AN EVALUATION OF THE SOLE USE OF SHORT-ANSWER TESTS IN APPRENTICESHIP EXAMINATIONS

Report to the Department of Education and Science

Thomas Kellaghan, Mark Morgan, Maeve Fitzpatrick and David Millar

with an Appendix by Gerhard Kohn

Educational Research Centre
St Patrick's College, Dublin

September 2002

CONTENTS

Summary 1

Terms of Reference 4

Acknowledgements 6

I. Introduction 7

The Context of the Study 7

Views about Apprenticeship Short-Answer Items 11

Procedure 13

Outline of Report 15

II. Considerations in Test Development 16

1. The Purpose or the Nature of the Inferences to be Made from Test Scores Should be Identified 16

2. The Various Components of a Curriculum Should be Appropriately Represented in Items of the Test 25

3. Tests Should be Reliable, that is They Should Consistently Place Students in the Same Relative Position if the Test were Given Repeatedly 27

4. The Effects of Tests on Teaching and Learning Should be Taken into Account in Their Design 30

5. Economy and Feasibility Should be Taken into Account in Test Design 33
6. Account Should to be Taken of the Fact That Apprenticeship Examinations are Administered Frequently (Every Term) and in Several Locations 33
7. Account Should be Taken of How to Facilitate Students who are Unsuccessful 33
III. Standards 35
IV. Item Types 40
Item Types 41
Does Choice of Item Type Matter? 44
V. Conclusion 54
Guidelines for Test Construction 54
Implications of the Guidelines for Apprenticeship Examinations 56
Recommendations 59
References 66
Appendix: Final Assessment/Testing in Vocational Education and Training in France, Germany, and the Netherlands 73

AN EVALUATION OF THE SOLE USE OF SHORT-ANSWER TESTS IN APPRENTICESHIP EXAMINATIONS

SUMMARY

The sole use of short-answer tests for Phases 4 and 6 (off-the-job) in apprenticeship examinations has been a matter of debate for some time. While FÁS favours their use, the Institutes of Technology (IT) in which apprentices spend Phases 4 and 6 do not. The study described in this report considered aspects of their use for the following crafts: Carpentry, Electrical, Fitter, Toolmaker, and Blocklayer. The study procedure involved a review of literature on test development and, in particular, on the format of test items; a review of documentation relating to apprenticeship examinations; meetings with stakeholders; analyses of a number of Phase 4 and Phase 6 examinations to determine the level of response (recall, understanding, application) required by test items; key learning point categories represented in tests; and the extent to which curriculum activities were represented. Analyses of examinees' responses were also carried out to determine reliabilities of tests and the difficulty level of individual items.

Seven considerations associated with 'best practice' in the construction of tests to assess students' achievements at the end of a phase of study – whatever item format is used – are outlined. Following this, a brief commentary on standards is presented. It would seem that interpretation of the term 'standard' when applied to apprenticeship examinations differs considerably from interpretation in much contemporary discourse on the topic.

Item formats are considered in the context of: curriculum coverage in a test; the typical cognitive demands and the range of cognitions they elicit; their

relationship to construct-irrelevant factors; their effects on teaching and learning; reliability; and cost. It is concluded that a mixture of item formats will facilitate the achievement of wide coverage of the content of a curriculum in a test, as well as of a wide range of cognitive activity.

On the basis of the review of 'best practice' in constructing tests and in the use of varying item formats, the following recommendations for apprenticeship tests are made:

1. Tests should comprise more than one item type (short-answer and essay-type).
2. The number of short-answer items should be increased to provide more extended curriculum coverage. Examinees would be required to respond to all such items.
3. Essay-type items should be designed to elicit higher-order cognitive processes involving comprehension, analysis, evaluation, problem-solving, and students' ability to organize and apply knowledge.
4. Examinees should be provided with a choice of prompts in essay-type items to allow them to choose an area in which they can best express themselves.
5. Care will be required in essay-type examinations to ensure that examinees' ability to demonstrate their knowledge is not inhibited by language difficulties.
6. A system of moderation of the examination process will be required.
7. Steps should be taken to enhance the reliability of tests.
8. The number of marks allocated to examinee responses on each examination paper should be increased (to 100).
9. The number of marks allocated to questions will vary depending on the difficulty of a question and the demands that it makes on examinees.
10. The pass mark should be reduced to 50%. The credit mark should also be reduced.
11. Examination arrangements should be such that they facilitate students who need to repeat an examination.

12. Consideration should be given to issues relating to the administration of the examinations (e.g., whether they will be devised centrally or locally, release of examination papers, use of item banks).

AN EVALUATION OF THE SOLE USE OF SHORT-ANSWER TESTS IN APPRENTICESHIP EXAMINATIONS

TERMS OF REFERENCE

The training of craft apprentices follows an orderly programme agreed between the Department of Education and Science and FÁS. This programme includes two phases of block-release by employers to Institutes of Technology and a small number of other educational institutions.

During and at the end of these phases in the educational system, apprentices are evaluated by means of assessments or examinations. The methodology used in these written tests is the subject of the evaluation. This research project is confined to written tests and does not embrace practical tests.

The contractor shall carry out an evaluation of the sole use of short-answer question papers to assess their efficiency and effectiveness in assessing the cognitive attributes, skills and competencies required by crafts persons in the designated trades.

The research project is to be confined to the following skills/trade areas: Carpentry, Electrical, Fitter, Toolmaker, Blocklayer.

Four components of the evaluation are envisaged:

1. An examination of the extent to which tests currently in use reflect the range of knowledge/skills in the curriculum. Breadth and depth of coverage will be considered.
2. Since assessment is standards based, it is appropriate to examine methods by which standards for these national exams are set and the adequacy of procedures to establish that successful students have achieved an appropriate level of

competence. The study does not embrace the process of curriculum development or review.

3. In the examination of current tests, it is proposed that an analysis of the performance of students in recent test administrations be carried out to determine the nature of score distributions, item difficulty levels, and item discrimination levels.
4. Procedures for assessment in the Irish system will be compared with systems in a number of other countries. The contractor should identify a representative set of systems, both confined and not confined to the sole use of short answer questions for this comparison.

ACKNOWLEDGEMENTS

The authors acknowledge the support of all stakeholders in apprenticeship examinations who provided material and information for this study. FÁS made available a wide range of documentation and the Institutes of Technology provided data on student examination performance. Personnel from the Department of Education, FÁS, the Institutes, and the Teachers Union of Ireland made themselves available for discussion.

Thanks are also due to Maurice Doran for his support and assistance during the study and to Hilary Walshe who typed the manuscript.

I. INTRODUCTION

THE CONTEXT OF THE STUDY

The context of the study described in this report is the Irish Apprenticeship System, which came into force in 1994. A National Apprenticeship Advisory Committee was set up by the Board of FÁS to oversee its implementation. Terms of reference were determined by the FÁS Board, to which the Committee reports. Membership of the 25-strong Committee comprises: a Chairperson nominated by FÁS, eight members from the FÁS board, five employer representatives, five trade union representatives, one member representing Institutes of Technology, one member representing FÁS executive, one member representing FÁS staff, one member representing the Department of Enterprise and Employment, one member representing the Department of Education and Science, and one member representing the Dublin Institute of Technology.

Apprenticeships are open to individuals over the age of 16 who have a minimum of a grade D in five subjects in the Junior Certificate Examination or its equivalent. Individuals over the age of 25 may qualify if they have relevant experience of at least three years and are successful at interview (Ambrosio et al, 1995). Actually, more than half of apprentices have completed the Leaving Certificate. In 2000, the figure was 53.1%, but there were large differences between

trades. While 66.8% of apprentices in printing trades and 63.7% in electrical trades had taken the Leaving Certificate Examination, the percentages for the construction (43.1%) and motor (43.3%) trades were lower (Kerr, 2002).

In June 2002, there were 25,431 registered apprentices. Of these, 7,306 were preparing to become electricians, 4,984 carpenters, 2,732 plumbers, 1,572 motor mechanics, 1,298 bricklayers, and 1,145 fitters. There has been considerable growth in the number of apprentices between 1996 and 2001, during which time the number in electrical trades grew by 132% and in carpentry by 129%. Apprentices in other trades also increased substantially in number. The increase has come at a time when enrolment in some third-level courses in Institutes of Technology has been falling (McDonagh, 2001).

Almost all (99.5%) apprentices are male, despite a commitment in the Programme for Economic and Social Progress (1991) to increase female participation in non-traditional areas of apprenticeship and training. While females are also in a minority in apprenticeships in other countries, a greater proportion is found in traditional male occupations than is the case in Ireland. The socioeconomic background of apprentices more closely resembles the background of the general population than is the case with entrants to other sectors of post-secondary education: while children of higher professional classes are underrepresented among apprentices, and children of skilled manual classes somewhat overrepresented, the children of other socioeconomic groups are equitably represented (McDonagh, 2001).

Curricula are developed for each trade by subject-matter experts representing employers, trade unions, FÁS, and staff of Institutes of Technology. These were based on an occupational analysis of each trade following a postal survey of employers and interviews with employees to identify the skills, knowledge, and attitudes required of

craftpersons. Identified skills were ranked in order of importance and frequency of use and were used to form an occupational profile of each trade which contained four areas of skill: core (essential skills required by all craftpersons in a trade); specialist (applicable to specialist sectors); common (required by a trade, but also common to other trades within a family or group of trades); and personal (applicable to all trades and incorporating the practical application of abilities such as report writing and customer relations) (ESF, 1995).

Apprenticeships are described as 'standards-based' and consist of seven phases spread over four years in which the apprentice spends Phases 1, 3, 5, and 7 with an employer on the job; Phase 2 (consisting of 20 weeks of basis training) in a FÁS Training Centre; and Phases 4 and 6 (lasting 10 or 11 weeks, depending on the trade) in an Institute of Technology. (A small number of trades operate to a different phasing duration.) At the end of each phase, apprentices are required to demonstrate a satisfactory level of competence. The focus of the present study is the assessment procedure following formal instruction in Institutes of Technology in phases 4 and 6 for apprentices in the Carpentry, Electrical, Fitting, Toolmaking, and Bricklaying trades. These trades were selected as representative of trades in general and because they involve large numbers of apprentices.

Assessments generally comprise a practical examination, at least one theory examination, and an examination in related subjects such as drawing and applied mathematics. Draft questions and marking directions, which should adhere to a number of criteria specified by FÁS, are submitted by teachers in Institutes of Technology to the relevant FÁS Certification and Standards project teacher/manager, who then assembles the tests and makes them available to colleges.

The appropriateness of the sole use of short-answer questions in these assessments, which FÁS (1999) states have 'a recognised advantage of being generally easy to construct and mark' (p. 4), has been a matter of controversy. In June 1998, the Department of Education and Science agreed to their use for a trial period, on condition that 'an independent evaluation of the sole use of short-answer question papers be undertaken to assess their efficacy and effectiveness in assessing the cognitive attributes, skills, and competencies required by craftspersons in the designated trades' (letter from Denis Healy, Assistant Secretary General, to Donal Kerr, Manager, Certification and Standards, FÁS, 17 July 1998).

The study described in this report was designed to evaluate the sole use of short-answer question papers in examinations to assess candidates' achievements. Two preliminary comments about the study may be made. First, although the term 'short-answer question' is used to describe the apprenticeship tests, the questions are not really short-answer as the term is usually employed. One need only consider the fact that students are usually required to answer only 10 to 20 questions in periods ranging from an hour and a half to three hours to realize that much more time is allowed for an answer that would be the case when what are traditionally called short-answer questions are used (a minute to a minute and a half per question). Secondly, the evaluation could not be carried out without reference to the context in which the examinations are held and without some analysis of those examinations, since there are no rules about the type or types of item that are most appropriate in examining students in all situations.

The main focus of the study is an analysis of assessment procedures and of the characteristics of item types (their ability to assess students' achievements, the cognitive demands they make on students, and their effects on teaching and learning). The analysis is supported, where available, by empirical data. In addressing the task

posed for the study, an examination of some aspects of the current method of assessment was carried out. Apart from meeting the need to contextualize the study, such an examination was considered necessary since it was not clear whether concerns expressed about the use of short-answer question papers in apprenticeship examinations were based on principle about the inadequacy of this procedure, or if they arose because of perceived inadequacies in its implementation.

VIEWS ABOUT APPRENTICESHIP SHORT-ANSWER ITEMS

According to the ESF (1999) report on apprenticeship and traineeship, the apprenticeship assessment system is in need of 'refinement' (p. xi). Criticisms by Institute of Technology (IT) staff have focused on the use of the short-answer form in Phase 4 and 6 examinations. There is considerable agreement on these in a submission that was prepared for our study by the Teachers' Union of Ireland, in a published paper of a survey of IT teacher views (O'Connor & Harvey, 2001), and in the views of staff obtained in interview for the present study.

Some of the criticisms would seem to be applicable to the short-answer form in any circumstances; some are not intrinsic to this form of item, and apply to the way it is being used at Phases 4 and 6; other criticisms, while based on the present system, could be regarded as relating to problems that are likely to arise when short-answer items are used.

Criticisms that were made of the short-answer form that might be considered intrinsic to this form of assessment included:

the fact that the same marks are allocated to all items despite obvious differences in the complexity and level of difficulty of items;

the inability to give credit for 'partially' correct responses; and

the lack of student choice of items to respond to on the examination paper.

It would be unusual in a short-answer test to depart from any of these conventions.

Criticisms that relate to the specific use of short-answer papers at Phases 4 and 6 included:

the reuse of examination papers;

the prohibition on officially releasing papers after an examination is taken (which, if released, would provide students with guidance in their study);

leakage of examination questions/papers;

lack of standardization in marking across colleges;

a pass mark (70%) that is considered excessively high;

lack of recognition of more than three levels of performance (fail, pass, credit).

It is difficult to say whether further common criticisms apply only to perceptions of the present system or represent a more basic criticism of the short-answer format. The criticisms are:

that the tests provide inadequate curriculum coverage;

that the knowledge that is assessed is superficial;

that the focus is on recall, with the result that a correct response does not necessarily mean that the student understands a concept or can apply it;

that higher-order knowledge is not assessed and that, as a consequence, candidates are not given the opportunity to display such knowledge or their diagnostic and problem-solving skills; and

that the backwash of the assessment on teaching and learning results in inadequate attention to more advanced forms of knowledge and communication skills.

Some of the criticisms probably arise from the fact that the form of Phase 4 and 6 tests differs so much from the form used in Institutes of Technology in other

programmes of study and, indeed, differs from the form that they were accustomed to in the Junior and Senior Trade Examinations for apprentices which had been administered by the Department of Education prior to the introduction of the new standards-based apprenticeship system.

PROCEDURE

The following steps were taken in carrying out the study.

1. Literature on test development and, in particular, on the format of test items was reviewed.
2. Documentation relating to apprenticeship examinations was reviewed.
3. Meetings were held with staff in FÁS, Institutes of Technology, the Department of Education and Science, and the Teachers' Union of Ireland (TUI) to discuss issues surrounding the apprenticeship examinations. A written submission was received from the TUI.
4. While a comprehensive review of existing apprenticeship examinations was not carried out, a number of analyses on a sample of examinations were.
   (a) Analyses of examinations were carried out by IT staff to a specification provided by the investigators to determine the level of response (recall, understanding, application) required of examinees in each of the following: Carpenter/Joiner, Phase 4; Carpenter/Joiner, Phase 6; Electrical Science, Phase 6; Electrical Craft Practice, Phase 6 (four examinations for each subject).
   (b) Analyses of examinations were carried out by IT staff to a specification of the investigators to determine the key learning point categories represented in

each of the following: Electrical Science, Phase 6; Electrical Craft Practice, Phase 6 (four examinations for each subject).

(c) Analyses of examinations were carried out by IT staff to a specification of the investigators to determine the extent to which curriculum activities were represented in each of the following: Electrical Science, Phase 6; Electrical Craft Practice, Phase 6 (four examinations for each subject).

(d) In the study proposal, it was stated that an analysis of performance in recent test administrations would be carried out to estimate the reliabilities of tests and to determine the 'difficulty' level of individual items. For this component of the study, colleges provided data for examinations in Brickwork (Phase 6), Carpenter/Joiner (Phases 4 and 6), Electrical (Phases 4 and 6), and Fitter (Phase 6). Altogether 86 sets of examination results were received. When students from different Institutes of Technology sat the same examination paper, the results from the Institutes were combined. For each of 15 tests, means and reliability indices (alpha coefficient and standard error of measurement) were calculated, as well as the difficulty level of individual test items. In many cases, the number that took an individual test was small, largely because the test which students took varied by Institute of Technology. Reported findings should be interpreted in light of the fact that the numbers of students for whom examination data were available were small except in the case of Electrical (Phase 6) for which data were available for 351 examinees and Carpenter/Joiner (Phase 6) for which data were available for 188 examinees. The numbers available for other examinations ranged from 48 to 80.

