selection described above, in which assessment materials had been evaluated qualitatively. As in the first round, each set was judged by two panellists familiar with the subject area.

Results

Results of the surveys on teachers' assessment practices

The following summary covers both surveys. For each point, we specify whether the results were obtained in 1998 ($n = 49$) or 1999 ($n = 165$). Percentages are rounded to the nearest 5% for the sake of readability.

Goals, conceptions and accents. Teachers primarily target Goal 1 (subject knowledge), Goal 2 (independent learning) and Goal 4 (research steps) (80–90% to a moderate or strong degree in both 1998 and 1999). They find it important that they pay explicit attention to research as part of their discipline (70% to a strong degree, 25% to a moderate degree) (1999). They do not equally emphasize all 10 research steps (1998), laying the most emphasis on drawing conclusions (95% to a moderate or strong degree) and the least emphasis on evaluation (60%), personal point of view (50%), and oral presentation (25%).

Assignments. Most teachers create their own student assignments (70%) and/or use examples from a textbook (65%). The main requirements applied when creating or selecting assignments (nominated by 90–95% of teachers in both 1998 and 1999) were to provide students freedom of choice, to give assigments that are solvable in varied ways, to match the knowledge level of the students, to be challenging, and to ensure that the assignment can be completed within a certain time. Of secondary importance (nominated by 70–85% of teachers) were the requirements to return testable results, to provide the opportunity for reflection, to conform to examination requirements, and to make use of the materials available at the school. The least important requirements (nominated by 55–65% of teachers) were to provide a realistic context and to cover the subject matter. The assigments teachers give are rather pre-structured (1998 and 1999). A constant 60–70% of teachers moderately or strongly structure the choice of subject matter, material and means, and the location of, and the time to be spent on, the research. In 1999, the same question was posed in greater detail for each research step, focusing on the first and the last assignments that students receive in the final years before examination. An average of 60–70% of teachers are inclined to structure the diverse parts in the first assignment, and about 50% are inclined to do so in the final assignment.

Instruction and assistance. About 70% of teachers first teach students component skills and about half have students first research an easy topic, then tackle topics of increasing difficulty (1998). Most teachers take into account differences between students by giving freedom of choice concerning the topic of the research (80%) and/or providing tailored amounts of assistance (70%) (1998). Teachers assist students most frequently during the first steps, covering problem and question formulation (80%) and planning (65%), and during the final steps, covering evaluation (65%) and reporting (50%) (1999).

Timing of and conditions for assessment. Teachers vary as to when they assess their students' work (1998). A few (5%) do so after each step, 10% after the formulation of a research question, 40% after a plan of action has been formulated, 10% after submission of a concept report, 85% after submission of the final report, and 25% after discussing the final report. Teachers say they take various measures to support assessment quality (1999). Almost all teachers (95%) say they give students sufficient time to carry out the research, and 75% try to ensure that students have equal access to sources of information. About 65% of teachers say that they take steps to ensure that students have sufficient material to do the research, that precise instructions are provided, and that cheating is prevented.

Establishment and communication of assessment criteria. Teachers show considerable variety in the extent to which they document matters concerning assessment. Half the teachers record only the assessment criteria, 30% also record the maximum score per criterion, and 20% additionally document a scoring model (1998). Teachers differ strongly as to how often they communicate the assessment criteria and scoring scheme to their students (1998). About half the teachers do this all the time, and the other half do this some of the time. Teachers clarify the assessment in several ways, mostly (85–95%) by explaining the purpose of the assessment, by explaining the criteria by which the research projects will be judged, and/or by consulting students at each stage of assessment (1999).

Assessment approaches. Teachers can employ various approaches: for example, assessing globally or analytically, assessing a research project as a whole or in terms of separate steps, and either weighting or not weighting subscores. The data from both 1998 and 1999 show that teachers do indeed use many approaches. However, teachers make sparse use of colleagues as coassessors: only 35% do so frequently or regularly; the same is true of the use of peer assessment and self-assessment by students (10% frequently or regularly).

