DOCUMENT RESUME

ED 387 490 TM 023 246

AUTHOR Dochy, F. J. R. C.; Bouwens, M. R. J.
TITLE The Construction of Knowledge State Tests, Knowledge Profiles and the Measurement of Value Added.
INSTITUTION Open Univ., Heerlen (Netherlands). Centre for Educational Technological Innovation.
REPORT NO ISBN-90-358-0864-9; OTIC-RR-27
PUB DATE 91
NOTE 39p.
AVAILABLE FROM Open University, Secretariaat OTIC/COP, Postbus 2960, 6401 DL Heerlen, The Netherlands.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Gains; *College Students; Economics; *Evaluation Methods; Foreign Countries; Higher Education; Item Banks; *Knowledge Level; *Prior Learning; *Profiles; Student Evaluation; *Test Construction; Test Reliability; Test Use; Test Validity
IDENTIFIERS Netherlands; *Value Added

ABSTRACT

This report deals with the question of obtaining insights into states of prior knowledge and exploring knowledge state tests and knowledge state profiles. The first part introduces a few basic terms important in the context of prior knowledge state. Evaluation and the validity, reliability, and usefulness of tests are considered in a prior knowledge context. "Value added" is discussed as students' gains on tests of knowledge and skill, a measurable indication of the effects of instruction. In the second part of the report, various knowledge state tests are discussed, and the results obtained in a domain-specific knowledge state test are reviewed. The domain-specific test was constructed from the economics test item bank at the University of Limburg (the Netherlands). Results of 536 first-year students were analyzed to obtain knowledge profiles that may be used to compare groups of students or to determine whether prior knowledge profiles of good and poor students differ. Further study will clarify whether prior knowledge state profiles can help in the determination of the effectiveness of educational efforts. Six figures illustrate the analysis. (Contains 47 references.) (SLD)

The construction of knowledge state tests, knowledge profiles and the measurement of the value added

F.J.R.C. Dochy

M.R.J. Bouwens

OTIC RESEARCH REPORTS.

The Open University is responsible for developing and offering open, higher distance education in which special attention is paid to innovations in educational technology. The research in this field is concentrated in "OTIC", that is the Centre for Educational Technological Innovations (Onderwijs Technologisch Innovatie Centrum).
OTIC is also engaged in running projects for other institutes. Here the Centre makes use of OTIC's knowledge and experience acquired in research and development.
The series of OTIC Research Reports consists of publications of the OTIC research projects and aims mainly at an audience of fellow researchers.

RESEARCH PROJECT 'PRIOR KNOWLEDGE'.

'The role of the Prior Knowledge State during the learning process of adult students in a modular educational system with applications in interactive electronic learning systems'.

This research project started from the idea that if the specific prior knowledge is taken into account, in a modular educational system, students will have the opportunity of following different learning paths in a more efficient way. The research is directed at a clear definition of the problems and their solutions.

CIP- gegevens koninklijke bibliotheek, Den Haag

Dochy, F.J.R.C.
Bouwens, M.R.J.

The construction of knowledge state
tests, knowledge profiles and the
measurement of the value added
F.J.R.C. Dochy, M.R.J. Bouwens.
-Heerlen: Open University,
Educational Technology Innovation Centre (OTIC)
- Ill. - (OTIC research report 27)
Met lit. opg., reg.
ISBN 90 358 0864-9- compl.

Reference: knowledge state tests, knowledge profiles

OTIC Research Reports are available at:

the Open University
secretariat OTIC
postbus 2960
6401 DL Heerlen
Telephone 045-762261 / 471

Educational Technology Innovation Centre

Open University

The construction of knowledge state tests, knowledge profiles and the measurement of the value added

OTIC Research Report 27

F.J.R.C. Dochy

M.R.J. Bouwens

Table of contents page

Introduction

There can be no doubt that the possession of prior knowledge facilitates learning (Dochy, 1988). In order to be able to study the precise mechanisms of this phenomenon and to draw conclusions of relevance to students and educators, an insight into the prior knowledge state must first be obtained. Aspects that spring to mind are the volume of knowledge, its accessibility, the type of knowledge, and the knowledge state profile within a specific domain.