5. The Terms of Reference specified that procedures for assessment in the Irish system would be compared with systems in a number of other countries. Gerhard

Kohn of Human Resources Development Consulting Services, Darmstadt, Germany, was retained to carry out this aspect of the study. His report of apprenticeship systems in France, Germany, and the Netherlands is attached as an appendix to this report.

OUTLINE OF REPORT

In Section II, broad issues relating to test construction are outlined. A number of these relate to the requirement that the end result of an assessment procedure is that the information it provides can be considered valid and reliable. However, in striving to attain ideals of validity and reliability, the constraints under which tests are developed and administered have to be recognized, and some trade-offs may be required. The section includes a description of data obtained from analyses of the short-answer tests currently in use in apprenticeship examinations.

Given the prominence accorded to the 'standards-based' aspect of apprenticeship examinations, a brief commentary on standards is provided in Section III.

In Section IV, following a description of item types, research evidence relevant to answering the question 'Does choice of item matter?' is presented. Item format is considered with reference to curriculum coverage (an important issue for test validity considered in Section II), the cognitive demands and the range of cognitions elicited by an item format, the operation of construct-irrelevant factors, the effect of item format on teaching and learning, reliability, and cost.

Conclusions of the study and recommendations arising from them are presented in Section V.

II. CONSIDERATIONS IN TEST DEVELOPMENT

In this section, we outline seven considerations that the construction of tests to assess students' achievements at the end of a phase of study should take into account. Some apply to all test development; others are specific to the circumstances in which apprenticeship examinations are conducted. For each consideration we cite evidence when relevant from the FÁS (1999) specification for the setting and marking of examination questions and papers. Data from analyses of examination papers and examinees' responses are available for three of the considerations.

1. THE PURPOSE OR THE NATURE OF THE INFERENCES TO BE MADE FROM TEST SCORES SHOULD BE IDENTIFIED.

A statement of purpose provides an overall framework for test specification and for writing items (Millman & Greene, 1989). Tests often serve more than one purpose, and achieving a balance between them can be difficult. Two purposes relevant to apprenticeship examinations are considered here: to make decisions about examinee proficiency in a curriculum domain, and to make decisions about expected individual examinee performance in a future education or work environment.

(i). Tests are used to make decisions about examinee proficiency in a curriculum domain after a period of instruction.

If a student's performance on a test is to provide evidence of his/her proficiency, the items in the test should represent the objectives or skills about which one wishes to make inferences. In test development, this is achieved by specifying the boundaries and structures of the construct to be assessed which will usually be defined in terms of knowledge and skills.

There are a number of ways in which the objective of a curriculum (the attainment of which a test will seek to assess) might be characterized. One characterization involves two major objectives:

(a) students will acquire knowledge about a specific subject content area and the way knowledge is organized or structured (achievement), and

(b) students will develop cognitive processing skills which enable them to use the knowledge acquired (e.g., to solve problems, to think creatively, to evaluate the merits of competing solutions to a problem, to continue learning) (ability).

The relevance of domain content is usually established following curriculum analysis by professionals in the relevant field. Such professionals may also serve to establish the relevance of domain processes selected for an assessment. However, their work needs to be supplemented by empirical studies that provide evidence of response consistencies in performance regularities that support the view that ostensibly sampled processes are actually engaged in (Messick, 1995, 1998).

Other taxonomies of learning outcomes are available. For example, the National Qualifications Authority of Ireland (2002) document, Towards a National Framework of Qualifications, specifies three 'strands' of outcome: knowledge, know-

how and skill, and competence. It also distinguishes between declarative knowledge and procedural knowledge.

More detailed taxonomies are available, such as one which distinguishes between observation, data gathering, and recalling, and the consequences of these activities in terms of intellectual (cognitive) processes (interpreting, comparing, classifying, generalizing, inferring, analyzing, synthesizing, hypothesizing, predicting, evaluating), skills (psychomotor) (imitating, patterning, mastering, applying, improving), and attitudes and values (affective) (responding, complying, accepting, preferring, integrating) (Hannah & Michaelis, 1977).

Whatever classification is used, it is important to bear in mind that the areas are functionally related, so that each is iteratively contingent on the development of the others, that they will be intertwined in teaching, and that all areas should be assessed. Thus, in terms of our original classification, the domains for test-score inference should include information about students' acquisition of achievement (the amount and nature of knowledge an individual has acquired) and ability (the development of cognitive processing skills), both of which can involve higher level cognitive activities.

If written tests alone are used, it is not, of course, possible to directly assess students' 'competence', as defined in the Towards a National Framework of Qualifications document: 'the effective and creative demonstration and deployment of knowledge and skill in human situations' (National Qualifications Authority of Ireland, 2002, p.28). Written tests can, however, assess the knowledge students will need to display competence and they may also assess their ability to describe how they would apply that knowledge in specific situations. At the same time, it should be appreciated that no test can fully capture the range of outcomes of a complex

curriculum (Frederiksen, 1984; Haney & Madaus, 1989; Wiggins, 1989). In addition to the apprenticeship written tests, which are the subject of this study, all apprentices take practical assessments as part of the overall examination procedure; thus, the achievements that are assessed are not limited to those assessed in the written tests.

Determination of the content of a test that has as its purpose making decisions about examinees' proficiency in a curriculum domain is a complex issue, and requires consideration of a variety of questions: What specific areas of subject matter should be included? Which uses of knowledge structure should be assessed? Which cognitive planning, monitoring, and evaluation processes are relevant to the learning situation? Which subject areas, knowledge uses, and cognitive processes warrant greatest emphasis in the test, and which are less important? (Millman & Greene, 1989).

To address these issues, the FÁS (1999) directions for test development specify that the short answer tests should relate directly to, and take account of, the intended depth of treatment of the subject as indicated in curriculum objectives. They should include all aspects of that knowledge (e.g., theory/principles, maths and science), and they should ensure the broadest possible testing of the underpinning knowledge (measure the breath of knowledge) and that students cannot pass by answering a minority of questions (p. 5).

It should be noted, however, that specification of curriculum objectives is not unproblematic, and there would appear to be a divergence of view between FÁS and staff in Institutes of Technology on the nature of the curriculum, and, by implication, its objectives. FÁS has adopted what has been called an 'object-matter' focus, in which theory and practical elements should be combined in instruction. It is claimed that this approach suits craftspersons who are 'good with their hands' or less academic. Although the connection is not obvious, it is also claimed that

achievements are best assessed in short-answer tests which may include multiple-choice tests.

Staff in Institutes of Technology, on the other hand, take the view that, a 'subject-matter' focus, as is the case in their other educational programmes, is preferable to equip students with problem-solving skills that they can apply from first principles (rather than learning a way of doing something that has been demonstrated by an instructor), and that this focus requires a more extended response from students than a short-answer question can provide (though this view is not held by all IT staff). These points of view raise an issue that has been recognized in the preparation of professional workers for many fields: how students are to be taught (and assessed) in a way that will assist the integration and application in practice of 'theoretical' knowledge.

To get some indication of the level of response required of examinees in apprenticeship examinations, items in a number of examinations were categorized as requiring recall, understanding, or application. There were considerable differences between craft areas and phases in the extent to which these processes were involved, probably reflecting differences in the nature of curricula in the areas. In four Phase 4 carpenter/joiner examinations, 73.4% of items were judged to require application, 16.3% understanding, and 10.2% recall. The situation was somewhat different for the Phase 6 carpenter/joiner examinations. In analyses of four such papers, almost half the items (47.5%) were judged to require understanding, 30.0% application, and 22.5% recall.

In four Phase 6 Electrical Science examinations, all but three items were judged to require understanding; one was judged to require recall and two application. The situation was similar for four Phase 6 Electrical Craft Practice examinations, in which only two items (requiring recall) were not judged to require understanding.

Tests were also examined to determine the range of key learning categories that were assessed (Maths, Drawing, Craft Related Knowledge, Science, Personal Skills, Skills, Hazards). In Phase 4 carpenter/joiner examinations, the greatest number of items (38%) were judged to assess Related Knowledge, followed by a combination of Skills/Related Knowledge (32%) and Mathematics (26%). There was some variation from test to test in the key learning point categories that were represented. For example, in one test, eight items were judged to test Related Knowledge; in another only two did.

In four Phase 6 Electrical Science papers, Related Knowledge was again the key learning category assessed (60.5% of items), followed by Science (23.7% of items). Other categories assessed by a small number of items were Science + Mathematics, Skills, and Related Knowledge + Hazards. In the Phase 6 Electrical Craft Practice papers, Related Knowledge occupied an even more salient position; almost 9 out of 10 items were judged to assess this category.

(ii). <u>Tests are used to make decisions about expected individual examinee performance in a future education or work environment.</u>

If it is intended to use the results of a test to predict examinee performance in a future education or work environment, the appropriate source of test content will be an analysis of the cognitive requirements of that setting. These requirements may be identified through

(a) job analysis of employment settings;

(b) a consideration of cognitive indicators known or hypothesized to be positively related to criterion requirements; even if the test that is built on these considerations is verbal in nature (comprised of written pencil-and-paper

questions), the items will be considered as proxies for actual performance in a real-life situation; and

(c) an examination of the relationship between performance on a test and performance on a criterion variable. The criterion may be performance on the job or it may be more general professional development. It is not easy in practice to establish the extent to which a test successfully predicts future performance because of the difficulty in establishing the validity of the criterion variable; unreliability in measuring it; and lack of information on criterion performance for individuals who are judged unsuccessful on the initial test. In the case of FÁS apprenticeships, sufficient numbers of trainees had not completed training and entered the labour market at the time the ESF (1999) study was carried out to allow conclusions about the performance of those trained in the standards-based scheme. Lack of numbers, however, is no longer an issue as there are now almost 10,000 graduates of the standards-based system to whom National Craft Certificates have been awarded.

It is obviously important in apprenticeships that what students learn (and what is assessed) should be linked to their later performance in work. Otherwise, one would not have separate programmes of study for different crafts. Although it is much easier to link later work requirements to what occurs in on-the-job phases of apprenticeships than to what occurs in off-the-job phases, FÁS (1999) tests are designed to verify that 'a candidate has sufficient grasp of and is able to apply knowledge which is essential to ensure his/her competence at a particular task or job'. Again, the importance of assessing competence for later work performance is indicated when it is stated that examination of the theoretical and mathematical aspects of a programme should relate to industrial practice and that the design of questions should reflect this. Furthermore,

questions should elicit ability to apply theoretical and mathematical aspects and should avoid testing recall of formulae, concepts, or rules.

Later work in a particular craft, of course, is not the only future situation that is relevant to training experiences and assessment. It is also important that qualifications fit into a framework that promotes and maintains opportunities for transfer and progression (ITAC, n.d.). Indeed, progression obligations emphasized in NQA1 legislation suggest that apprenticeships should include preparatory material for progression opportunities (McDonagh, 2001). While it should be possible to demonstrate a clear relationship between an award and relevant occupational or professional standards, and while awards should be relevant to the labour market, these should function within a framework which also caters for future progression and for economic activity other than direct employment (for example, 'self-employment, business start-up, community-based and other socioeconomic activity, including personalized pathways of development') (National Qualifications Authority of Ireland, 2002, p.17). Thus, the need is identified to position traineeships within a qualifications framework to assure trainees that avenues of progression exist which permit advancement within a specific sector, or, alternatively, transfer to other sectors of employment (ESF, 1999).

These considerations also underline the role of vocational training in students' general educational development. The director of CEDEFOP has pointed out that

> vocational training is first of all a form of education. The cognitive process, the mechanisms of learning, the fundamental pedagogical principles to be applied are not all that different, whether we are dealing with general education or professional training, initial or continuing training, compulsory schooling or training that has been freely chosen and engaged in.

> …. The lines drawn between various types of education, principally of course between general education and vocational training, are largely artificial, having more to do with ideological and political considerations than educational ones.
>
> Moreover, we know how important general education is for the quality of vocational training: in today's world, adaptability in the face of uncertainty, creativity, an open mind, the capacity to learn and the ability to manage interpersonal relations have become universal requirements. (CEDEFOP INFO about Vocational Training in the European Union, 2000, 2, p.1)

A number of implications of these views for curricula, instruction, and assessment that are designed to promote life-long learning may be noted.

First, as jobs are becoming more complex, there is a need to develop the ability to transfer knowledge and skills to new situations. This involves the integration of context-specific knowledge and general skills.

Secondly, students need to learn how to learn; such learning is built on habits of systematic observation, analysis, and a questioning attitude.

Thirdly, students need to be reflective both of their own practice and their own learning.

Fourthly, students should develop thinking and problem-solving skills. This can be assisted by having students make their thinking skills explicit, which occurs when they articulate the knowledge, reasoning, or problem-solving skills they are using. Whether or not one considers the role of vocational preparation in the context of possible student progression, students will benefit from being able to organize what they are learning into 'schemas', 'maps', or 'networks' which link a variety of concepts (Attwell & Brown, 2000).

Fifthly, students should be prepared for work that is changing in nature, as the economy becomes increasingly information-based, knowledge-based, and international, and as production technologies and techniques become increasingly complex. We would expect the nature and rate of change to vary by trade.

2. THE VARIOUS COMPONENTS OF A CURRICULUM SHOULD BE APPROPRIATELY REPRESENTED IN ITEMS OF THE TEST.

This validity requirement is met by differentially allocating numbers of test items to content components on the basis of their conceptual importance and/or by applying weights in scoring that adjust for differences in item or subtest parameters or both.

This consideration may be regarded as supplementing the one relating to the use of a test to make decisions about examinee proficiency discussed above. There we saw that the tasks of an assessment procedure should be relevant to the domain being measured. There is, however, a further requirement that they should be representative of the domain. This means that all important parts of the construct domain should be appropriately represented. Lack of adequate representation, in which an assessment is narrow and fails to include important dimensions of a domain, is a threat to construct validity. Construct validity is also threatened when an assessment is too broad and contains variance that is irrelevant to the construct that is being measured (Messick, 1993, 1998). Language is particularly important in considering construct-irrelevant factors. It is clear that validity is affected when the linguistic requirements of a test interfere with an examinee's ability to demonstrate knowledge of the construct that is being assessed. We shall return to this issue when considering how item formats may contribute to construct-irrelevant variance.

To address the issue of representativeness of the components of a curriculum in an assessment, FÁS (1999) specifies that the proportion of questions on a paper should reflect the amount of time allocated to the modules on which the questions were based, and that questions should take account of the unit(s) and key learning points(s)/topic(s) within each module. It also requires examinees to answer 70% of questions correctly.

Measures of the time allocated to different areas in the Phase 4 and Phase 6 curriculum were not available for the present study. However, items in examinations were examined to determine the extent to which activities described in the curriculum were represented. These activities vary by craft and phase. Examples from Electrical Science (Phase 6) are: identifying logic gate symbols; describe the operation and layout of a PLC system. Examples for carpenter/joiner (Phase 6) are: set out hipped roofs with unequal pitches; calculate the volume, mass and density of given shapes. Analyses are available for only four examinations in Electrical Science (Phase 6) and four in Electrical Craft Practice (Phase 6). In the former, about half the activities were represented in examinations; about half were not. In the latter, less than half were represented in tests. Given that the number of curriculum activities in some crafts is considerably greater than the number of items in examinations, it would be difficult to include all activities in tests. However, some activities were represented more than once in all tests. While this presumably reflects the importance of these activities in the curriculum, it has the negative effect of reducing curriculum coverage in the tests.

In general, the activities represented in different tests were similar. It would seem that an effort was made to keep tests parallel in the activities that were assessed.

3. TESTS SHOULD BE RELIABLE, THAT IS THEY SHOULD CONSISTENTLY PLACE STUDENTS IN THE SAME RELATIVE POSITION IF THE TEST WERE GIVEN REPEATEDLY.

FÁS (1999) addresses issues of reliability in its directions for test development when it says that questions should clearly indicate what is required in the answer (e.g., list at least three, state four, calculate correctly); and that marking criteria must indicate clearly the answers required to pass questions.

Data on reliability that can be obtained from examinees' performance in the apprenticeship examinations are limited. As is normally the case in such examinations, repeated measures on the same examinees are not available. In the absence of data on individuals' performance from one occasion to another, estimates of reliability have to be based on internal analysis of performance on a single occasion. Two procedures are available. One is a measure of the internal consistency of a test (alpha coefficient), and one involves estimating the 'error' associated with each person's score (standard error of measurement). The former is considered in this section, the latter in Chapter III (Standards).