Assessment of process and product; norms (1998). Teachers differ in the components and characteristics they include in their assessment: 60% weight the research question heavily, 45% the research plan, 50% the execution of the research, 30% the results, 60% the conclusions, 50% the structure of the report, 30% the appearance of the report, and 70% the level of the content. Cooperation among students is not assessed by 30% of teachers, and is assessed implicitly by 45% and explicitly by 25%. Most teachers (70–80%) determine assessment criteria, norms and passing grades themselves. About half the teachers base their norms and passing grades on the performance of the class at hand, other classes of the same age level, and/or the previous year's students (norm-oriented), about 40% on the subject matter and examination requirements (criterion-oriented), and about 20% on individual student progress.

Giving feedback. An open-ended question as to how the teacher provides feedback to students (1998) generated diverse answers along a number of dimensions: teachers emphasize to their students what went well or what went poorly, and they give feedback individually or to the total class, in written or oral form, after each step or occasionally, immediately or later, with or without the possibility of revision, and with or without instructions or suggestions for this.

Bottlenecks (1998). According to the teachers, their students have between a fair amount and a great deal of difficulty formulating a problem and a research question (80%), creating a research plan, setting up a time schedule, drawing conclusions and evaluating the research (70%), collecting and processing data (60%), and reporting results (50%). The teachers themselves experience a lack of time (80%) and a shortage of modern equipment at school (70%), and have trouble assessing the individual contributions to group work and including students in the assessment (both 60%).

Differences between subject areas. There are some statistically significant differences between subject areas. Physics and biology teachers value research more highly than economics and geography teachers, more often first teach their students certain component skills (prior to assigning a complete research project), more often permit students to choose the topic of their research, and more often permit them to work together. History teachers take a middle position in this regard, except for the tendency to permit students to work together, which they do the least. Economics teachers, who have the least experience with student research, teach students component skills the least often, set the highest requirements, and are the most inclined to begin with pre-structured assignments, to take measures to ensure the assessment quality and to inform students beforehand about the assessment criteria and scoring. Such measures are undertaken least often by physics teachers. Biology teachers, who have the most experience with student research, report the fewest bottlenecks.

In addition, the individual differences between teachers are great. On most four-point Likert scales the range of the scale scores (the means of the item scores) is 1.5 points or more, and on quite a number of scales the scale scores vary between the extremes of 1.0 and 4.0.

Results of the panel judgement rounds on the quality of teachers' assessment practice

First judgement round. The panel found considerable consistency between the assignment instructions and the assistance given, and between teachers' goals and assistance. In a number of cases, this was difficult to assess because the assistance given was insufficiently documented. The panel found less consistency between goals and assignments, between assignments and assessment criteria, and between assistance and criteria. The least consistency was found between goals and criteria. According to the panel, teachers tended to formulate more goals than could be identified in the assignments, and often decidedly more than could be recognized in the criteria. The panel also found that in a number of cases, the criteria were unclearly formulated.

A supplementary question was posed to the panellists: is the way in which teachers have been assessing adequate in view of the psychometric quality criteria? An important part of the panel's observations dealt with *validity*. In a number of cases, it was not clear to the panel which knowledge and skills were intended and subsequently measured by the teachers. It was also noted that teachers' goals were not clearly reflected in the assessment criteria, and that they failed to measure the intended skills. Conspicuous also were the question marks regularly placed by the panel regarding *acceptability*, owing to its view that the assessment was not objective or not fair and/or failed to provide insight. The assessments' *discriminating value* was often judged to be small. The panel took into consideration the fact that many student projects receive the same grade through the simple addition of a number of component scores covering diverse aspects of the process, rather than on the basis of a clear model of the skills in question.