In this connection, great attention must be paid to the generally recognized "requisite and ... foundational role played by domain knowledge" (Alexander and Judy, 1989). The authors explicitly refer to the importance of the Prior Knowledge States Project and the Knowledge Acquisition Support System Project in a domain-specific context. "Research in cognitive psychology during the past two decades has produced two undisputed findings about academic performance. First, those who know more about a particular domain generally understand and remember better than those with only limited background knowledge (e.g. Chi, 1985; Glaser, 1984). Second, those who monitor and regulate their cognitive processing appropriately during task performance do better than those who do not engage in such strategic processing (e.g. Flavell, 1981; Garner, 1987)".

This report deals with the question of obtaining insights into states of knowledge; more specifically it looks at knowledge state tests and knowledge state profiles. In the first part we introduce a few basic terms which are of importance in this context. The term 'evaluation' cannot be omitted here. The validity, reliability and usefulness of tests are then explained. In addition, the construction, use and analysis of tests are treated in more detail, with the aim of arriving at a concrete definition of the concepts involved. In the second part of the report, the various knowledge state tests are discussed and a picture is drawn from the results obtained in a domain-specific knowledge state test involving 536 participants.

1. Tests, test construction and the analysis of test results

1.1. Evaluating, measuring and marking

During the past twenty years much attention has been devoted in the literature to the problems associated with the construction of tests and their evaluation.
The main themes can be summarized as follows:
- a concept of evaluation that involves more than just the determination of academic results;
- the search for a pedagogically based method for constructing instruments to determine academic results (de Corte et al., 1976).

The evaluation of study results or the 'science of education measurement' (Mellenbergh, 1986) has developed into a specialist area with its own jargon during the last decades; terms such as summative evaluation, formative evaluation and product process evaluation are just a few examples.

The term 'evaluate' is frequently encountered in education measurement.
The evaluation of education is a process consisting of describing, collecting and presenting useful information with the aim of assessing (the value of) alternative decisions (van Os, 1987). Evaluation does not stand alone; it is directed towards decision making. In its most general and most far-reaching form, this can mean the continuation or termination of an activity; more specifically, doing away with a course of study or not. It must be possible to make a choice, or, to put it another way, the things we are dealing with must be theoretically or practically capable of being changed (Camstra, 1977).
On the one hand, 'information' implies more than facts or data; it covers everything that reduces the uncertainties associated with a choice of alternative decisions, and this includes opinions and impressions. On the other hand it is more limited; information must be structured with an eye to the choice that will subsequently have to be made. Useful information must be described in the light of the choice of alternative decisions with which we will be faced, and the limits (criteria, standards) imposed in this context. Collection and presentation are respectively concerned with the more technical part of the evaluation and the report of the decision taker (van Os, 1987).
However, the word 'evaluate' can indicate widely different activities. For example, "let's just evaluate things", "trainees must be carefully evaluated" and "the final evaluation was very difficult".
It is clear that the term 'evaluate' refers to an informal discussion in the first case, an assessment of behaviour in the second, and a written test in the third.
If we look at the etymological meaning of 'evaluate', we find that it is the same as 'value', 'to determine or fix the value of'.

If evaluation is to be properly understood, it is essential to appreciate the difference between the measurement aspect and the assessment aspect.

The terms evaluation and measurement are sometimes confused, or not sufficiently differentiated from one another.
A measurement of behaviour or performance is a quantitative description of these factors. Numbers are used to describe observations of empirical phenomena, in this case behaviour. Evaluation encompasses more than measurement and consequently includes measurement (de Corte et al., 1976):

- Evaluation includes not only a description of behaviour, but also an assessment of it as a function of a certain standard or criterion.
- With regard to description, evaluation includes both qualitative and quantitative description of behaviour.

* **Evaluation = describing + assessing behaviour**

  (quantitative = measuring, or qualitative)

* **Measuring = quantitatively describing observations associated with empirical phenomena, e.g. behaviour**

The attribution of a value obtained by means of an instrument involves a separate decision. A statement such as "I got 16 points" means nothing as it stands. To understand it, we need to know how many points could have been obtained, how many points others obtained, and whether the number of points is sufficient or not.
The value of a measurement is determined by the intrinsic significance we attach to it. However, 'attaching a value' to something is strongly dependent on the context and is therefore subject to individual (subjective) fluctuations.
Determining a value, therefore, must not be equated with taking a measurement; the information used to determine the value has a substantial bearing on the result.