Limitations of these procedures should be acknowledged. Firstly, use of the alpha coefficient makes the assumption that a test measures a single underlying trait. It could be argued in the case of the apprenticeship examinations that a variety of achievements and abilities, as represented in levels of achievement (recall, understanding, application), in key learning points (e.g., maths, science, related knowledge), or in a variety of activities, are being assessed. However, this argument is weakened by the fact that related knowledge is the key learning point category assessed in most items, as well as by the fact that performance on the test is represented by a single score.

A problem also arises when the standard error of measurement of the mean is used when a test is being used to make 'competence' or 'mastery' decisions. In this case, it would be preferable to calculate a standard error for scores close to the score defined as indicating mastery. However, this does not seem to be a serious problem in the present situation, as for all 15 tests for which performance data were available, the mean score was close to the 'pass' (or 'mastery') score.

A problem with the use of both reliability estimates with the apprenticeship data arises from the fact that the number of items in tests and the number of students who took some of the tests were small. The maximum score in tests analysed ranged from 10 to 36; FÁS (1999) specifies a maximum of 30 items (with corresponding maximum scores of 30), generally allowing students three to four minutes to answer a question. The number of students for whom data were available ranged from 48 to 351. Small numbers of items and/or examinees would tend to be associated with low estimates of reliability.

Table 1 provides data for three separate examinations in Brickwork, Phase 6; two examinations in Carpentry, Phase 4 and one in Carpentry, Phase 6; two in Electrical, Phase 4 and three in Electrical, Phase 6; and four in Fitter, Phase 6. When students in different ITs took the same examination, data from the centres were combined.

There was great variability in the alpha coefficient value for the 15 tests, which ranged from .20 to .94, with a median value of .62 (Table 1). In general, the higher the maximum score on a test, the greater the reliability, though there are exceptions. While tests with a maximum score of 36 are among those with the highest reliability values, the test with the highest value had a maximum score of 20. That more than number of items is involved in determining reliability can be seen from the fact that the alpha value of tests with maximum scores of 10 ranged from extremely low to moderately low. In

general, fitter and electrical tests had higher alpha values than carpentry and brickwork tests. This suggests that items in the former tests are more homogeneous than in the latter and are measuring a common latent attribute.

Table 1

Means, Standard Errors, and Coefficient Alphas for 15 Apprenticeship Tests*

Test	N	Max Score	M	SE	α
Brickwork 6	48	16	12.56	1.306	.454
Brickwork 6	48	16	12.69	1.463	.477
Brickwork 6	48	16	13.06	1.376	.577
Carpentry 4	49	10	7.78	1.217	.198
Carpentry 4	49	10	8.65	0.967	.537
Carpentry 6	188	10	7.82	1.232	.379
Electrical 4	62	20	12.44	1.593	.935
Electrical 4	62	20	16.03	1.551	.760
Electrical 6	351	20	15.79	1.664	.616
Electrical 6	355	20	14.22	1.843	.751
Electrical 6	355	20	13.84	1.896	.716
Fitter 6	80	15	12.05	1.405	.613
Fitter 6	80	15	11.81	1.425	.657
Fitter 6	81	36	28.51	2.221	.806
Fitter 6	81	36	26.49	2.370	.822

* Data from some ITs diverge from dichotomous scoring (0, 1), allowing for partial credits (0.5).

Also relevant to a consideration of reliability is whether a students' performance is affected by the version of a test he/she takes. Examinations of the data in Table 1 suggest that that may be the case. It will be noted that there are considerable differences in the mean scores of examinees on some tests that were designed to be parallel. An extreme case is to be found in the Electrical Phase 4 test, on which the mean score on one was 62.2%, while on another it was 80.15%. This difference could mean that the achievements of students taught in one course were lower than the achievements of students taught in another course. It could also mean that one group of examinees was

set an examination that was more difficult than that set for another group, raising concern about the equivalence of tasks designed to be representative of the construct domain, and about generalizability of the performance of examinees on the assessments to the broader construct domain.

FÁS (1999) specified that questions should, as far as possible, be of equal level of difficulty, though it does not specify the level, or how it might be determined. Inspection of the items in tests suggests that some items require more complex procedures than others and thus are unlikely to be of equal difficulty. This view is confirmed by statistical analysis of the performance of students on 15 tests. In general, the percentage of examinees that got individual items correct varied from the 50s to the 90s. Thus, there was a good deal of variation in difficulty level using the criterion of percentage (or proportion) getting items correct. In a few instances, there were 'very difficult' items for which the percentage of examinees who got the items correct was as low as 25 and 31.

4. THE EFFECTS OF TESTS ON TEACHING AND LEARNING SHOULD BE TAKEN INTO ACCOUNT IN THEIR DESIGN.

There is ample evidence from many countries that when sanctions are attached to test performance, teachers and students will look to tests for clues about what is important to teach, with the result that the content and format of past tests will impact on teaching and learning (Kellaghan, Madaus, & Raczek, 1996; Madaus & Kellaghan, 1992). Many commentators perceive this as a positive aspect of testing. If the objectives and skills to be measured are carefully chosen, and if the test truly measures them, the goals of instruction will become explicit and well-defined targets for teachers and students on which they can focus their efforts. Furthermore, the tests

will provide students and teachers with standards of expected achievement (see, e.g., Eisemon, 1990; Frederiksen, 1984).

However, positive effects have to be balanced by negative, albeit unintended, effects of tests on teaching and learning, several of which have been documented. First, the fact that teachers and students attend in class and in study to topics that are likely to appear in tests or examinations will result in a narrowing of the curriculum and the exclusion of curriculum areas (both cognitive and non-cognitive) that are not examined, which in turn will result in a restriction in student achievements. One would expect that when the focus in teaching is on the content or format of a specific test, which represents only a small sample of the achievement domain, teaching will become less representative of the domain, and this will be reflected in students' achievements (Koretz, 1995). The negative impact will be more pronounced if the knowledge and skills required to do well on a test are for the most part ones relating to the recall or recognition of factual information rather than the ability to synthesize data or apply principles to new situations.

Second, tests to which high stakes are attached are likely to result in considerable effort being invested in test preparation activities. This is evidenced, not only in the use of a wide range of test-preparation practices, ranging from 'test-wiseness' to actually teaching test items, it also can be seen in the more general activities of teaching. Thus, one would expect teaching methods to vary depending on whether the examination requires students to select a correct answer (as in multiple-choice items), to construct a response in a short answer, or to construct a more extended response. The format can narrow the focus of instruction, study, and learning to the detriment of other skills. We will consider some evidence relating to this in the section 'Does Choice of Item Type Matter?'

Third, high stakes tests affect the nature of students' learning – their goals, learning strategies, involvement in learning tasks, and attitudes to learning, in particular attitudes towards improving their competence. While one would hope that students would develop self-regulating learning and problem-solving strategies and exhibit a preference for challenging work and risk-taking, high stakes examinations tend to promote the use of strategies that are superficial or short-term, such as memorizing and rehearsing, and the avoidance of challenging tasks and risk-taking (Kellaghan, Madaus, & Raczek, 1996).

It is very difficult to avoid these consequences when important sanctions are attached to performance on a test. The primary concern of the test developer is that any negative impact should not derive from any source of test invalidity, such as construct underrepresentation or the presence of construct-irrelevant factors (Messick, 1998). As we saw in considering views about apprenticeship examinations, the facts that the examinations were considered to provide inadequate curriculum coverage and that the knowledge assessed was superficial were regarded as creating an undesirable backwash on teaching and learning in which more advanced forms of knowledge and communication skills received inadequate attention.

A high level of competence in test design is required to minimize possible negative consequences and to ensure that the positive effects of examinations are emphasized; that the objectives and skills to be measured are carefully chosen to represent the domain of achievement and ability; that the test truly measures them; that the goals of instruction are made explicit and are translated into well-defined targets for teachers and students to focus their efforts; and that the examinations provide students and teachers with standards of expected achievement.

5. ECONOMY AND FEASIBILITY SHOULD BE TAKEN INTO ACCOUNT IN TEST DESIGN.

While performance assessments or oral examinations might recommend themselves as a superior way of assessing students' proficiencies, constraints of time and personnel may mean that they are not feasible.

6. ACCOUNT SHOULD BE TAKEN OF THE FACT THAT APPRENTICESHIP EXAMINATIONS ARE ADMINISTERED FREQUENTLY (EVERY TERM) AND IN SEVERAL LOCATIONS.

Security problems arise if the same test is used in several locations and/or on different occasions. Comparability problems arise if different tests are used. Further, the onus in developing a large number of tests is considerable.

7. ACCOUNT SHOULD BE TAKEN OF HOW TO FACILITATE STUDENTS WHO ARE UNSUCCESSFUL.

A substantial number of students do not pass the Phase 4 and 6 examinations, though the number is not high in the context of other post-secondary institutions. According to the ESF (1999) evaluation report of apprenticeship and traineeship, 15% (or almost 1 in 6) Phase 6 and former apprentices had been obliged to repeat Phase 4 assessments. Analyses of data for the present study indicate that 18.5% of candidates failed a Phase 4 examination (range on four examinations: 4.1% to 35.5%) and 20.5% a Phase 6 examination (range on six examinations: 8.2% to 33.7%). (Pass-fail analyses were not carried out for examinations for Fitter since a pass-fail decision was not made on the basis of performance on the tests analysed for this study.)

An apprentice who is unsuccessful in an assessment is given the opportunity to repeat it twice. It would appear that provision for helping students prepare for repeats is limited (ESF, 1999).

Although FÁS provided a procedures manual for apprentices wishing to appeal off-the-job assessments in which the Services to Business Manager of the region in which the apprentice is employed and the Manager of Certification and Standards in FÁS are involved, the ESF (1999) report concluded that 'no effective appeals procedure exists in relation to Phases Four and Six' (p. 101).

The failure rates for off-the-job phases contrast with the situation in on-the-job phases. Concern has been expressed about assessment during these latter phases. Employers are required to certify that an apprentice is capable of doing specified tasks; yet there appear to be no failures or repeats. One difficulty that has been identified is that employers may not engage in the type of work for which the apprentice must be assessed. There are also indications that some employers simply presume that an apprentice can perform the tasks to a satisfactory standard and award a result accordingly (O'Connor, 2000).

III. STANDARDS

Several countries across the globe have in recent years adopted a 'standards-based' approach in their vocational education and training systems (e.g., Australia, England, Scotland, New Zealand) (Gunning, 2000). In some countries, the term is used almost synonymously with the terms 'competence' and 'outcome'. Whatever the precise term, all systems specify that to obtain a qualification, an individual must demonstrate that he(she) has acquired predetermined levels of knowledge, skill, and understanding appropriate to employment, progression, or self-development.

Various reasons have been given for adopting a standards-based approach: the need to improve international competitiveness; the desire to relate vocational education and training more closely to employment needs; the proliferation of qualifications and the lack of relationships between them; and the need for national portability (Gunning, 2000). Perceived benefits of the system include the fact that clear targets are provided for learners and instructors and the facility it provides to respond rapidly to changing economic and employment needs (Gunning, 2000).

A standards-based system has also been adopted in Ireland. As elsewhere, the system may be contrasted with the system that preceded it, which only required an apprentice to serve a specified period of time to qualify as a craftsperson. While there was provision for attendance at a one-year off-the-job course or at three 'block release' courses, and to sit for Junior and Senior Trade Examinations, not all

apprentices were able to do this. No mandatory assessment of competence was required of apprentices that had served the required period of time (O'Connor, 2000).

To address this situation, the 1986 White Paper on Manpower Policy set the objective, among others, of developing an apprenticeship system that would be based on standards achieved rather than time served. In the Programme for Economic and Social Progress (1991), government and social parties agreed to the introduction of the new system, following which new curricula and mandatory assessments were developed and introduced. According to O'Connor (2000), the new system 'ensures that every apprentice attains a predetermined level of competence.'

Reference is made in FÁS (1999) documentation to 'criterion-referencing' and 'a predetermined level of competence.' It says that assessment 'measures a trainee's performance against external criteria and aims at a level of competence which is predetermined and based on prevailing social and economic standards.' However, it is not clear how 'prevailing social and economic standards' are operationalized in terms of the specifications for specific assessments or in determining levels of competence.

Standard setting involves the development and adoption of a mark scale and identification of points on the scale with particular performance standards, with the intention of enhancing the inferences that are warranted from test scores (Wiliam, 1996). The objective is to map scores on an assessment task to 'performance levels.' That is, particular types of knowledge and skills are matched with scores on a task to provide a picture of what students classified at varying levels of proficiency know and can do. It is now generally accepted that standard setting is a complex, difficult, and to some extent arbitrary procedure. It usually comprises several components: the identification and selection of stakeholders who will act as panellists, the training of panellists, choice of a standard-setting method (of which there are many), reference to

empirical data on student performance, and the review and revision of judgments made by panellists (see Cizek, 2001). No evidence was obtained in the present investigation that these procedures were followed. Thus, pass and credit marks seem to have been determined without reference to the knowledge, skills, and abilities of examinees that performance at a particular level represented.

According to FÁS (1999) specifications, a pass mark is achieved on the basis of 70% of questions answered correctly and a credit on the basis of 85% answered correctly, though these figures do not always apply in practice. In tests for which the total score is 10 (e.g., Brick/Stonelayer Phase 4; Carpenter/Joiner, Phase 4), while the pass mark is 70%, the credit mark is 90%. In tests for which the total score is 16 (e.g., Brick/Stonelayer, Phase 6), the credit mark is 81 percent. The high pass mark is not associated with any identified knowledge or skills, but is prescribed so that 'a candidate cannot pass by virtue of answering a minority of heavily weighted questions which deal, perhaps, with a limited area of the required underpinning knowledge.' There has been criticism of the high mark required to pass from staff in Institutes of Technology who are more accustomed to arriving at a pass decision on the basis of a lower percentage.

The choice of percentage correct to determine levels of competence in the apprenticeship tests gives rise to a number of issues. First, the choice of pass and credit marks does not take into account the fact that examinees' scores are dependent on, and so can vary, with the level of difficulty of items. Kane (1994) points out that

> the tradition of requiring 70% correct on tests seems especially arbitrary, because we know that, for any group of examinees, we can probably make the items easy enough so that everyone gets more than 70% correct or difficult enough so that nobody gets more than 70% correct. (p. 426)

Second, although all items are accorded the same value, in apprenticeship tests they vary in the complexity of the responses they require and, using the criterion of percentages of examinees who get an item correct, are not equal in difficulty. In theory, an examinee who got more 'difficult' items correct and less 'easy' items incorrect could obtain a lower total mark than an examinee who got more 'easy' items correct and less 'difficult' items incorrect.

Third, no justification is provided for choosing 70% as the cut-score, or how a score of this magnitude guarantees that an examinee exhibited competence. Neither is evidence provided that students scoring below 70% lack competence. Indeed, in the view of staff in Institutes of Technology, some students, who on the basis of the examination fail, have the competence to continue with their course.

Finally, an issue relating to the two standards (pass, credit) specified for test performance in the examinations arises from the standard error of measurement of examinees' test performance. For all tests which had less than a possible total score of 35, the credit score fell within 2 standard errors of the pass score. Thus, whether or not an individual's score fell at the pass or credit level could be due to measurement error (95% level of confidence).

A problem with the apprenticeship examinations is that a direct relationship between passing scores, performance standards, and levels of competence seems to be assumed. However, to interpret a passing score in terms of a performance standard, it is necessary to demonstrate that it represents a level of skills or achievement in some area, which in turn is taken to represent a desired level of competence. Furthermore, the performance standard should be appropriate in that it is considered just high enough to meet the purposes of the decision process that is based on it (Kane, 1994). The critical issue in standards then is not a particular score on a test, but the fact that they represent a

construct that is shared in a community of interpreters (Wiliam, 1996). As already noted, specification of that concept follows a complex process involving selection and training of a group of stakeholders, choice of a standard-setting method, and the review of judgments in the light of empirical data.

IV. ITEM TYPES

Having considered general factors that are relevant in the development of tests and the organization of an assessment system (in Section II), we now turn to the specific issue of item format.

The issue of item format arises because it is possible to distinguish between the construct domain of an assessment (its substance, content, boundaries and structures, and interrelationships among its elements) and how it is measured (its form or format). In general terms, measurement methods can be categorized as pencil-and-paper exercises or as performance or simulation exercises. As this review does not extend to practical examinations, our concern is with pencil-and-paper items. These can be further sub-divided into

selected response items and

constructed supply-type items.