Second judgement round. In the second round, the panellists answered 18 pre-structured questions per set of materials about a number of criteria regarding the quality of the teachers' assessment practice (see Table 2). They could indicate the degree to which each criterion was met on a four-point scale, ranging from 'not at all' to 'to a great degree'; they could also indicate their inability to assess this. Finally, they had the option of clarifying their answer. Table 2 contains the sums of the percentages of panellists who responded 'to a moderate degree' and 'to a great degree'. Approximately 20% of most questions were answered with 'I cannot judge'. Exceptions were Questions 1 (6%), 16 (36%) and 17 (38%). The percentages in the table are not corrected for this.

The panellists valued most the comprehensibility of the teacher's criteria (Question 1), their relevance to the teacher's goals (3,5), their suitability for assessment of the product (7), their applicability to the assignment (14), and the way the teacher handled the criteria (16,17). The panellists were critical of how teachers' criteria suited the task of assessing subject content accuracy (8) and students' knowledge and skills (11,12), and the degree to which there was a question of improper criteria (13) and of a sensitivity to errors (18). Explanations added by the panellists showed that the most critical issue was that in a number of cases, assessment criteria were lacking or were not sufficiently explicit. Additionally, inconsistencies were detected between goals and criteria, between assignments and criteria, and among the

Table 2. The panel's judgements of teachers' assessment practices using Likert-type questions about psychometric quality criteria

	% to a moderate or great degree
1. Are the teacher's assessment criteria comprehensible?	67
2. Are the criteria formulated unambiguously?	54
3. Are the criteria relevant in view of the learning goals the teacher is striving for with the assignment?	71
4. Are all the learning goals the teacher is striving for covered?	58
5. Are the criteria relevant in view of the goals the teacher is striving for by assessing the assignment?	65
6. Are the criteria suitable for assessing the research process?	54
7. Are the criteria suitable for assessing the research product (the report)?	73
8. Are the criteria suitable for assessing the correctness with respect to content?	46
9. Are the criteria suitable for assessing the depth or quality level with respect to content?	58
10. Are the criteria suitable for assessing the quality of the research as research?	56
11. Are the criteria suitable for assessing the students' knowledge on doing research?	40
12. Are the criteria suitable for assessing the students' research skills?	46
13. Is there question of improper criteria (assessing other knowledge, skills or attitudes)?	54
14. Are the criteria applicable in view of the assignment given to the students?	67
15. Should the teacher, in view of his/her assessment goals, criteria, etc., have been using more interim assessments (during the inquiry process) than he/she actually did?	50
16. Is the way the teacher handles the criteria (the scoring, the use of weights [if any], etc.) relevant in view of the goals the teacher is striving for by assessing the assignment?	60
17. Is the way the teacher handles the criteria efficient?	60
18. Is this set of criteria prone to making faults and mistakes?	58

*The four scale points were: not (coded as 1), to a little degree (2), to a moderate degree (3), to a great degree (4). Shown are the sums of the latter two.

criteria. Furthermore, in a number of cases, it was not entirely clear how many of the criteria had been made known to the students beforehand. Finally, submitted assignments did not always represent research projects. To summarize: regarding *reliability*, the assessment practices that were judged by the panel satisfied some aspects well (see Question 1), some moderately (2), and some poorly (18). Similarly, they satisfied some aspects of *validity* well (3, 5, 7 and 14), some moderately (4, 6 and 10), and some poorly (11, 12 and 13). They also satisfied some aspects of *acceptability* well (16 and 17), some moderately (9 and 15), and some poorly (8).

Discussion

Teachers' assessment practices

Our first question was: how do upper secondary school teachers in the natural and social sciences assess the research skills of their students? To answer this question we conducted two surveys, generating a total of 214 responses from heads of departments in the subject areas of physics, biology, history, economics and geography.

Teachers find it important that systematic attention be paid to research skills in their discipline (95% to a moderate or strong degree). They focus on several goals, above all developing knowledge of subject matter, stimulating independent learning, and providing research experience (80–90%). Teachers conceive research skills chiefly in terms of consecutive steps, laying the most emphasis on two steps: the formulation of the research question and the drawing of conclusions (95%).