Often, however, measuring the knowledge state of a student will not involve an element of assessment. The purpose is not usually to decide about pass or failure, but to gain an insight into the level of knowledge in order to use this as a basis for making certain choices. This means that more than numerical descriptions are required. The construction of a knowledge state profile is related to the domain-specific context and is thus dependent on content.

In the past, it was strongly emphasized that, to be fair to students, their (study) performance should be measured as objectively as possible. It is therefore very important that a great deal of care is devoted to choosing a suitable measuring procedure which will provide useful information. Useful information is information that fulfils scientific and practical criteria. We will now deal with the often cited criteria validity, reliability and usefulness in more detail.

1.2. Validity, reliability and usefulness

Validity

A test must, above all, be valid (Dousma, Horsten, 1980).
We define the validity of a test as: the degree to which the test measures that which it was intended to measure.
The validity of a test is impaired by the occurrence of systematic errors. The more substantial the error component, the less valid is the test, i.e. the correspondence between the result and the characteristics of the measured construct disappears.
This means that the test and the test questions as a whole have to fulfil two criteria, viz. balance and relevance.

Balance

In this context the question arises: is the number of questions per subject and level sufficient to achieve the specified aims, or are there too few questions to enable the material to be covered? In the language of study methodology, we speak of content invalidity. The content invalidity of an instrument is determined by assessing the degree to which the content of the items of which it is composed is representative, for the purpose of drawing conclusions in the given context. The content of an instrument cannot be calculated, it can only be determined on the basis of assessments, by experts or test persons, of the instrument and that which it is intended to measure.

Relevance

Relevance is the degree to which the test questions measure that which the questions are meant to measure. In education measurement, use is often made of the correspondence of test items and objectives, or of the expertise of specialists. Relevance can also be operationalized in terms of criterion and/or construct validity. The criterion or the theoretical construct are then the relative measure for relevance. Criterion validity is the relationship between a test score and a series of test scores with one or more criterion variables (Meerling, 1981). The greater the degree of correlation between an instrument and an external criterion, the greater is the validity of the instrument. One problem here is the choice of criterion. According to Carmines and Zeller (1979), there is no 'single criterion-related validity coefficient'. There are as many validity coefficients as there are criteria.
In the case of construct validity, a theory exists over the latent construct. This theory describes how the latent construct functions under certain circumstances. In the theory, points of contact are found for the construction of the test, viz. the relationships between variables or concepts.
An important condition for a good test has been fulfilled when the validity of the test is assured.

b. Reliability

Besides being valid, a test must be as reliable and useful as possible.
A test is reliable if the same results are always obtained from repeated measurements of the same object under the same conditions.
Reliability is understood as: the degree to which the test gives consistent results irrespective of its aim
This means that the results from a certain group of students must correspond to the results obtained from the same students in other tests on the same subject at the same level.
The following criteria are of importance for the reliability of the test.

Objectivity: are the questions so clearly formulated, and are the possible answers so unambiguous, that an assessor who observes the rules cannot influence the student's score?

Specificity: are the questions so formulated that they can only be answered satisfactorily by persons who have sufficiently mastered the material?

Difficulty: can the test and the questions as a whole differentiate between students who have a good grasp of the material and those with a less adequate grasp?

Test length: does the test comprise enough questions to exclude the possibility of chance successes?

c. Usefulness

The criteria with regard to the usefulness of tests are related to the degree to which the instrument must fulfil a number of requirements, some of them practical. The criteria to be observed here are as follows:

Efficiency: how much of the tester's limited time do the construction and the scoring of the test require; how much of the student's time is needed to complete the answers; how much information does the test supply in relationship to the (short) time that the student spends answering the questions?

Fairness: is the test so constructed, and are the results processed in such a fashion, that each student has an equally good chance of demonstrating his or her ability and progress in the field covered by the test?

Time available: is the test of the right length relative to the time available, so that the speed at which the student works has no influence (positive or negative) on the result?