The latter can be further sub-divided into

short-answer supply items and

essay-type items.

ITEM TYPES

Selected Response Items

In selected (or fixed) response items, the examinee is given the correct answer/solution to a question/problem as well as alternative (incorrect) answers/solutions, and is instructed to select the correct answer/solution from the options. The items include multiple-choice items (in which four or five response options are usually provided), alternate choice (true–false, etc) items, and matching items. They are often called 'objective' because items can be scored with significant certainty; since the 'correct' or 'best' answer is predetermined, the application of a key or scoring guide is simply a matter of comparing students' responses to this answer (Rodriguez, 2002).

Multiple-choice items (and many short-answer constructed response items) are based on a number of assumptions.

First, complex skills can be decomposed and isolated from their applied contexts.

Second, items in the test are based on a limited range of well-structured algorithmic problems.

Third, the scoring scheme is based on a view of learning in which skills and knowledge can be incrementally added (Bennett, 1993).

Fourth, multiple-choice items are constrained in the kinds of thinking and higher-order cognitive processing that they can asses, since they tend to emphasize recall and convergent thinking and to de-emphasize synthesis and divergent thinking (Messick, 1987; 1998; Wainer & Thissen, 1993).

These assumptions conflict with the increasing emphasis in recent years on 'situated cognition' which is regarded as context-specific in nature requiring a domain-specific

knowledge base. Since problems in the real world are often unique and poorly structured, they require skills that are highly integrated and tied to conditions of application.

Construct (Supply-Type or Free Response) Items

Constructed (supply-type or free response) items require the examinee to generate or construct a response, and normally more than one correct or unique 'correct' answer is possible; even if it is not, the fact that examinees provide answers in their own words means that human judgment is required to decide whether or not a response is acceptable. For that reason, such items are often referred to as 'subjective'. Although constructed items are frequently categorized as short-answer supply items and essay-type items, in fact they range from well-structured decontextualized tasks to ones requiring processes involved in solving deeply situated ill-structured problems (Snow, 1993), and include short-answer, reordering/rearranging, substitution/correction, simple completion/close procedures, computation, complex completions, problem exercises, and restricted and extended-response essay items (Messick, 1998; Rodriguez, 2002). Thus, they represent a graduated continuum of test formats, which can vary greatly and cover a wide range of tasks from relatively minor variations of the tasks involved in multiple-choice items to extended projects and complex performances. If we limit the tasks to ones involving pencil and paper, at one extreme an examinee may be required to respond by writing a word or short sentence; at the other extreme, by writing an extended essay.

These extremes represent a wide range in the complexity of the manifest responses of examinees, as well as in the knowledge of structures, processing strategies, and self-regulating functions that they demand (Martinez, 1999). At one extreme, many completion and short-answer items will not differ greatly from

multiple-choice items; responses are constructed only in the sense that they require recall; and they do not yield a scorable record of an extended process or product. Such items differ considerably from ones at the other extreme that require the mental assembly of a new conceptual product or the provision of a response that has qualities of novelty and complexity. Openness with respect to response possibilities allows examinees to exhibit cognitive structures and skills that are difficult to assess within the limits of the multiple-choice format (e.g., shaping or restructuring a problem; developing alternative strategies to solving a problem; facility in using interconnected rules rather than fragmentary pieces of information).

Short-Answer Supply Items. While the short-answer supply item had almost disappeared by the 1960s (Wesman, 1971), there has been a revival in its use in recent years. Tests composed of short-answer constructed response (or supply) items present the examinee with many items with minimal, but varied, contexts intended in the aggregate to measure 'achievement' evinced across multiple learning situations. Alternative solutions are not presented, and so a response must be generated rather than chosen from a list of options (Cronbach, 1984; Osterlind, 1998). Sometimes, short-answer supply items are considered extensions of the multiple-choice format. Stems for the two types of item may be equivalent; the essential distinction resides in the examinee response of recognition or recall.

It should be noted that most of the items in apprenticeship examinations are not short-answer as that term is normally used (in which the examinee is required to respond by writing a word, phrase, or short sentence). Some are of this nature, but many require operations of varying complexity to arrive at a solution. Some involve mathematical calculations. About half the questions in the Phase 6 Carpenter Joiner examinations that were reviewed required the examinee to draw or sketch. Thus, use

of the term ‘short-answer’ to describe the apprenticeship examinations may be regarded as a misnomer. Certainly, some of the items require complex completions by the examinee. The examinations are probably more correctly categorized as requiring more extended constructed free response items which, however, do not involve the degree of extension involved in traditional essay-type examinations.

Essay-type Items. The extended essay is probably the test format that has been longest in use (Bloom, Madaus, & Hastings, 1981; Coffman, 1971). An essay requires an examinee to select appropriate information from his/her knowledge and background and to explain, discuss, compare, or analyze phenomena or topics. Examinee responses are composed by the student and usually take the form of a series of sentences.

An advantage of the essay-type item is that it can tap high levels of reasoning such as are required in inference, in the organization of ideas, and in making comparisons and contrasts. The tasks set in an essay can be complex requiring, for example, an identification and description of a problem and how one may address it. Thus, in its ideal form at any rate, the extended essay provides opportunities to tap the complex structuring of multiple basic and higher-order skills and knowledge, which may be embedded in rich problem contexts that allow the examinee to engage in and display extended or demanding forms of reasoning and judgment.

DOES CHOICE OF ITEM TYPE MATTER?

It has been argued that ‘the form of the task can be important as the substance’ (Cronbach, 1984). Furthermore, substance and form are not independent: substance dictates form, and form in turn affects substance (Millman & Greene, 1989).

Despite these views, it should be acknowledged at the outset that empirical evidence relating to item type is limited. Most research has compared the multiple-choice format with simpler forms of the constructed response format, and even that research has been constrained by the particular design features it adopted (see Bridgeman, 1992; Martinez, 1999; Thissen, Wainer, & Wang, 1994; Traub, 1993). For example, most studies were designed specifically to assess the proficiency dimensions that are tapped in multiple-choice tests. Even in studies that were not designed in this way, constructed response items were composed to be direct counterparts of multiple-choice questions. However, if the multiple-choice format is in fact most likely to assess lower level skills, transforming items to a constructed format would not be likely to affect what is being assessed. A further serious disadvantage of studies of item types is that the tasks most likely to show differences (e.g., complex performance tasks) have received little attention (Bennett, 1993; Snow, 1993).

The research evidence indicates that, in general, students who take two tests using different item types will achieve a similar rank order on both tasks. However, the extent to which this is so will depend on how similar the tests are. The correlations between performances are larger where comparisons involve tests that were designed to be construct equivalent and used a similar item format than when they were written to be construct different and employed different item formats. Relationships are weakest when the comparison involves performance on a multiple-choice test (with its greater sampling of the content domain) and performance on an extended constructed response test (with its greater depth of process) (Rodriguez, 2002).

Does Item Format Affect Curriculum Coverage in a Test?

As already noted, an important feature of the content validity of a test is that it should adequately represent the objectives of a curriculum. There are clearly differences associated with item format in the nature of the content and processes that are measured (Frederiksen, 1984). Multiple-choice items provide broad domain sampling and, of all item types, offer the opportunity of obtaining the largest sample of content per unit of time testing. In the discrete constructed response format involving short responses, such as written words, numbers, or phrases, sampling of content per unit of time can also be high. In more extended constructed responses (e.g., written essays), the sampling of content per unit of time decreases. Given time constraints, it is usually possible for an examinee to address only a limited number of topics. However, while the essay may not be an effective means of assessing a wide range of curriculum content, it may have an important role in sampling cognitive skills (e.g., students' ability to organize or apply knowledge). It is also considered to replicate more faithfully the tasks examinees face in academic and work settings (Lukhele, Thissen, & Wainer, 1994).

Does Item Format Affect the Typical Cognitive Demands and the Range of Cognitions They Elicit?

Even if tests using different formats correlate as high as their respective reliabilities may allow, yielding similar rank ordering of examinees, this cannot be taken as necessarily implying psychological equivalence. Despite their psychometric equivalence, based on correlational and covariance-structure methods and Item Response Theory (see Lukhele et al, 1994; Wainer, Wang, & Thissen, 1994), tests using different formats may call into play distinct reasoning processes and knowledge,

and so cannot be regarded as measuring the same attributes (Bennett, 1993; Snow, 1993). A number of studies support the view that test formats (particularly when multiple-choice and free-response formats are compared) appear to measure different abilities (Cronbach, 1941; Thissen, Wainer, & Wang, 1994; Traub & Fisher, 1977; Ward, 1982; Ward, Frederiksen, & Carlson, 1980).

While certain aspects of constructs that appear in curricula would seem to lie outside the range of item formats (e.g., problem finding, aspects of performance that are tacitly employed such as metacognition and self-regulating abilities), some item formats are less successful than others in eliciting some aspects of cognitive activity. Thus, a change in format could change the nature of the construct that is assessed from, for example, recognition to recall, or from factual knowledge to higher order thinking skills (Bennett, 1993).

Tests composed of multiple-choice items are often regarded as measuring recall and superficial and lower-level cognitive processes. While this is probably true of most tests, multiple-choice items can be written to elicit more complex cognitive performances (involving, for example, comprehension, prediction, analysis, evaluation, and problem-solving) (Haladyna, 1994, 1997). Indeed, it has been argued that what is measured by multiple-choice items is more a function of their content than of their form (Ebel, 1970). However, writing items that will assess complex aspects of cognition is very difficult, and some commentators believe that the cognitive range of multiple-choice tests is limited by their format. A similar conclusion can be reached about short-answer constructed responses.

Of all item types, multiple-choice items involve the lowest degree of response construction. Short-answer formats involve more complex construction. These can vary from provision of a single word or short sentence to drawing a graph from given

data, giving reasons (e.g., why condensation forms on windows), producing a drawing, and writing a geometric proof.

The short-answer supply-type item is regarded as superior to the multiple-choice item insofar as a higher level of competence is indicated by producing a correct response than by identifying it from a list of options. Certainly, an examinee that knows and can produce the correct response would be expected to be able to recognize it, while it does not follow that an examinee that can recognize a correct response would be able to produce it.

However, there are a number of disadvantages associated with the short-answer free response item (Wesman, 1971). First, it is most appropriate when the answer is a single word, name, symbol, or formula. It is less suitable for statements of generalizations, principles, etc. Hence, the focus in the short-answer item is usually on memory for facts. Second, it is extremely difficult to phrase short-answer items so that the same correct answer will be given by all who know the answer. The problem arises because words and phrases have many synonyms and near-equivalents. Third, there are usually degrees of appropriateness to a precise or accurate answer, with the result that it is often difficult to draw a line between marginally acceptable and marginally unacceptable answers.

The cognitive range that can be tapped by items extends as one proceeds from multiple-choice to discrete constructed responses to extended performance constructed responses. So does structural fidelity which can be regarded as a continuum of the distance between a criterion measure and an unobservable criterion, which can be assessed through cognitive analysis, content analysis, or logical analysis (Haladyna, 1998). Longer essays require the greatest degree of construction.

It should be noted that the cognitive demands of an item depend not only on the question or task, but also on the prior experience of an examinee. For example, if a topic has been well covered in a course, all that might be required would be recall on the part of the student, while for a student without this prior experience, the task might require application of knowledge.

Do Construct-Irrelevant Factors Operate Differently with Different Item Formats?

An examinee's performance on a test is susceptible to contamination by a number of factors. The factors contribute to variance in examinees' scores, even though they are irrelevant from the point of view of the construct that is being measured. The question of interest in the present context is: do construct-irrelevant factors contribute differently to variation in scores for different types of item format?

One factor that may affect examinees' scores is their proficiency in format-specific strategies. The multiple-choice format is regarded as being particularly vulnerable to examinees' test-taking strategies. These can exhibit themselves in examinees' ability to capitalize on information embedded in response options or in the ability to use response elimination, in which the number of options from which to choose is reduced by eliminating implausible ones, thus increasing the probability of responding correctly by chance.

Constructed response formats are not immune from contamination by examinees' use of test-taking strategies. An examinee may, for example, apply a prepared template to an essay or write in a way that capitalizes on what he/she knows, while at the same time concealing gaps in knowledge or giving the impression that such gaps do not exist.

The role of language as a construct-irrelevant factor is particularly important in the context of extended constructed responses. Essays measure writing skills as well as students' knowledge of, and ability to apply, curriculum content. They require verbal

ability to read, comprehend, process, and produce, regardless of the content area being assessed. This can give rise to a validity problem, in, for example, a mathematics or science test, when examinees' content-specific knowledge and abilities may be confounded with their verbal skills. It is also problematic when it cannot be assumed that the literacy level of all examinees is similar. The issue that needs to be addressed is whether the verbal abilities which an examinee needs to respond reflect the construct domain being measured or reflect a construct-irrelevant source of variation. In addressing the issue, it may be worthwhile bearing in mind that 47% of apprentices have a Junior Certificate and 53% a Leaving Certificate.

In some situations, verbal ability forms an integral part of the construct being measured. For example, if a task requires the examinee to read and/or write instructions or directions, ability in reading and in comprehending material is obviously required, and its influence on examinees' achievement scores will reflect a valid source of variation.

In other cases, verbal ability, though not an integral part of the construct being measured, is correlated with it. One can assume that in any curriculum area, verbal skills will be involved as, for example, in knowledge and understanding of facts, concepts, principles, and procedures, as well as in the ability to apply concepts to the analysis of complex tasks and problems.

The influence of verbal ability on the validity of an assessment will depend on how curriculum outcomes are defined. A case can be made for minimizing its role in articulating these outcomes, and in their assessment, in such subjects as mathematics, science, and drawing. At the very least, care should be taken to define outcomes in such content areas and to set examinations with reference to the verbal abilities

(including their reading ability) of the students for whom the curriculum and assessment are designed (Ryan & deMark, 2002).

Do Item Formats Differentially Affect Teaching and Learning?

As noted in Section II, when important consequences are attached to test performance, teachers and students use their expectations concerning tests to guide their teaching and learning. Furthermore, much time and effort may be invested in actual preparation to take the test. The effect of all this is that teaching processes and learning outcomes are distorted, leading to the neglect of topics and cognitive processes that are not tested (Madaus & Kellaghan, 1992).

The effects of test item format on teaching and learning leading up to a test has been a matter of much speculation and some research (Martinez, 1999). Kinney and Eurich (1932) expressed the view that

> the use of the subjective (i.e., open-ended) examination stimulates the pupil to study in order to acquire an organized body of information, and observe the relationships and implications of the facts thus learned. (p. 543)

Other commentators have claimed that essay-type examinations encourage students to learn how to manage their ideas and express them effectively (Bloom et al, 1981). Meyer (1934), who a long time ago reported that the form of an examination expected by students determined the nature of their study, also claimed to have found that the highest levels of achievement on all types of examination followed study in anticipation of an essay examination. While this finding has not always been replicated (French, 1956; Sax & Collet, 1968; Vallence, 1947), the available evidence seems to favour the view that response formats affect anticipatory learning, with richer outcomes being associated with extended constructed response

formats. If it is known that students will be required to demonstrate competence in problem-solving, graphing, essay organization, and writing, these skills will be emphasized in classrooms (Lukhele et al, 1994). For example, students will pay more attention to the semantic organization of learning material when preparing for a test that requires recall rather than recognition, and students who expect a constructed response test will attend more to the structure of curriculum content in their study, compared to students who expect a multiple-choice test. Again, students who are studying for essay tests tend to search for main points and to strengthen their grasp of subject matter at a global level (Martinez, 1999).

Does the Reliability of a Test Vary with Item Format?

Tests with multiple-choice items and tests with discrete constructed response items have the highest reliability. In fact, reliability can be greater in some short-answer constructed-response scores than in multiple-choice scores, perhaps because in the former, guessing is not possible and clues are not available in options (Rodriguez, 2002). Short-answer free response tests, however, can vary greatly in their reliability. Much depends on the quality of items and of scoring keys. A high level of unreliability in scoring is often associated with essays. Since no single response or pattern of responses can be listed as correct, the accuracy and quality of an examinee's work can only be judged subjectively by one that is skilled in the subject and in examining. FÁS (1999) has attempted to address this issue in its specifications for tests which say that model answers for calculation questions must indicate whether the correct answer is required or whether the correct computational process is sufficient; and that model answers for definition or description questions

should indicate the key elements required (with allowance for answers in which elements are expressed in words that differ from the model answer).