The research assignments that teachers give to students are designed by the teachers themselves (70%) or taken from a textbook (65%). According to teachers, assignments must comply with a number of demands, among which are that they provide students with freedom of topic choice, are solvable in a number of ways, are sufficiently challenging, can be completed within a certain time, return testable results, provide the opportunity for reflection, conform to examination requirements, and make use of the materials available at the school (70–90%). The assignments teachers actually use typically show a high degree of prior structuring (65%).

Teachers vary greatly in the amount and type of assistance and feedback they give. Teachers also vary greatly in a number of other aspects of assessment: frequency, performance criteria and scoring (in particular whether or not the research process is also assessed, implicitly or explicitly), communication with students (the degree of explication before the assigment is prepared), the assessment approach (global or analytical), whether the assessment process involves working with a colleague as co-assessor, whether students are given a role in the assessment (using peer- and/or self-assessment), whether and which measures are taken to support assessment quality, the weighting of assessment components, and whether norm-referenced or criterion-referenced norms are used.

The quality of teachers' assessment practices

Our second question was: what is the quality of the teachers' assessment practices? A panel of experts assessed in two rounds respectively 26 and 24 teacher-submitted assignments together with associated information and materials, in a qualitative respectively quantitative manner.

According to the panel in the first round, there was substantial consistency between assignments and assistance, and between goals and assistance. Less consistency was evident between goals and assignments, between assignments and assessment criteria, and between assistance and criteria. The least consistency was found between goals and criteria. Teachers quite often formulated more goals than can be recognized in the assignments and often considerably more goals than can be recognized in the criteria. Moreover, in a number of cases, the panel found the criteria poorly formulated.

In the second round, the assessment practices of teachers measured up variously well, moderately and poorly, with regard to each of the main quality criteria, reliability, validity and acceptability (see Table 2; the quality criteria were not or to a small degree met in 30–60% of the cases). An important part of the panel's observations dealt with validity. In a number of cases, it was not entirely clear which knowledge and skills were intended to be assessed. Also, teachers' goals failed to find valid expression in the criteria, and factors other than those intended were measured. Furthermore, question marks placed by panellists indicated acceptability problems because the assessment was not objective or not fair and/or failed to provide insight. The discriminating power was often judged to be small, owing to the fact that students' projects were judged by the teachers using the simple addition of a number of component scores covering diverse aspects. The panellists' remarks showed that they differed as to how far teachers can be expected to tailor their assessment process to meet all of the psychometric quality criteria.

Generalizability

As to the generalizability of these findings, not all subject areas were involved in the study, but those participating in the surveys were well spread over the natural and social sciences. The teachers who submitted materials were less focused on subject matter and more interested in research as such than the other respondents to the surveys, and they were relatively systematic in the design and assessment of their research assignments. Thus, the self-selection appears to be a conservative one: the quality of the average teacher's assessment practice is more likely to be overestimated than underestimated.

Interpretations and consequences

The results give rise to several interpretations.

The participating teachers appear to support the conception of research in terms of consecutive steps as incorporated in the national examination requirements, and they give students quite pre-structured assignments and quite a lot of assistance. This could mean that they primarily value maintaining control over the student activities.

The teachers believe that assignments should give students freedom of topic choice, should present problems that are capable of being solved in a variety of ways and should be challenging. At the same time, the teachers believe that assignments should conform to examination requirements, should be able to be completed within a certain time, and should make use of the materials available in schools. It could well be difficult to satisfy all of these demands simultaneously.

The greatest differences among the teachers showed up in various aspects of assessment. This might be an indication that assessment is the task domain receiving the least support from textbooks and from colleagues.

Teachers while assessing typically weigh heavily the research questions and the corresponding conclusions, that is, the disciplinary content. This could be at the expense of developing students' research skills.

That many teachers view research in terms of consecutive steps, give pre-structured assignments and heavily weigh the disciplinary content possibly indicates that they are not yet very familiar with the phenomenon of research themselves. This would confirm the results of the two Dutch studies described earlier (Van Tilburg & Verloop, 2000; Van Rens & Dekkers, 2000), that teachers have little knowledge of research.

In developing their instructions, assignments and assessments, there is currently no conclusive theoretical model that teachers can turn to for the development of research skills, and they must rely on the examination requirements and their own goals. The results of our study indicate that they try to make the assessment fair by giving students enough time, equal access to information, sufficient material and precise instruction. Nevertheless, according to the panel, there are grounds for serious concern regarding the clarity of the assessment criteria, the consistency between teachers' goals, assignments, assistance and criteria, and the validity and acceptability of teachers' assessment practices.

The question now is how to evaluate these results. Firstly, in this study we focused on a certain form of research, namely, 'investigation' (Duggan & Gott, 1995). The results do not have to be valid for other forms of student research, such as standard laboratory experiments and exploratory practical work (Meester & Kirschner, 1995). Secondly, it appeared that the psychometric requirement to base the assessment on an explicit theoretical model of the skill (Messick, 1984, 1995) can not be met, because a good model is not yet available. To be sure, the examination requirements and also the teachers unfold research in a number of steps but these steps are not skills in the classic psychometric sense. Moreover, thinking in terms of concrete practical steps does no justice to the more conceptual aspects of doing research (Kirschner, 1992; Gott & Duggan, 1996; White & Frederiksen, 1998). Thirdly, the teachers are consistent in developing their assessments around these steps and the steps are also general enough to accommodate the variation in students' choices and products. However, a critical judgement of the present practice of 'authentic assessment' (Messick, 1994, 1995; Haertel, 1999), as conducted by our panel, reveals some problems. These problems include the reliability and quality control (mostly, the students' teacher is the only assessor), the validity (in a number of cases it is not clear whether the teacher is assessing knowledge and skills concerning research), and the acceptability (the criteria are not always clear, known by the students, and consistent). In spite of the Dutch code for assessment in secondary education (Creemers-van Wees *et al.*, 1997), teachers in the Netherlands obviously have difficulties in fulfilling their assessment task, just as Stiggins and Bridgeford (1985) found in the USA. Looking at Figure 1, the main problems seem to concentrate at the quality criteria process (2b), interpretation (2d), generalizability (2g), transparency (3b), and discriminating power (4b).

Our conclusion is that the relevance of the psychometric quality criteria deserves critical reflection and differentiation. One point of view is that teachers cannot reasonably be required to meet the quality criteria in view of their existing competencies and the time and facilities they have available. However, in our view, the standards by which those practices are to be judged depend on the role of the assessment: whether it is formative or summative, for the students, the teacher and/or others. In all cases validity should have priority (cf. Linn *et al.*, 1991; Linn, 1994), because assessment is all about the intended knowledge or skills. Transparancy, lack of bias, and functionality should always be satisfactory. In formative assessment, reliability, objectivity, and equality might get less priority, because the results only indicate possibilities for improvement. Workability and efficiency might get a high priority, so that feedback can be frequently given. In summative assessment reliability, objectivity and equality are also important. For high-stake purposes, special measures for quality assurance should be routine. These might include using standards in the form of model products or model answers on each performance level, or at least anchor points in the form of detailed descriptions (fostering objectivity); using more than one assessor, such as the students' own teacher and a colleague (controlling for reliability); and using information from more than one source, such as a final research assignment and the average score on a number of earlier assignments (contributing to equality and generalizability).

Assessing students' knowledge and skills is a complex and wide-ranging part of the educational process, particularly in view of the recent trends towards new goals, new views of learning and teaching, and new assessment strategies. The popularity of assessing complex skills using more authentic assignments is not without complications: assignments giving students options for topic choice can decrease the equality of the assessment between students, and assignments with multiple good answers can make it difficult to develop a scoring scheme. Whoever seeks to evaluate assessment practices in schools must address and do justice to these facts. Yet, we doubt whether the teachers' assessment practices described here are sufficiently defensible. The teachers do their best, but the acceptability of their assessments, the instructiveness for the students and the validity in view of the learning goals are improvable.