Are There Differences in Cost Associated with Item Formats?

Multiple-choice tests are expensive to develop but are inexpensive to administer and score. At the other extreme, essay-type tests are not expensive to develop, but are expensive to score. Discrete constructed response items occupy an intermediate position. When the number of examinees is small, however, cost will not be an important issue in scoring.

V. CONCLUSION

GUIDELINES FOR TEST CONSTRUCTION

We conclude by enumerating nine guidelines which should be kept in mind when tests are constructed and administered for apprenticeships, as indeed would be the case in tests administered in any similar situation. Following that, we briefly consider the implications of some of these guidelines for apprenticeship examinations before listing a number of recommendations for the construction of apprenticeship examinations.

1. Tests should be designed to provide evidence about examinee proficiency in the content of curricula and in the way knowledge is organized or structured in specific curriculum areas, as well as about the range of cognitive skills that enable individuals to use the acquired knowledge.
2. The items on a test should cover all the important components of a curriculum, with those that are considered more important being accorded greater weight.
3. An assessment should seek evidence that would be useful in informing decisions regarding the future education and career environments of examinees.
4. Tests should be constructed to consistently place students in the same relative position if given repeatedly.
5. An examinee should be awarded the same grade whatever version of a test he (or she) takes.

6. The range of scores that performance on a test yields should be sufficient to allow accurate discrimination between examinees judged to be performing at different levels of proficiency.
7. If tests are presented as standards-based, the procedures that were used to match scores to performance levels (in terms of the particular types of knowledge and skills which students classified at varying levels of proficiency have acquired) should be described.
8. Since the content and format of tests will inevitably impact on teaching and learning when sanctions are attached to performance, care should be taken to ensure that any negative impact should not derive from any source of test invalidity, such as construct underrepresentation or the presence of construct-irrelevant factors.
9. In constructing a test, consideration should be given to the many assessment formats that are available, and the appropriateness of each to serve the purpose of an assessment, together with its strengths and weaknesses. Since no single format can be considered appropriate for all educational purposes (Rodriguez, 2002; Wainer & Thissen, 1993), no format should be preferred over all others in all respects and for all purposes. Because of this, there has been considerable interest in recent years in developing tests that use more than one type of item (particularly in the United States where free-response items are being added to multiple-choice items). Thus, the decision that will normally be made is not either/or, but what mixture of item formats will yield the best possible effect in combination. Judgment of appropriateness will depend on a variety of factors, ranging from a consideration of the knowledge and skills that are to be assessed to

issues of economy and feasibility. Key trade-offs that may be involved also need to be taken into account (Martinez, 1999).

IMPLICATIONS OF THE GUIDELINES FOR APPRENTICESHIP EXAMINATIONS

Analysis of apprenticeship tests revealed variation in the extent to which tests in different crafts assess students' ability to recall, understand, and apply knowledge. The small number of items that depended on recall in all tests indicates that the tests are assessing knowledge at a level that is higher than the level that is usually associated with short-answer items, and runs counter to a general perception that the focus is on recall to the neglect of higher-order knowledge. The fact that many items in tests are considered to assess examinees' understanding supports the view that test items require a higher level of construction on the part of examinees than is normally required in short-answer tests.

Content coverage is difficult to achieve if a test contains only a small number of items, whatever item format is used. While this is a particular problem for essay-type items, even in the case of the apprenticeship examinations that were analysed for the present study, the small number of items in tests precluded an assessment of a wide range of curriculum content. Improved content coverage could be achieved by using short-answer questions that require a response that could be provided in a shorter period of time. However, as many of these items would most likely require recall rather than higher levels of understanding or application, they would need to be supplemented by items that require more elaborate constructions and thought processes (e.g., students' ability to organize or apply knowledge).

It is generally agreed that attention, encoding, and working memory processes in learning differ as a function of expected test format (d'Ydewalle, Swerts, &

DeCorte, 1983). Furthermore, the range of cognitive functions (involving knowledge, procedures, schemas, and self-regulatory skills) that is elicited by a test increases with the complexity of the response that is required. All construct-valid functions measured by multiple-choice items can be tapped by constructed response items, but the reverse is not true. In particular, the constructed response format can assess complex performance and divergent production. While simpler and short constructed items require only the product of thought processes, written essays can provide a rich base of examinee cognition, including strategy selection, cognitive structuring, the organization of items, paraphrasing, elaboration, and reasoning processes. Furthermore, both the observable performance of examinees and the recorded trace of that performance are potentially richer in constructed response items (Martinez, 1999).

While extended constructed responses have advantages, they also have disadvantages associated with content coverage, reliability, and construct-irrelevant variance associated with examinees' verbal abilities. The use of more conventional essay-type items would thus require great care to ensure that the reliability of the assessments was not reduced. As we saw in considering item types, the highest levels of reliability are achieved with multiple-choice and short-answer items, the lowest with items that require more extended constructions by examinees. This is an obvious problem if one is interested in assessments that require complex expressions that allow insights into knowledge structures, processing functions, abilities, and dispositions that mark the novice-expert axis. However, important though reliability is, it should not take precedence over validity.

As well as addressing the technical issues raised in this report, a review of the apprenticeship assessment system should be sensitive to differences between FÁS and the Institutes of Technology in their values and priorities. For example, in considering

the relevance of future criterion settings to apprenticeship examinations for off-the-job curricula, there would appear to be a difference between the views of FÁS and of staff in Institutes of Technology about the criterion that is most relevant. While the former emphasize competence in a particular defined task or job, the latter emphasize general educational development as a preparation for a dynamically changing work environment and also place greater emphasis on the extent to which the competencies that students acquire fit into a framework that promotes and maintains opportunities for transfer and progression in accordance with the Qualifications Act. It was not possible in the present study to reach a conclusion about the appropriateness of assessment procedures in this context, except to say that curricula and assessments should take into account the need to prepare students for a changing work environment that will require problem-solving and adaptation skills.

Differences between FÁS and the Institutes of Technology in their perceptions of what apprentices are being prepared for may flow over into their perceptions of what is most appropriate in assessment. In the 'training' approach which seems to be favoured by FÁS, the work of a 'craftsperson' normally leads to actions and products that are well defined, and it is usually not too difficult to judge whether or not a product reaches identifiable standards. In the 'educational' approach which seems to be favoured in the Institutes, on the other hand, 'professionals' will vary in the way they pursue desired ends, and the results of the actions they take are generally less predictable, and indeed in many cases may be difficult to describe precisely. The distinction is reflected in views of assessment in which FÁS favours 'objectivity', high levels of specification, standardized procedures, and rigid 'standards', while the Institutes of Technology favour a less well-defined procedure which allows for greater variation, flexibility, and human judgment. While the latter view leaves itself

open to charges of subjectivity and lack of comparability, the more 'objective' approach involving detailed specification and standards cannot be regarded as immune to the same charges. Any procedure to establish standards involves human judgment, and so is subject to error. In addressing the two traditions, the most important objective to work towards might be an understanding of standards as constructs that are shared in a community of interpreters.

RECOMMENDATIONS

In making recommendations, consideration might be given to practice in other countries. However, traditions and practice vary very much from country to country. This is exemplified in the practices of the countries described in the Appendix (prepared by Human Resources Development Consulting Services, Darmstadt). However, there are some similarities. In France, Germany, and the Netherlands, apprenticeship examinations have written, oral, and practical components, and the written component in each country comprises a variety of item types. On the other hand, systems differ in the way assessment is managed. In France, the procedures are basically the same as those used for public examinations in secondary schools. Like public examinations, apprenticeship examinations, are the responsibility of Academies. They are administered in examination centres and are marked by teachers and professionals nominated by the Academie. The approach fits well in a system that is centralized, and that values, and provides for, progression. By contrast, there is much less control, and consequently less uniformity, in apprenticeship examinations in the Netherlands, where examinations are set and administered by each training institution, which establishes pass marks and issue certificates. In Germany, the responsibility for assessment varies for occupations.

Because of differences in tradition and practice, no universally agreed system of assessment that might serve as a blue-print exists. In light of this situation, our recommendations are based for the most part on the guidelines presented in the first section of this chapter in the belief that following them would bring the apprenticeship examination system into line with present-day views of 'best practice' in assessment.

1. Tests should comprise more than one item type. A combination of short-answer items and more extended essay-type items is proposed. The short-answer items would require a more limited response than the ones at present in use and would be designed to provide broad coverage of curriculum content, probably at a recall/understanding level. The extended essay-type items would require more indepth responses. They would assess a wider range of examinee cognition, including strategy selection, cognitive structuring, elaboration, and reasoning processes; they would assess preparedness for adapting to future and changing technologies; they would take into account the potential impact of the examinations on teaching and learning; and they would prepare students interested in progression.

   Obviously, the number of essay-type questions will be quite small. The problems created by this can be partly overcome by requiring students to provide a number of relatively short answers rather than a few very long ones. Furthermore, questions can be constructed so that responses are contingent on responses to previous sections of a question, allowing the examinee to demonstrate his/her ability to negotiate solutions through component tasks. Reliability will be enhanced if the form and length of the answer that is required

are specified, if points that should be addressed are identified, and if some structure is prescribed (see, e.g., Verma, Chhatwal, & Singh, 1997).

2. The number of short-answer questions should be increased to provide more extended curriculum coverage. This would be achieved by the use of more conventional short-answer items (i.e., ones not involving complex processes or steps). Examinees would be required to respond to all such items.
3. Essay-type items should be designed to elicit higher-order cognitive processes involving comprehension, analysis, evaluation, problem-solving, and students' ability to organize and apply knowledge.
4. Examinees should be provided with a choice of prompts in the essay-type items. While this practice raises questions about comparability, on balance it would seem worthwhile since it allows examinees to choose an area in which they can express themselves. It also allows students of varying levels of ability to choose a topic at an appropriate level.
5. Particular care will be required in essay-type examinations to ensure that examinees' ability to demonstrate their knowledge is not inhibited by language difficulties. To address this issue, the verbal demands of examinations should be kept to a minimum, and examinations should be set with the verbal abilities of the students in mind.
6. A system of moderation of the examination process will be required since examinations are administered and scored in a number of institutions.
7. Steps should be taken to enhance the reliability of tests. Reliability is enhanced (a) by ensuring that items and directions are unambiguous; (b) by having procedures in place to increase agreement among scorers; and (c) by increasing the discrimination power of items (i.e., how well individual items discriminate

between examinees who do well on the test as a whole and examinees who do poorly) (Bloom et al, 1981).

Since unreliability is a feature of essay-type examinations, particular attention will need to be given to procedures to reduce it. A number of procedures are available. In one, componential scoring procedures are prescribed (analytic scoring). Typically, key components of a domain are distinguished in a marking scheme; operational criteria are developed for each component; scores are assigned to each; and total scores are derived from some combination of individual component scores. In an alternative set of approaches, more global scoring procedures are used (holistic scoring). This involves the development of a single set of scoring criteria. For example, a number of sample/model responses may be provided for an essay, each reflecting greater proficiency in a domain (ranging, for example, from 'virtually none' to 'near perfect'). Both approaches require the development of scoring criteria based on a clear conceptualization of proficiency, competence, or 'mastery'; careful selection and training of scorers; and the operation of inter-rater and rater-trainer reliability checks.

8. The number of marks allocated to examinee responses on each examination paper should be increased (to 100). This would help improve overall reliability as well as increasing the range of distinctions that could be made reliably between levels of examinee performance.
9. The number of marks allocated to questions will vary depending on the difficulty of a question and the demands that it makes on examinees. For example, a single mark might be allocated to each of 30 short-answer questions and the remaining marks distributed among questions which demand more extended and complex responses.

10. The pass mark should be reduced to 50%. There are a number of reasons for this recommendation. First, for statistical reasons, a 50% threshold will result in the lowest amount of error. Secondly, the figure is not so low that it might suggest to examinees and users that achievement levels are too low. Thirdly, requiring a pass mark of 70% can actually depress standards, as teaching and learning will tend to focus on minimum competence, and high achieving students (e.g., with Leaving Certificate higher mathematics) will not be sufficiently challenged. Paradoxical though it may seem, reducing the pass mark could impact positively on teaching and learning and on student standards. It is perhaps for reasons such as these that the pass mark in German apprenticeship examinations is 50 percent. The credit mark should also be reduced, but should not fall within two standard errors of the pass mark. The cut-scores at present in use (70%, 85%) do not take sufficient account of measurement error.
11. Examination arrangements should be such that they facilitate students who need to repeat an examination. The conditions under which students repeat should be flexible, and students should be assisted in their preparations for the examination. The availability of an item bank would facilitate the organization of repeat examinations.
12. Consideration should be given to issues relating to the administration of the examinations, and the most appropriate procedure should be determined by stakeholders. A number of issues will need to be addressed. Will tests be centrally devised? Will only one test be available for all examinees at a particular phase at a particular time, or will a choice of tests be available? Will tests be constructed within institutions (with appropriate methods for moderation to ensure comparability across colleges)? Will examination papers be available to students after they have taken an examination?

Any response to these questions is likely to be associated with advantages and disadvantages. A unique centrally devised test means that all students respond to the same tasks. This will enhance comparability, though not assure it as variation may occur in marking. Further, if the same test is used from year to year, comparability over time is enhanced. However, the use of a centralized test is also associated with lack of flexibility in administration and leakage of questions. Tests constructed within institutions, by contrast, have the advantage that they follow traditional practice in post-secondary education, capitalize on instructors' knowledge of curricula and students, and allow a good deal of flexibility in administration. They would seem most appropriate when extended essay-type questions are used. They do, however, raise problems of comparability, and require some form of moderation.

The release or withholding of test papers after an examination is an issue whether examinations are set centrally or locally. Advantages in not releasing them are associated with comparability and savings in test construction. However, in addition to problems of leakage, the withholding of tests means that the tests do not serve the function of providing students with explicit targets for their study or the standards they are expected to achieve.

In considering item formats, it was recommended in this report that a variety be used in apprenticeship examinations. Consideration might also be given to variation in administrative arrangements. For example, short-answer items might be centrally provided in an item bank while essay-type questions would be constructed and scored locally. However, the construction of an item bank would require considerable investment since items would need to be clearly identified with the curriculum area being assessed; the level of response they require; the key learning categories being assessed; and their difficulty level.

In considering the variety of options and trade-offs that are available in the design of an assessment, it should be borne in mind that all procedures have disadvantages as well as advantages. The aim should be to maximize advantages and minimize disadvantages in light of the objectives of the assessment and the conditions in which it is administered.

REFERENCES

Ambrosio, T., Byrne, N.M.T., Oliveira, T., Page, K.W., & Richini, P. (1995). Teachers and trainers in vocational training. Volume 2: Ireland, Italy, Portugal. Thessaloniki: CEDEFOP – European Centre for the Development of Vocational Training.

Attwell, G., & Brown, A. (2000). The acquisition of skills and qualifications for lifelong learning, trends and challenges across Europe. In B. Sellin (Ed.), European trends in the development of occupations and qualifications. Findings of research, studies and analyses for policy and practice, Vol II (pp. 163-187). Thessaloniki: CEDEFOP – European Centre for the Development of Vocational Training.

Bennett, R.E. (1993). On the meanings of constructed response. In R.E. Bennett & W.C. Ward (Eds), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale N.J.: Lawrence Erlbaum.

Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). Evaluation to improve learning. New York: McGraw Hill.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Cizek, G.J. (Ed.). (2001). Setting performance standards. Concepts, methods, and perspectives. Mahwah N.J.: Lawrence Erlbaum.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.; pp. 271-302). Washington D.C.: American Council on Education.

Cronbach, L.J. (1941). An experimental comparison of the multiple true-false and multiple-choice tests. Journal of Educational Psychology, 33, 401-415.

Cronbach, L.J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

d'Ydewalle, G., Swerts, A., & DeCorte, E. (1983). Study time and test performance as a function of test expectations. Contemporary Educational Psychology, 8, 55-67.

Ebel, R.L. (1970). The case for true-false items. School Review, 78, 373-389.

ESF (European Social Fund). Programme Evaluation Unit. (1995). Standards based apprenticeships programme. Preliminary evaluation. Dublin: Author.

ESF (European Social Fund). Programme Evaluation Unit. (1999). Apprenticeship and traineeship. Evaluation report. Dublin: Author.

Eisemon, T.O. (1990). Examination policies to strengthen primary schooling in African countries. International Journal of Educational Development, 10, 69-82.

FÁS. (1999). Specification for the setting of written short answer papers, short answer questions, marking criteria and supplementary instructions. Dublin: Author.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

French, J.W. (1956). The effects of essay tests on student motivation. Research Bulletin no 4. Princeton N.J: Educational Testing Service.