First, and at the very least, we should expect teachers to make their assessment criteria explicit and to communicate them to, and discuss them with, their students. Secondly, goals, assignments, assistance, and assessments should be tailored to each other. Teachers' practices should be educationally consistent. Thirdly, formative and summative assessments might be distinguished more clearly. Integrating education and assessment can have positive effects on students' learning activities, processes and results. Communicating and discussing the assessment criteria can help to clarify the learning goals and their relevance, can motivate the students and can make realistic and instructive feedback possible (Crooks, 1988; Snow & Lohman, 1989; Boekaerts, 1991; Rowe & Hill, 1996). However, while integrating education and assessment can improve the assessments' transparency and acceptability, it also makes the assessment more vulnerable. Therefore, in high-stake summative assessment, special measures for quality assurance should be routine (see above). Fourthly, individual teachers could be supported in their instruction and assessment of students' research skills. Their own uncertainty concerning research could be decreased by having structured discussions with each other leading to a clearer vision of the meaning and characteristics of research, the learning goals to be set for the students, and the assessment criteria to be used.

It can be expected that teachers will increasingly have to judge students' skills using more authentic assignments. Teachers' own professional skills should therefore develop in this direction. Currently, teachers typically receive little training and support in matters of assessment. Improvements should be possible by incorporating support in textbooks and other teaching materials, giving teachers sufficient time and explicit responsibilities, and improving conditions at school for learning and development; for example, by facilitating commencement and use of a systematic collection of assignments, assessment criteria, students' model products and teachers' experiences in using them.

The results of this study give cause for follow-up research, at least along the following lines. First, concerning the quality of teachers' assessment practices, there is a need for explicit, documented standards setting out acceptable minimum levels of reliability, validity, acceptability and practical utility. To develop such standards in a realistic and useful way requires further empirical research (cf. Van der Schaaf *et al.*, 2003). Secondly, good assignments are crucial for both education in general and assessment in particular: a key question is how to develop assignments that meet such diverse requirements as authenticity and controllability. The claims in the literature that the newer forms of assessment are more valid and suitable for classroom use still await further empirical confirmation. Thirdly, research is necessary to determine the extent to which research skills can be applied across all disciplines or are discipline-specific, and whether and how the development of subject knowledge and research skills could go hand in hand.

Acknowledgements

This research was funded by the Dutch Organization for Scientific Research (A575–35–007).

References

Alexander, P. A., Schallert, D. L. & Hare, V. C. (1991) Coming to terms: how researchers in learning and literacy talk about knowledge, *Review of Educational Research,* 61(3), 315–343.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990) *Standards for teacher competence in educational assessment of students* (Washington DC, AFT, NCME & NEA).

Anderson, J. R. (1982) Acquisition of cognitive skill, *Psychological Review,* 89, 369–406.

Arter, J. (1993) *Designing scoring rubrics for permance assessments: the heart of the matter* (Portland, OR, Northwest Regional Educational Laboratory, Test Center).

Baker, E. L., Abedi, R. L., Linn, R. L. & Niemi, D. (1995) Dimensionality and generalizability of domain-independent performance assessment, *Journal of Educational Research,* 89(4), 197–205.

Bereiter, C. & Scardamalia, M. (1993) *Surpassing ourselves. An inquiry into the nature and implications of expertise* (Chicago, IL, Open Court).

Birenbaum, M. (1994) Toward adaptive assessment—the student's angle, *Studies in Educational Evaluation,* 20, 239–255.

Boekaerts, M. (1991) Subjective competence, appraisals and self-assessment, *Learning and Instruction,* 1, 1–17.

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(3), 377–391.

Colley, A. M. & Beech, J. R. (1989) *Acquisition and performance of cognitive skills* (Chichester, John Wiley & Sons).

Cowie, B. & Bell, B. (1999) A model of formative assessment in science education, *Assessment in Education,* 6(1), 101–115.

Creemers-van Wees, L. M. C.M., Knuver, J. W. M., Vos, H. J. & Van der Linden, W. J. (1997) *Toetsen, beoordelen en beslissen in het voortgezet onderwijs* [Testing, assessing and deciding in secondary education] (Enschede, OCTO).