Gunning, D. (2000). Competence-based training in Scotland, England, Australia and New Zealand. Final Report. Glasgow: Scottish Qualifications Authority.

Haladyna, T.M. (1994). Developing and validating multiple-choice items. Hillsdale N.J.: Lawrence Erlbaum.

Haladyna, T.M. (1997). Writing test items to evaluate higher order thinking. Boston: Allyn & Bacon.

Haladyna, T.M. (1998). Fidelity and proximity to criterion: When should we use multiple-choice? Paper presented at annual meeting of the American Educational Research Association, San Diego.

Haney, W., & Madaus, G.F. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, 70, 683-687.

Hannah, L.S., & Michaelis, J.U. (1977). A comprehensive framework for instructional objectives. Reading MA: Addison-Wesley.

ITAC (Institutes of Technology Apprenticeship Committee). (n.d.). Some contexts of apprenticeships and progression. Unpublished paper.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Kellaghan, T., Madaus, G.F., & Raczek, A. (1996). The use of external examinations to improve student motivation. Washington D.C.: American Educational Research Association.

Kerr, D. (2002). Post-apprenticeship progression: FÁS perspective on transfer and progression for craftpersons. Paper presented at Institutes of Technology seminar, Sligo.

Kinney, L.B., & Eurich, A.C. (1932). A summary of investigations comparing different types of tests. School and Society, 32, 540-544.

Koretz, D. (1995). Sometimes a cigar is only a cigar, and often a test is only a test. In D. Ravitch (Ed.), Debating the future of American education. Do we need national standards and assessments? (pp. 154-166). Washington D.C.: Brookings Institute.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 31, 234-250.

Madaus, G.F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P.W. Jackson (Ed.), Handbook of research on curriculum (pp. 119-154). New York: Macmillan.

Martinez, M.E. (1999). Cognition and the question of test item format. Educational Psychologist, 34, 207-218.

McDonagh, S. (2001). Institutes of Technology: Some new directions. Paper presented at Institutes of Technology seminar.

Messick, S. (1987). Assessment in the schools: Purposes and consequences. Princeton N.J.: Educational Testing Service.

Messick, S. (1993). Trait equivalence as construct validity of score interpretations across multiple methods of measurement. In R.E. Bennett & W.C. Ward (Eds), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 61-73). Hillsdale N.J.: Lawrence Erlbaum.

Messick, S. (1995). Validity of psychological assessments: Validation of inferences from person's responses and performance as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M.D. Hakel (Ed.), Beyond multiple choice. Evaluating alternatives to traditional testing for selection (pp. 59-74). Mahwah N.J.: Lawrence Erlbaum.

Meyer, G. (1934). An experimental study of the old and new types of examination. Journal of Educational Psychology, 26, 30-40.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. Linn (Ed.), Educational measurement (3rd ed.; pp. 335-366). New York: American Council on Education/Macmillan.

National Qualifications Authority of Ireland. (2002). Towards a national framework of qualifications – Establishment of policies and criteria. Dublin: Author.

O'Connor, L. (October 2000). The evolution of standards-based training in Ireland. Techdirections, 30-33.

O'Connor, L., & Harvey, N. (2001). Apprenticeship training in Ireland: From time-served to standards based: Potential and limitations for the construction industry. Journal of European Industrial Training, 25, 332-342.

Osterlind, S.J. (1998). Constructing test items: Multiple-choice, constructed-response, performance, and other formats (2nd ed.). Boston: Kluwer Academic.

Programme for Economic and Social Progress. (1991). Dublin: Stationery Office.

Rodriguez, M.C. (2002). Choosing an item format. In G. Tindal & T.M. Haladyna (Eds), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation (pp. 213-231). Mahwah N.J.: Lawrence Erlbaum.

Ryan, J.M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladya (Eds), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation (pp. 67-88). Mahwah N.J.: Lawrence Erlbaum.

Sax, G., & Collet, L.S. (1968). An empirical comparison of recall and multiple-choice tests on student achievement. Journal of Educational Measurement, 51, 169-173.

Snow, R.E. (1993). Construct validity and constructed-response tests. In R.E. Bennett & W.C. Ward (Eds), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 45-60). Hillsdale N.J.: Lawrence Erlbaum.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. Journal of Educational Measurement, 31, 113-123.

Traub, R.E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R.E. Bennett & W.C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 29-43). Hillsdale N.J.: Lawrence Erlbaum.

Traub, R.E., & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. Applied Psychological Measurement, 1, 355-369.

Tyler, R.W. (1966). What testing does to teachers and students. In A. Anastasi (Ed.), Testing problems in perspective. Washington D.C.: American Council on Education.

Vallence, T. R. (1947). A comparison of essay and objective examinations as learning experiences. Journal of Educational Research, 41, 279-288.

Verma, M., Chhatwal, J., & Singh, T. (1997). Reliability of essay type questions – Effect of structuring. Assessment in Education, 4, 265-270.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Towards a Marxist theory of test construction. Applied Measurement in Education, 6, 103-118.

Wainer, H., Wang, X-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? Journal of Educational Measurement, 31, 183-199.

Ward, W.C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6, 1-11.

Ward, W.C., Frederiksen, N., & Carlson, S.B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.), Educational measurement (3rd ed.; pp. 335-366)). New York: American Council on Education/Macmillan.

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

Wiliam, D. (1996). Meanings and consequences in standard setting. Assessment in Education, 3, 287-307.

APPENDIX 1

FINAL ASSESSMENT/TESTING IN VOCATIONAL EDUCATION AND TRAINING IN FRANCE, GERMANY, AND THE NETHERLANDS

Gerhard Kohn

Human Resources Development Consulting Services

Darmstadt

Final Assessment / Testing in Vocational Education and Training in France				
No.	**Categories/aspects**	**Guiding questions**		**The French case**
1	**Legal basis** for assessment / examination of apprentices (mainstream)	1.1	Which legal regulations and rules are in place to prepare, conduct and evaluate assessment / testing?	For each diploma = vocational qualification (a total of 744 for all types of specialisations and levels : CAP, BEP, Bac technologique, Bac professionnel, BP, BTS), there is one *referential* (*le référentiel*) laid down in a decree by the Minister of National Education, based on a proposal elaborated by a Consultative Professional Commission of the corresponding economic branch. The referential includes three parts 1) the *referential of professional activities* (*le référentiel des activités professionnelles*) 2) the *certification referential* (*le référentiel de certification*) and 3) the *examination regulation (le règlement d'examen)*. The referential defines the global and operational capacities (knowledge and skills, associated knowledge (savoirs et savoirs-faire, savoirs associés) which constitute the diploma. The referential contains further an examination regulation which defines the modalities of evaluation. The total of the qualifications evaluated with the candidate include the following components: the capacities, the knowledge, the know-how and the skills (capacités, connaissances, savoirs et savoirs faire). The *examination regulation* defines for every diploma the registration conditions, the tests and the execution of these tests. The regulation stipulates, for the different subjects, the objectives, the contents, the evaluation criteria, the form (written / oral / practical / control during the training process), the evaluation environment and the coefficient related to the whole examination The *Consultative Professional Commissions* (a total of 17) are composed of 4 groups of experts : 1) representatives of employers (10 seats), 2) of employees (10 seats), 3) the public authorities : one representative per competent Ministry, 2 inspectors from National Education, 1 representative from CEREQ (national R&D institute for TVET) and 1 from AFPA (the national association for continuing training), 4) qualified personalities (11 seats) : 6 rep. from the teachers' unions, 2 rep. from the parents' organisations, 2 rep. from the assemblies of Chambers of Commerce and Industry and of Crafts Chambers, and finally 1 rep. from the technical education counsellors.

No.	Categories/aspects	Guiding questions		The French case
1	**Legal basis** (cont.)	1.1		The national register of professional certifications: All diploma and titles with a professional objective delivered in the name of the State and created after opinion of the consultative bodies, associating the representative organisations of employers and workers, are registered by law in this register. They are classified by field of activity and by level. A national commission of professional certification is created for this purpose and placed with the Prime Minister. This commission establishes and updates the national register of professional certifications. The commission is in charge of the renewal and adaptation of diploma and titles to the evolution of qualifications and of work organisation. (« Loi de modernisation sociale » of 17 January 2002, Article 134)
		1.2	Which are the institutions in charge?	The Rectorate of the Academy (Ministry of national education (28 Academies all over France)
2	**Modes of exam/assessment**	2.1	Which modes of exams /assessment exist (formative, summative, mix)?	The diploma can be obtained globally or progressively : a) The global form : It means that the examination takes place at the end of training. *b)* The progressive form : The pupil or the apprentice obtains progressively and step by step during his training process the elements of the diploma referring to the units according to two modalities : either punctual control or continuous control, with at the minimum one final test, if the training is accomplished in a recognised public institute. The final exams are exclusively “summative” = they establish the acquired knowledge and skills and those missing, for each individual candidate per field of knowledge and skills. The exams in the course of training are of a mixed character (“summative” and “formative”).

No.	Categories/aspects	Guiding questions		The French case
2	**Modes of exam/assessment** (cont.)	2.1		*The Validation of Acquired Experience (VAE)* : (Social Modernisation Law, issued on 17 January 2002, Articles 133-146) The "validation of acquired experience" is a proper track for obtaining professional diploma and titles, parallel to school and university education, apprenticeship and continuing training. The validation is assured by a jury whose composition guarantees a significant presence of qualified representatives from the professions concerned. The jury can attribute the totality of a diploma or title. Otherwise, it states as to the coverage of the validation and, in case of a partial validation, it determines the types of knowledge and capabilities which are subject to an additional control. The jury expresses its vote in view of a file constituted by the candidate, following an interview with the candidate and, eventually, a real or simulated professional situation, if this type of procedure is foreseen by the authority delivering the certification. (Article 134)
3	**Test item writing**	3.1	Which types of items are used – for both parts: written and practical assessment, and oral, if applicalbe (e.g. standardised items, item banks)? Are there any test item data banks?	The examination regulation, which is a part of the « referential » (guidelines) for each diploma, determine 1) the tests at the end of the training, 2) the controls in the course of the training (CCF). The tests are composed of written, oral and practical tests (e.g. standardised tests). The choice of tests (oral / written / practical) is established in the diploma decree. The disciplinary inspector of the Academy (for the Vocational Aptitude Certificate – CAP, and the Vocational Education Diploma - BEP; both corresponding to the skilled worker or employee level) or the general inspector (for the Vocational Baccalaureate – Bac pro, and the Superior Technician Diploma - BTS) determine the "evaluation support" (requirements, contents etc.). Test writing includes the elaboration of the "evaluation support". «The test situations aim at questioning the candidate per «sondage» by means of tests, multiple choice questionnaires, questionnaires with open and closed questions, exercises which are based on course contents, case studies, problem situations etc. allowing to control the candidate over a large spectrum of knowledge described in the referentials.»[1]

[1] Source: Cellule nationale de professionalisation – groupe interministériel. Repères sur la certification et la validation des acquis. Nouveaux services emploi-jeunes, mars 2000

No.	Categories/aspects	Guiding questions		The French case
3	**Test item writing** (cont.)	3.2	Who develops test items (written, practical, oral)?	An item selection committee writes the test items and the evaluation supports. The latter vary per examination session. The group works under the responsibility of an inspector of the Ministry of national education (regional level for the CAP and BEP diploma; national level for the Bac pro and BTS diploma).
4	**Basis for item writing**	4.1	Are test items based on standards, or on curricula, or on training plans?	The tests are based on the levels of requirements, as they are established by the « referential for vocational activities ». The final qualification corresponds to a group of vocational activities established by the branch.[2]
5	**Test paper composition**	5.1	Which are the components of test papers (eg. written tests, performance tests, oral tests)?	There are written, oral and practical tests, and the control in the course of training (CCF). The composition is determined by the examination regulation (component of the referential) which has legal value. The tests determine : 1) the competences, 2) the knowledge, 3) the know-how 4) the skills (capacités, connaissances, savoirs et savoirs faire)/[3]
6	**Conduction of tests**	6.1	Who is responsible for running exams/assessments?	The administrative authority in charge of the examinations is the Exams and « concours » Division of the Academy rectorate, depending of the Ministry of national education.
		6.2	Who is in charge of monitoring and evaluation?	Monitoring and evaluation are in charge of teachers nominated by the Exams and « concours » Division, in collaboration with the inspectors' corps.
		6.3	Who establishes pass marks and grading schemes?	The evaluation is in charge of correctors nominated by the Academy rector. The correctors are either teachers or representatives of the professional branch concerned. The correctors evaluate alone or as a group. The jury validates, depending on the examination regulation, and the rector certifies, in the name of the Minister of national education.

La terminologie française n'utilise pas le terme « standard » ou un équivalent.
[3]Voir annexe „Définitions"

No.	Categories/aspects	Guiding questions		The French case
6	**<u>Conduction of tests</u>** (cont.)	6.4	Who is represented on assessment/testing panels or committees (teachers / trainers vs. employers / employees representatives)?	Half of the jury members are teachers, half are representatives of the professional branch.
		6.5	Which is the size of the examination candidate group?	The size of the examination candidate group varies according to the type of test.
		6.6	Where are the examinations conducted (VET college, enterprise, other)?	The candidates coming from public and private establishments of a region are examined in schools which have been declared as « examination centres » by the Academy rector (e.g. Technological Lycées). The practical tests can take place in a professional environment, e.g. in a retail shop for a sales CAP.
7	**<u>Certification / awarding bodies</u>**	7.1	Who issues the certificates?	The Academy rector signs the diploma in the name of the Minister of national education.
		7.2	What are the contents of certificates?	The diploma mentions the type of diploma and the specialisation selected. The candidates who have not reached the requirements of the diploma, receive a certificate for the certification units they have passed. The certificate is valid during 5 years. The candidate can present him or herself within this period and pass the tests which are missing.

No.	Categories/aspects	Guiding questions		The French case
8	**Access** to exams and **progression routes** after exams / certification	8.1	Who has access to exams?	The access to the exams is open to all persons providing the enrolment conditions for the exams, established in the referential governing the diploma chosen by the candidate : CAP, BEP, Bac technologique, Bac professionnel, BP, BT, BTS a) For secondary education students leading to the corresponding diploma b) For the apprentices preparing for the same diploma c) For free candidates under certain conditions d) For adults in continuing training e) For candidates for the validation of acquired experience (VAE)
		8.2	What are the progression routes after examination / certification?	a) Graduates of the diplomas CAP and BEP are entitled to embark on the Bac professionnel track (two additional years of studies) or to enter in active life. Graduates of the BEP diploma have also the right to embark on the track of general secondary education which leads to the Baccalauréat. b) Graduates of the diplomas Bac technologique and Bac professionnel are entitled to enter a university or another post-Bac training path, or to enter in active life.

France

Explanation of Abbreviations and Technical Terms

Abbreviations

Baccalauréat Professionnel (Bac pro)	Vocational baccalauréat
Brevet d'Études Professionnelles (BEP)	Certificate in vocational studies
Brevet des Métiers d'Art (BMA)	Certificate in art studies
Brevet Professionnel (BP)	Vocational certificate
Brevet de Technicien (BT)	Technician certificate
Baccalauréat Technologique (BTn)	Technological baccalauréat
Brevet de Technicien Supérieur (BTS)	Higher technician certificate
Certificat d'Aptitude Professionnelle (CAP)	Certificate of vocational competence
Diplôme des Métiers d'Art (DMA)	Diploma in art studies
Diplôme Supérieur d'Art Appliqué (DSSA)	Higher diploma in applied art
Diplôme Supérieur de Technicien (DST)	Higher technician diploma
Mention Complémentaire (M.C.)	Supplementary reference

Definitions

Levels of qualification and training

Frame of reference of vocational activities[4]
"A document describing the content and methodology of tasks and activities, conditions of practice, aims, objectives or goals. In the context of national education, this description is based on the type of employment, to the extent that it combines the analysis of vocational situations that are sufficiently close to constitute an entity, an occupation or a generic profession in one or several vocational sectors. This description refers to practice rather than competence."

Frame of reference of a diploma[5]
"A document that provides an exact inventory of abilities, skills and knowledge required to secure the desired diploma. It identifies the situations in which these can be

[4] Source: Cellule Nationale de Professionalisation. Groupe Interministériel. Repères sur la certification et la validation des acquis. Nouveaux services emploi-jeunes, mars 2000

[5] Idem.

evaluated, the levels to be attained, and the criteria for success in assessing the performance of a trainee. This description is not a syllabus but an evaluation instrument. It indicates what is to be evaluated, and the method and the instruments to be used in the evaluation."

Validation of vocational attainments[6]

"A specific mode of awarding vocational or technological diplomas by granting exemptions from tests or constituent units of the diploma, in accordance with an assessment of the knowledge and skills based on an analysis of a written and/or oral description of the activity involved. A diploma cannot be obtained solely by this method."

Modes of evaluation[7]

Formative evaluation

It "facilitates the trainee throughout the duration of the training, in analysing difficulties, and in identifying points of reference to help consolidate his attainments and to formulate training needs. It allows the tutor or the trainer to make adjustments to the training. This type of evaluation implies a process of continuous assessment in which it is possible to check if trainees have acquired the skills and knowledge in the course of the instruction process."

Summative evaluation

It "makes a list of attainments at a particular juncture or at the end of the training period. It makes little or no change to the planning of the training. This type of evaluation is distinguished from the preceding one by its focus on a specific time in the training process. It also tries to estimate the variation that exists in a range of the most important attainments."

Certified evaluation

Its "function is to evaluate attainments from the point of view of the award of the qualification or diploma envisaged in the situations provided for in the examinations. It is based then on the recognition or validation of attainments. This definition superimposes three concepts that come within the scope of an evaluation of candidates for a national diploma. In fact, in this particular case it refers to the same official authority (the Ministry of National Education by delegation to the DLC) which organises the tests of attainment through its examination service and which thus guarantees on the one hand their reliability through the supervision of the organisation of the tests, and on the other their validity insofar as they conform to the reference points that determined the knowledge to be acquired. Recognition is further strengthened by the fact that this authority is the source of the instruments that provide the indicators and norms that determine the value of this measure."

Abilities[8]

Competence: Collection of skills and knowledge utilized in an activity and adapted to the needs of an employment situation.

---

[6] Idem.

[7] Idem.

[8] Documents méthodologiques pour l'élaboration des diplômes : Référentiel des activités professionnelles et référentiel de certification du domaine professionnel, CPC 93/1, Ministère de l'Education Nationale, sans année.

Knowledge: The sum of an individual's knowledge relating to objects and to the environment, to the properties of objects, and to laws relevant to the environment.

Understanding: The sum of an individual's knowledge. Understanding is used here in a generic sense; know-how and knowledge are particular manifestations of understanding.

Know-how: The sum of an individual's knowledge relevant to activity in a technical and social milieu. A particular characteristic of this knowledge is that it can only be constructed and stored by actual activity (in a real or simulated situation) and can only be reactivated in the course of an activity. Therefore, know-how can be comprehended by an external observer only through the activity itself and observable signs (words, actions, manipulation of objects, etc.).

Ability[9]

The concept of ability is defined in different ways by different authors. In the context of frames of reference, an ability is the sum of the skills that an individual applies in a variety of situations (e.g., to communicate, to be informed). An ability cannot be assessed. It concerns the axis of training along which students ought to progress. The axis applies to all disciplines (maths, literature …) of the same training course.

[9] Source: Bernard Porcher, Du referentiel à l'évaluation, Editions Foucher, Paris, 1992, p. 93.

Final Assessment/Testing in Vocational Education and Training in Germany

No.	Categories/aspects	Guiding questions		The German Case
1	**Purpose of assessment/testing in vocational education and training (VET)**	1.1	What is the <u>main</u> purpose of assessment/testing? ⇒ To provide evidence about occupational competences of candidates? ⇒ To provide access to further/higher education and training?	The main purpose of final assessment/testing in dual vocational education and training is to provide **evidence about the occupational competences** of the candidates. Candidates successfully passing the final assessment ("Facharbeiter-/ Fachangestelltenprüfung") prove to be competent through their final diploma/certificate (e.g. "Facharbeiterbrief") issued by the repective chamber. Employers use this certificate for recruitment, allocation, career decisions, in-company further training decisions, etc..
2	**Assessment/testing system in the context of the education and employment**	**<u>2.1</u>**	**<u>Is the assessment/testing system hooked up to the</u>** ⇒ Education system? ⇒ Employment system?	Final assessment/testing is hooked up to the German system of **chambers** ("competent bodies" according to the Federal Law on VET, 1969), hence it is managed by representatives of the employment system. Vocational colleges ("Berufsschulen") do not have a direct influence on final assessments; however, teachers of the vocational colleges are represented in the assessment panels composed by the chambers.
3	**General organization**	3.1	How is (final) assessment/testing generally organized? ⇒ Centrally? ⇒ De-centrally? ⇒ Combinations?	Implementation of final assessments/tests is centrally regulated (see No. 10 below) but it is organised de-centrally by chambers. However, for a number of occupations, the test items (both for written and practical (performance) tests) are developed centrally (e.g. by "lead chambers") and distributed to chambers for use during testing.

4	**General quality criteria for the overall process of assessment/testing**	4.1	Are there any general criteria to guide preparation, implementation and evaluation of final assessement/testing?	Internationally recognised quality criteria, such as “objectivity”, “validity”, “reliability” and “transparency” (of procedures and grading) are observed during test paper design and test implementation.
5	**Certificate/Diploma**	5.1	Which information do final certificates/ diploma provide/contain?	After final assessment/testing successful candidates are awarded a final diploma/certificate (“Facharbeiterbrief”, “Fachangestelltenbrief”) by the chamber. This certificate contains information on (a) personal details, (b) subjects of the final exams and the related achievements, (c) a statement that the final assessment/test has been passed successfully. In addition to this final certificate/diploma issued by the chamber, candidates will get a school leaving certificate issued by the vocational college (certifying the achievements during off-the-job courses at the college); and they get an “appreciation letter” (“Arbeitszeugnis”) issued by the employer.
6	**Contents of final assessment/ testing**	6.1	Where are the contents of final assessment/testing derived from? ⇒ Occupational profiles/standards? ⇒ Training curricula/plans?	The contents of final assessments/tests are mainly derived, from the ‘skeleton training plans’ (“Ausbildungsrahmenpläne”) which are part of the (national) Training Ordinances (“Ausbildungsordnungen” = statutory instruments). However, the skeleton curricula of the Vocational Colleges (which are developed on “Länder”-level) are also taken into consideration by test paper setters.
		6.2	By whom and how are the contents of assessment/testing selected from the wide scope of possible contents?	A tripartite (employers, unions, vocational colleges) assessment commission (“Prüfungskommission”) selects contents from skeleton training plans by means of consensus.

No.	Categories/aspects	Guiding questions		The German Case
7	**Procedures and instruments**	7.1	Which parts do final assessments/tests consist of (e.g. written tests, performance tests, oral tests)?	Usually, final assessment/testing consists of two main parts: (a) occupation-oriented theory/knowledge (b) occupational practice/competence Normally, the theory test is a written exam and the competence assessment is a performance assessment. Both can be complemented by an oral exam.
		7.2	Which methods/types of tests are used for written assessment/testing (if applicable)?	Usually written tests make use of "open answer type" as well as pre-structured types of test items (e.g. multiple choice items, matching items). However, the composition of test papers varies between occupational areas (e.g. technical occupations, business occupations).
		7.3	Which methods/types of tests are used for performance assessment/testing?	Practical assessment is done by allocating real work assignments in real work environments (including the aspect of "work planning") or in simulated work environments to candidates. If applicable, candidates may also be requested to manufacture a real product (e.g. in tool and dye making, carpentry).
		7.4	Who develops test items?	Tri-partite Commissions (employers, unions, vocational colleges) develop test items and test papers following the consensus principle. As mentioned above, in several occupational areas this is done by a "lead chamber" for a number of other associated chambers, while in other occupational areas the chambers compose their own commissions for their vicinity only.
		7.5	Are there any test item data banks?	Lead chambers have build up test item data banks from previously used items. Some of these items are published for exercise by candidates after having been used in assessment. However, it is difficult to draw an overall picture for all chambers.

		7.7	Are there any marking sheets/keys?	Usually, all test items and test papers are going along with answer keys and marking schemes for use by the assessment commission during assessment.
		7.8	Who is represented on assessment/ testing panels or committees (e.g. teachers/trainers vs. employers'/employees' representatives)?	Assessment commissions are composed of 1 to 2 representatives each of employers, trade unions and vocational colleges (or a multiple of these numbers).
		7.9	Who decides about the composition of assessment/testing panels, and selection of panel members?	In principle, the composition of assessment commissions is regulated by the Federal Law on Vocational Education and Training ("Berufsbildungsgesetz", BBIG, 1969). Actual composition of commissions is done by the competent chambers based on the provisions of the Federal Law.
		7.10	Are the members of assessment/testing panels/committees trained for their tasks?	There is no real "training" for commission members. However, newly appointed members are introduced into their tasks, and given guidance, by serving members.
		7.11	What are the minimum pass marks for (written, performance) assessment?	Generally, candidates must score at least 50 % of maximum sores in both written and practical/performance part of a final assessment/test. However, there may be certain subjects where a minimum pass mark of 50 % is compulsory, and cannot be compensated by better marks in other subjects.
8	**Candidates for assessment/ testing**	8.1	Are there any preconditions to participate in assessment/testing, and if yes, which are the preconditions?	There is a provision for external candidates, i.e. candidates who have not completed a formal dual training in a recognised occupation, but have at least double the time spent on the job (than is prescribed for the related apprenticeship) to participate in final assessment/testing. However, chambers are quite restrictive on this possibility.
		8.2	Are there any rules determining the ratio between numbers of candidates and size of the testing panel?	There a no general rules for this ratio. Actual ratios are based on experiences with previous tests.

9	**Infrastructure**	9.1	Where is final assessment/testing conducted? ⇒ In schools/training centres? ⇒ In companies/at real workplaces?	Practical (performance) testing is usually conducted in companies (in case of chambers of industry and commerce) or in chamber-owned training centres (in case of chambers of crafts). However, chambers may decide to conduct actual assessment in workshops of vocational colleges but they retain in any case the overall organisational responsibility.
		9.2	How and by whom is the final assessment/testing prepared?	Chambers issue tools, equipment and material lists based on the test paper and communicate these lists to the location (company, centre) where the assessment is planned to be conducted. Based on these specifications, it is then the companies or centres who will prepare the stage for actual assessment.
		9.3	Who finances final assessment/testing? (Are there any assessment fees to be paid by candidates / others?)	Chambers and companies involved in assessment are bearing all the cost of assessment. The State contributes by paying salaries of teaching staff of vocational colleges involved in assessment commissions. Candidates fees do also contribute but only to a minor degree to overall financing of assessments.
		9.4	Which are the most critical steps in the process of preparing, conducting and evaluating assessments/testing?	The most critical aspects in preparing and implementing final assessments are (a) time planning, and (b) finalising conclusions of assessments on the last day of any assessment cycle. (Candidates need to be told the final result on the last day of their assessment for labour-related legal reasons.)
10	**Legal basis**	10	Which legal regulations and rules are in place to prepare, conduct and evaluate assessment/testing?	Basically, it is the Federal Law on Vocational Education and Training (“Berufsbildungsgesetz, BBIG, 1969) that regulates final assessments/tests. Further aspects of final assessments are regulated in the Training Ordinances for each recognised occupation (approx. 350 recognised occupations).

*) See also: Reisse, W.: Assessment, examination and certification of individuals in vocational education and training. In: G. Kohn et. al: Compatibility of Vocational Qualification Systems. GTZ Publication No. 16, March 2000

Final Assessment/Testing in Vocational Education and Training in the Netherlands

No.	Categories/aspects	Guiding questions		The Dutch case
1	**Legal basis for assessment / examination of trainees / apprentices**	1.1	Which legal regulations and rules are in place to prepare, conduct and evaluate assessment / testing?	**Legal basis** The Adult Education and Training Act of 1996 (Wet Educatie en Beroepsonderwijs - WEB) sets the framework for the Durth VET system in force. Its objectives are: 1) to improve the quality of vocational education and training 2) to strengthen the ties with the labour market 3) to introduce a coherent qualification structure 4) to increase retention rates. The principal actors within the system are: 1) The national Ministry of Education, Culture and Science (Ministerie van Onderwijs, Cultuur en Wetenschappen) is the Government authority supervising the sector. 2) The 21 national branch organisations, composed of representatives from employers, workers and teachers, obtain the power to establish national qualification standards (eindtermen). 3) The individual public training centres on the regional level (44 ROCs, 15 AOCs and some specialised institutes) obtain the authority to conduct training and assessment, in line with the national qualification standards. 4) External evaluation bodies assess 50 % + 1 Partial Qualification (Deelkwalifikatie). (The law stipulates “the smallest majority”.) Before 1989/90, training for national qualification standards had to be the same in the whole country. The training institutes have a large degree of autonomy, but are subject to regular inspections from the Ministry of Education, through its provincial inspectorate. Each training institute is obliged to present every two years a “Quality Report”. The provincial school inspectorate reports every year.

No.	Categories/aspects	Guiding questions		The Dutch case
1	**Legal basis for assessment / examination of trainees / apprentices** (cont.)	1.1 cont		**The future of Dutch VET:** There is a debate going on in the Netherlands, after the report of an independent Evaluation Committee (headed by the former mayor of the Hague, Mr. Deetman), questioning the quality assurance within the VET system. The existing external "legitimation bodies" are not independent enough to ensure high quality training provision. A new Vocational Education and Training law is under preparation (WOT). One measure for the future will be to establish a national Quality Assurance body, (Kwaliteitscentrum Examineering - KCE) under the direct control of the Ministry of Education. Works are under way to establish this body. COLO, the national federation of the branch organisations for VET, the Ministry of Education and the national Council of training institutes (BEV Raad) will be on the board of the new body. Another measure is that new national standards will be set for examinations, coming into force from August 2003 (SEP = Sectorale Examineerings Platformen). It is expected that each training institute will have to pass through a certification procedure in the future. There will be a transition period to give vocational institutes time to adapt. At the same time, extra money will be made available for schools which undertake strong efforts to comply with the new quality standards ("stimuleerings regeling").
		1.2	Which are the institutions in charge?	The national qualifications (altogether some 700) do not require a specific type of examination or training. Each recognised public training institute is responsible for curriculum development, assessment and examination. The individual training institute is also obliged to ensure the external validation, as required by the VET act. The institutions (ROCs – Regional Training Centres; AOCs – Agricultural Training Centres) set their own examinations and award certificates on the basis of the exit qualifications laid down for each course within the qualification structure. These describe the qualities in terms of knowledge, understanding, skills and, where applicable, professional attitude, which those completing the course should possess with a view to their future career and role in society. The exit qualifications are divided up into a number of partial qualifications (deelkwalificatie). Each partial qualification represents a combination of exit qualifications, which is deemed to form a separate unit in terms of professional practice in the field concerned. Students who complete the whole course successfully, are awarded a diploma. A certificate is awarded for each partial qualification obtained.