Cronbach, L. J. (1988) Five perspectives on validity argument, in: H. Wainer & H. I. Braun (Eds) *Test validity* (Hillsdale, NJ, Lawrence Erlbaum).

Crooks, T. J. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research,* 58(4), 438–481.

Crooks, T. J. & Kane, M. T. (1996) Threats to the valid use of assessments, *Assessment in Education,* 3(3), 265–285.

De Groot, A. D. (1970) Some badly needed non-statistical concepts in applied psychometrics, *Nederlands Tijdschrift voor de Psychologie,* 25, 360–376.

Duggan, S. & Gott, R. (1995) The place of investigations in practical work in the UK National Curriculum for Science, *International Journal of Science Education,* 17(2), 137–147.

Fischer, K. W. (1980) A theory of cognitive development: the control and construction of hierarchies of skills, *Psychological Review,* 87(6), 477–531.

Fitts, P. M. & Posner, M. I. (1967) *Human performance* (Belmont, CA, Brooks/Cole).

Frederiksen, N., Glaser, R., Lesgold, A. & Shafto, M. G. (Eds) (1990) *Diagnostic monitoring of skill and knowledge acquisition* (Hillsdale, NJ, Lawrence Erlbaum).

Gee, B. & Clackson, S. G. (1992) The origin of practical work in the English school science curriculum, *SSR,* 73(265), 79–83.

Glaser, R. & Silver, E. (1994) Assessment, testing and instruction: retrospect and prospect, *Review of Research in Education,* 20, 393–419.

Goldstein, H. (1979) *The design and analysis of longitudinal studies* (London, Academic Press).

Gott, R. & Duggan, S. (1996) Practical work: its role in the understanding of evidence in science, *International Journal of Science Education,* 18(7), 791–806.

Haertel, E. (1985) Construct validity and criterion-referenced testing, *Review of Educational Research,* 55(1), 23–46.

Haertel, E. H. (1999) Performance assessment and educational reform, *Phi Delta Kappan,* 80(9), 662–666.

Hambleton, R. K. & Murphy, E. (1992) A psychometric perspective on authentic measurement, *Applied Measurement in Education,* 5(1), 1–16.

Hodson, D. (1992) Assessment of practical work, *Science & Education,* 1, 115–144.

Kane, M. T. (1992) An argument-based approach to validity, *Psychological Bulletin,* 112(3), 527–535.

Keeves, J. P. (1994) Methods of assessment in schools, in: T. Husén & N. Postlethwaite (Eds) *International encyclopedia of education* (2nd edn) (New York, Pergamon).

Kirschner, P. A. (1992) Epistemology, practical work and academic skills in science education, *Science & Education,* 1, 273–299.

Linn, R. L. (Ed) (1989) *Educational measurement* (3rd edn) (New York, Macmillan).

Linn, R. L. (1994) Performance assessment. Policy promises and technical measurement standards, *Educational Researcher,* 23(9), 4–14.

Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991) Complex, performance-based assessment: expectations and validation criteria, *Educational Researcher,* 20(8), 15–21.

Lock, R. (1989) Assessment of practical skills. Part 1. The relationships between component skills, *Research in Science & Technological Education,* 7(2), 221–233.

Lock, R. (1990) Assessment of practical skills. Part 2. Context dependency and construct validity, *Research in Science & Technological Education,* 8(1), 35–52.

Madaus, G. & Kellaghan, T. (1992) Curriculum evaluation and assessment, in: P. W. Jackson (Ed.) *Handbook of research on curriculum* (New York, Macmillan).

Meester, M. A. M. & Kirschner, P. A. (1995) Practical work at the open university of the Netherlands, *Journal of Science Education and Technology,* 4(2), 127–140.

Messick, S. (1984) The psychology of educational measurement, *Journal of Educational Measurement,* 21(3), 215–237.

Messick, S. (1989) Validity, in: R. L. Linn (Ed.) *Educational measurement* (3rd edn) (New York, Macmillan).

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher,* 23(2), 13–23.

Messick, S. (1995) Validity of psychological assessment. Validation of inferences from person's responses and performances as scientific inquiry into score meaning, *American Psychologist,* 50 (9), 741–749.

Millman, J. & Greene, J. (1989) The specification and development of tests of achievement and ability, in: R. L. Linn (Ed.) *Educational measurement* (3rd edn) (New York, Macmillan).

Mitroff, I. & Turoff, M. (1975) Philosophical and methodological foundations of Delphi, in: H. A. Linstone & M. Turoff (Eds) *The Delphi method: techniques and applications* (London, Addison-Wesley).

Moss, P. A. (1992) Shifting conceptions of validity in educational measurement: implications for performance assessment, *Review of Educational Research,* 62(3), 229–258.

Nitko, A. J. (1989) Designing tests that are integrated with instruction, in: R. L. Linn (Ed.) *Educational measurement* (3rd edn) (New York, Macmillan).

Novak, J. R., Herman, J. L. & Gearhart, M. (1996) Establishing validity for performance-based assessments: an illustration for collections of student writing, *Journal of Educational Research,* 89(4), 220–233.

Reed, G. F. (1968) Skill, in: E. A. Lunzer & J. F. Morris (Eds) *Development in human learning* (New York, Elsevier).

Rowe, K. J. & Hill, P. W. (1996) Assessing, recording and reporting students' educational progress: the case for 'subject profiles', *Assessment in Education,* 3(3), 309–352.

Seyfart, J. T., Simon, D. J. & Schlesinger, J. (1994) Assessing student performance: are our assumptions valid? Paper prescribed at the *Annual Meeting of the American Association of Colleges of Teacher Education, Chicago.*

Snow, R. E. & Lohman, D. F. (1989) Implications of cognitive psychology for educational measurement, in: R. L. Linn (Ed.) *Educational measurement* (3rd edn) (New York, Macmillan).

Stiggins, R. J. & Bridgeford, N. J. (1985) The ecology of classroom assessment, *Journal of Educational Measurement,* 22(4), 271–286.

Stokking, K. & Voeten, R. (2000) Valid classroom assessment of complex skills, in: R. Simons, J. van der Linden & T. Duffy (Eds) *New learning* (Boston, MA Kluwer).

Thair, M. & Treagust, D. F. (1997) A review of teacher development reforms in Indonesian secondary science: the effectiveness of practical work in biology, *Research in Science Education,* 27(4), 581–597.

Van der Schaaf, M. F., Stokking, K. M. & Verloop, N. (2003) Developing performance standards for teacher assessment by policy capturing, *Assessment and Evaluation in Higher Education,* 28(4), 395–410.

Van Rens, E. M. M. & Dekkers, P. J. J.M. (2000) Leren onderzoeken—de rol van de docent [Learning to do research—the role of the teacher], *Tijdschrift voor Didactiek der Beta-wetenschappen,* 17(1), 76–94.

Van Tilburg, P. A. & Verloop, N. (2000) Kennis van en opvattingen over het onderwijzen van onderzoeksvaardigheden [Knowledge of and opinions on the teaching of research skills], *Tijdschrift voor Didactiek der Beta-wetenschappen,* 17(1), 60–75.

White, B. Y. & Frederiksen, J. R. (1998) Inquiry, modeling, and metacognition: making science accessible to all students, *Cognition and Instruction,* 16(1), 3–118.

Wiggins, G. (1989) A true test: toward more authentic and equitable assessment, *Phi Delta Kappan,* 70, 703–713.

Wiggins, G. (1993) Assessment: authenticity, context, and validity, *Phi Delta Kappan,* 74, 200–214.

Willett, J. B. (1988) Questions and answers in the measurement of change, *Review of Research in Education,* 15, 345–422.

Wolf, D., Bixby, J., Glenn, J. & Gardner, H. (1991) To use their minds well: investigating new forms of student assessment, *Review of Educational Research,* 17, 31–74.