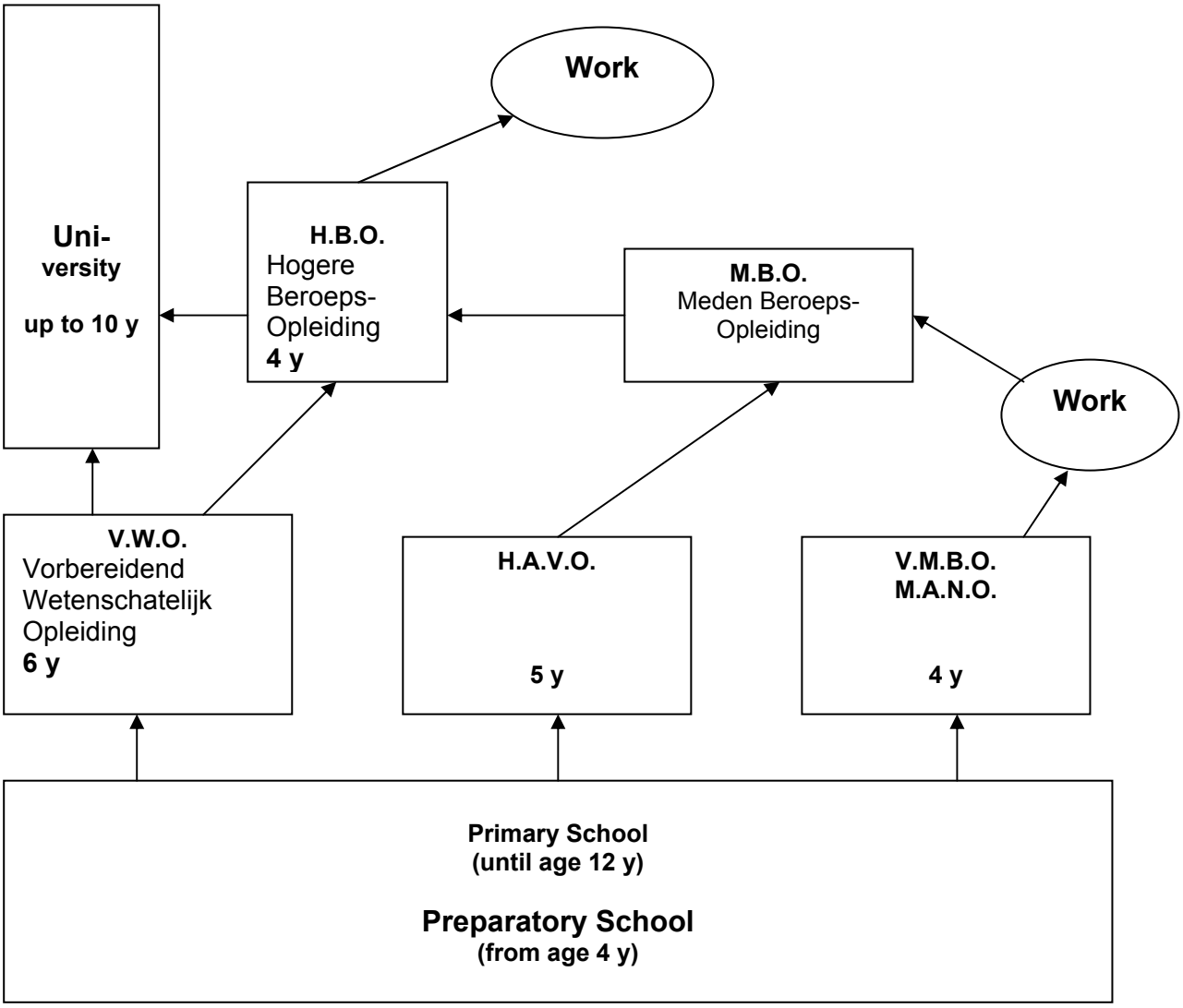
No.	Categories/aspects	Guiding questions		The Dutch case
1	**Legal basis for assessment / examination of trainees / apprentices** (cont.)	1.2 cont		Fifty percent plus one of all partial qualifications must be externally validated by an examining body recognised by the Minister of Education, Culture and Science. External validation serves to ensure that the content and level of the examinations conducted by the training institutes match the exit qualifications and that examination procedures are satisfactory. Most of the validation bodies belong to a national branch organisation for VET (LOB – Landelijke Organen Beroepsonderwijs). CELBE near s'HertogenBosch, presently the largest external validation body, belongs to three Regional Training Centres (ROCs = Regional Opleidings Centra) in the South (s'HertogenBosch, Tilburg and Nijmegen). The validation body does not attend the exams. It only verifies on the basis of a documentation provided by the training institute. The national vocational education bodies (LOBs) formulate the exit qualifications for each sector of employment, group of sectors or occupational group, which are then finalised by the Minister of Education, Culture and Science. They also put forward proposals concerning the division of exit qualifications into partial qualifications, the learning pathways and which partial qualifications are subject fo external validation. Employers' organisations, trade unions and educational institutions are represented on the boards of the national vocational training bodies on a proportional basis Each national body has an Education and Industry Committee (COB – Commissie Onderwijs Bedrijfsleven), comprising representatives of industry and the institutions (in equal numbers). Under the VET act (WEB), the examination syllabus is part of the teaching and examination regulations, a document setting out the main elements of teaching and the examinations to be held. These regulations are drawn up by the administration of the institution for each course offered and include the exit qualifications and the content and part of the examination. (Source: Website "Eurydice", inquiry on 19 February 2002)
2	**Modes of exam / assessment**	2.1	Which modes of exams /assessment exist (formative, summative, mix)?	There is no binding rule. The teachers have the authority to organise the evaluation of their students. However, they have to comply with the nationally established "eindtermen" and "deelkwalifikaties" The external validation body can reject too soft examination practices. One example: A teacher subdivided the evaluation in 200 very small examination units. There was nearly one test a week, and the external validation body had some trouble to verify, whether the totality of the tests complied with the national qualification or not.

No.	Categories/aspects	Guiding questions		The Dutch case
3	**Test item writing**	3.1	Which types of items are used : written / oral / practical assessment (e.g. standardised items, item banks)? Are there any test item data banks?	There are written, oral and practical test components. Teachers are free to use standardised tests and item banks. CITO (mainly in the fields of civil engineering) and ECABO (commercial and services qualifications) are organisations which maintain large item banks. The training institutes have to buy the test items from their own budget. Test items are selected according to the following criteria: Knowledge 1) Facts, 2) Concepts; Skills 1) Reproductive, 2) Productive. For each type of knowledge and skill, the contents are selected and it is decided whether they are tested a) written, oral or practical in school, or b) practical on a training site.
		3.2	Who develops test items (written, practical, oral)?	Teachers and specialised examination experts
4	**Basis for item writing**	4.1	Are test items based on standards, or on curricula, or on training plans?	Test items are based on the national qualifications (eindtermen) But the training institute and the individual teacher are free to develop their own curricula and testing / assessment process. There is no binding link between national qualifications (eindtermen) and test qualifications (toetstermen). The test qualifications are not clearly defined. Each national qualification (eindterm) is officially subdivided in several partial qualifications (deelkwalifikatie), as established in the CREBO (Centrale Registratie Beroepsonderwijs), the national register for vocational qualifications. Each recognised training institute (= vocational college) is free to develop its own training and examination concepts and tools, as long as it fits into the corresponding national qualification. Test items are linked with the curricula established by the individual training institute (vocational college). Assessment can be split in several parts. An example: The national qualification, for which the training institute prepares, consists of 15 “eindtermen”. The assessment process is divided in three exams. The first exam consists of 5 “eindtermen”, the second of 6 “eindtermen” and the hird of 4 “eindtermen”. 8 “eindtermen” will have to be evaluated by an external “legitimation body” (extern legitimeerd). Another example: the professional qualification “Nurse” is subdivided in 3 partial qualifications (“deelkwalifikaties”): 1) Chronically ill patients, 2) Planning of nursing care, 3) Clinical care. The training institute establishes a test schedule, the test content specification and the individual tests. CREBO divides the partial qualifications into three groups A, B and C. Group A contains obligatory assessment,Group B optional assessment, with a certain number of obligatory eindtermen, and Group C optional assessment. External validation is limited to the groups A and B. One example for C are the “doorstromkwalifikaties”. They lead to a higher level of vocational education and training, such as HBO (Hogere Beroepsonderwijs) = Higher VET.

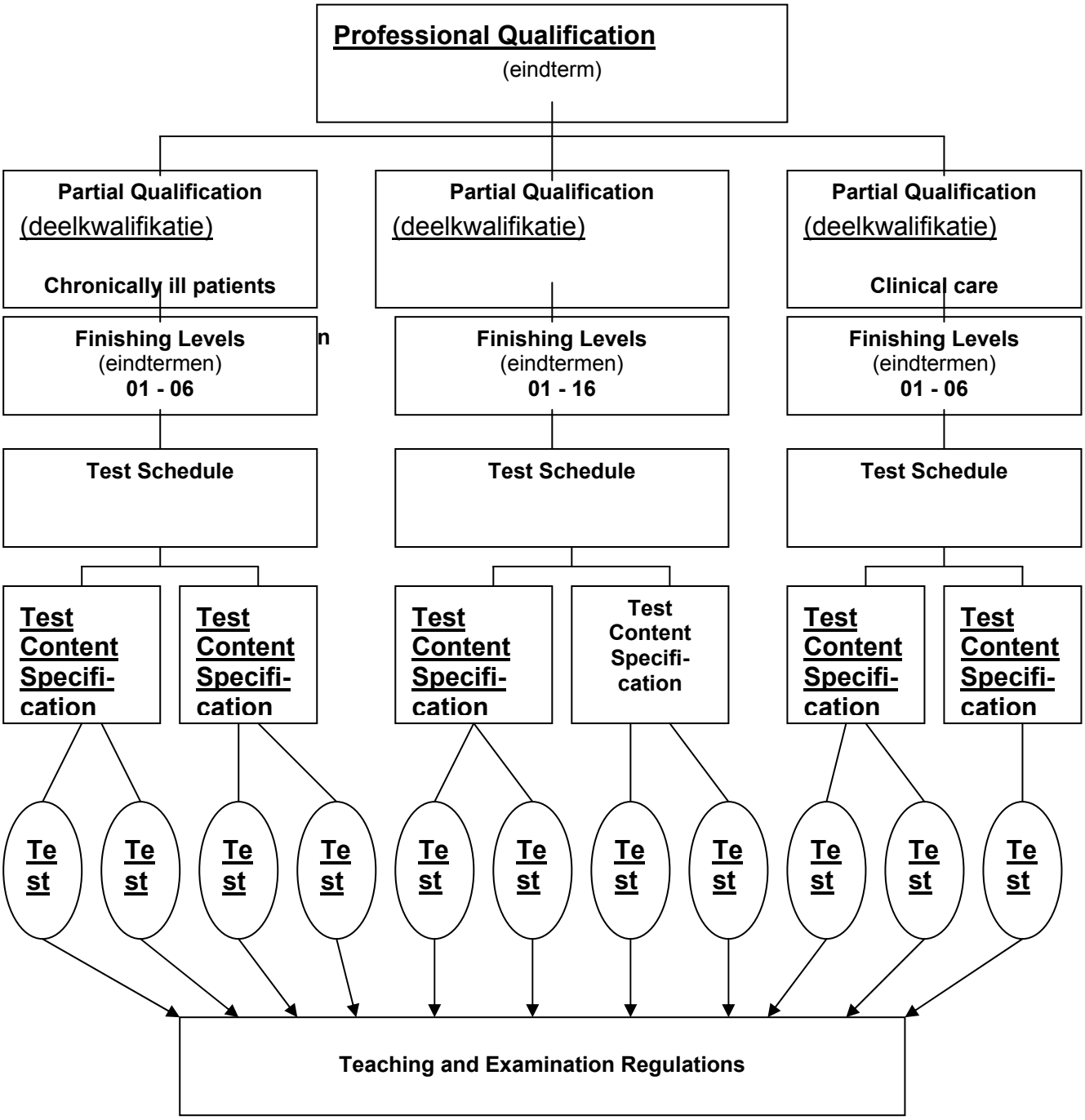
No.	Categories/aspects	Guiding questions		The Dutch case
5	**Test paper composition**	5.1	Which are the components of test papers (eg. written tests, performance tests, oral tests)?	Tests contain written, oral and practical performance tests. The individual training institute and teacher determine the mix between written tests, performance tests and oral tests. For practical performance tests, the real business environment can be chosen, e.g. for nurses in a hospital. If part of the training is conducted in an enterprise (training enterprise officially accredited by the corresponding branch organisation, the LOB), practical performance tests can take place in the enterprise.
6	**Conduction of tests**	6.1	Who is responsible for running exams/assessments?	The individual regional training institute (ROC, AOC or specialised training institute).
		6.2	Who is in charge of monitoring and evaluation?	The individual regional training institute (ROC, AOC or specialised training institute).
		6.3	Who establishes pass marks and grading schemes?	The individual regional training institute (ROC, AOC or specialised training institute).
		6.4	Who is represented on assessment/testing panels or committees (e.g. teachers/trainers vs. employers'/employees' representatives)?	The examination committee within the training institute is presided by the chief of the department concerned. Several teachers are members of the committee. It is up to the training institute to invite external practitioners to sit on the examination committee.
		6.5	Which is the size of the examination candidate group?	A class.
		6.6	Where are the exami-nations conducted: VET college, enterprise, other?	Practical tests may be done in school or on a practical training site.

No.	Categories/aspects	Guiding questions		The Dutch case
7	**Certification / awarding bodies**	7.1	Who issues the certificates?	The individual regional training institute (ROC, AOC or specialised training institute). The name of the corresponding branch organisation and of the Ministry of Education are mentioned on the certificate.
		7.2	What are the contents of certificates?	The certificate indicates the national vocational qualification, the corresponding LOB and the training institution.
8	**Access to exams and progression routes after exams/certification**	8.1	Who has access to exams?	1) Students and trainees registered in the regional training institute 2) Apprentices with a work contract, adults registering for examination. 3) Candidates from outside can add the missing partial qualifications to those which have been recognised in an assessment of their dossier (Accreditation of Prior Learning – APL, see 8.2).
		8.2	What are the progression routes after examination / certification?	1) Graduates from MBO (secondary level VET - Meden Beroepsonderwijs) are entitled to proceed to HBO (post-secondary level VET - Hoger Beroepsonderwijs), but not to university. The access to university is only possible via the HBO. 2) Candidates who have not accomplished all “deelkwalifikaties” belonging to one national qualification, receive a certificate per passed partial qualification. This enables them to go for work, but also to concentrate on the missing parts in further training. An example: A candidate who is following a training course for Level 3 can go down to a lower level and leave with a Level 2 certificate (“afstroming” in the Dutch terminology = “downstreaming”). 3) There is a special provision for professionals, who intend to obtain a certificate. Accreditation of prior learning (APL) is a proper part of the Dutch VET system. The Dutch term is “Erkenning van vorworven competenties” (EVC). It is up to the regional training institutes to organise the corresponding examinations.

The Dutch VET System



Example for test development



Final Assessment/Testing in Vocational Education and Training

Qualification levels in France and in the Netherlands[10]

The French Qualification Levels		The Dutch Qualification Levels	
Level VI	Personnel occupying positions, which require no training beyond the end of compulsory education.		
Level V a	Personnel occupying positions, which require short training lasting no longer than a year, leading, in particular, to the certificat d'éducation professionnelle (certificate of vocational education) or any other equivalent certificate or qualification.	**Level 1**	"Assistant", is responsible for his/her own activities. Work consists primarily of the application of automated routines and (to a limited extent) the application of standard procedures. It implies job-related skills and knowledge.
Level V	Personnel occupying position, which usually require a training level equivalent to the BEP or CAP.	**Level 2**	"Basic occupational practitioner", is responsible for his/her own activities. In addition, he/she and his/her colleagues share a collective responsibility and cooperate with colleagues. Work consists of applying automated routines and standard procedures. It implies occupation-related skills and knowledge.
Level IV	Personnel occupying supervisory staff positions or possessing a qualification level equivalent to a technical or technician baccalaureate or a technician diploma.	**Level 3**	"All-round practitioner", is responsible for his/her own activities and should account for his/her actions to his/her colleagues (non-hierarchical). In addition a worker has an explicit and hierarchical responsibility: he/she monitors and supervises the application of automated routines and standard procedures. His/her work comprises the application of standard procedures and combining standard procedures. In addition, he/she combines or devises procedures, in the light of work preparation and supervisory activities. It implies mainly occupational skills and knowledge.

[10] Anneke Westerhuis, European structures of qualification levels, CEDEFOP Reference series, Office for Official Publications of the European Communities, Luxemburg 2001; French Original in: Circulaire no. 67-300 du 11 juillet 1967 du ministère de l'Education nationale fixant la nomenclature interministérielle des niveaux de formation, in "Reconnaissance et validation des acquis, textes généraux, Ministère du travail, de l'emploi et de la formation professionnelle, Délégation à la formation professionnelle, Centre Inffo, Paris-La Défense, 1992

Level III	Personnel occupying positions, which usually require the higher technician diploma or a diploma from the IUTs or having successfully finished exams at the end of the first cycle of higher education.	**Level 4**	“Specialist or middle manager”, is responsible, for his/her own work and has to account for his/her actions to his/her colleagues (non-hierachical). In additio he/she bears explicit hierarchical responsibility; this responsibility concerns planning and/or administration and/or management and/or development of the whole production cycle. Furthermore he combines or devises new procedures. It implies specialist skills and knowledge and/or occupation-independent skills and knowledge.
Levels I and II	Personnel accupying positions, which usually require a level of training equal or superior to the university *licence* or diplomas of schools for professional engineer.	**Level 5**	« Occupational practicioner (professional) », is responsible for his own work and has to account for his own actions (to colleagues, non-hierarchical). Work can involve both applying and combining/devising complex standard procedures for a broad range of activities. In addition, an occupational practitioner bears excplicit hierarchical responsibility. This does not involve responsibility in an executive sense (i.e. monitoring and supervision), but rather responsibility in a formal, organisational sense. It implies specialised, occupation-independent skills and knowledge. A proefssional devises new procedures, tactical and strategic actions and has comprehensive skills with regard to policy development and execution.