1.3. Evaluation of objects, especially learning results

There are two ways of approaching the evaluation of an object (for example an instrument, an activity, a process).
- The object can be approached via its result as a function of a specific aim, enabling conclusions to be drawn about its yield.
- The object can be evaluated in the light of specific criteria that it has to fulfil.

In a sense, these two approaches are complementary. After all, if a certain object gives an inadequate yield, it can be examined more closely with a view to explaining its poor yield (de Corte, 1976).

Evaluation activities in education can be directed at a number of different objects. Most attention is paid to the performance of the students, the academic results. Academic results are simply the output of the learning process.
Evaluation activities that are concentrated on academic results are classed together under the term 'product evaluation'. Evaluation that is concentrated on education and the elements of which it is composed (subject matter, educational resources) is termed 'process evaluation'.
The assertion that academic results are the product of the learning process is only partially true. Academic results are in part the product of prior knowledge, and we have mentioned above that prior knowledge facilitates the learning process. The yield of the learning process must thus be described as the 'value added'.

The subject 'academic results as an object of evaluation' is central to the following exposition, whereby the learning process is understood as the acquisition of new behavioral possibilities. Behaviour must be understood in a broad sense here, and ranges from simply possessing knowledge to solving a problem.
Tests of prior knowledge states are directed at the learning results obtained at an earlier stage, which make up the 'knowledge state' of the student before he begins a subsequent learning process.

Academic results can be classified in different ways. Most classification systems for academic aims differentiate between three main categories, namely:
- cognitive aims: aims that are associated with intelligent functioning;
- dynamic-affective aims: aims that are associated with an emotional aspect and/or a degree of acceptance or rejection;
- psychomotor aims: aims where the emphasis is on motoric operations and sensory perception.

For further typification and operationalization of these categories, see Dochy and van Luyk (1987, p. 43).

1.4. The relationship between aims and tests

We have already stated that validity is a prerequisite for responsible testing. The various functions that tests can have in the education sphere, and the consequences of the associated decisions to be taken, clearly underline the great importance of valid testing. The relationship between the aims and the content of a test is of great importance in ensuring its validity. After all, the aims reveal what the creator of a specific academic course wanted to achieve with it, and the test provides information about the degree to which the aim was achieved. This means that not only the subject-specific content but also the desired level of command should be expressed in the test questions. Furthermore, the given aims must be equally represented in the sum of the test questions.

We shall restrict ourselves here to the cognitive aims, not least because aims of a cognitive nature are (still) pursued in the education sphere, and above all because tests are not usually the ideal instrument for evaluating affective and psychomotoric aims (Dochy and van Luyk, 1987).
The cognitive domain is the most intensely studied, and a number of authors (De Corte et al., 1981; De Block, 1975; Gagne, 1977) have developed classification systems that differ in details but can all be traced back to the following basic subdivision of behaviour levels:
1. knowing;
2. understanding/insight;
3. applying rules/procedures;
4. problem solving.

In a knowledge state test, the material that has been learned must be recognized or reproduced. In a test of understanding, the objective is to discover, by means of questions requiring examples, conclusions, comparisons, etc., whether the material that has been learned has really been understood.
Here we are dealing with so-called declarative knowledge: the knowledge of facts, the meaning of symbols, the concepts and principles of a specific specialist field (Dochy, 1988).

The ability to apply knowledge and to use it to solve problems is tested by setting questions that require rules or procedures to be applied.
This knowledge of actions, manipulations and operations is also referred to as procedural knowledge (Dochy, 1988).

The essential difference between the two types of knowledge is that procedural knowledge is directly associated with an action or operation, while declarative knowledge needs to be interpreted before it can be translated into action (Messick, 1984).

1.5. Functions of tests

Tests are suitable instruments for collecting information on the results of learning processes. This information is important for both student and teacher, and its consequences are dependent on the aim and function of the particular test.

We distinguish between three functions of tests: the predictive, the selective and the diagnostic.

The aim of predictive tests is the forecasting of future academic and study success. This involves testing aims with a predictive value for later activities. Of course, the content of a predictive test must be related to factors and characteristics that have been proved to have a predictive value. In higher education, a satisfactory assessment given at the end of a specific course is usually regarded as a requisite for admission to further courses, so that predictive and selective usage coincide (Dousma, Horsten, 1980).

The aim of selective tests is to enable a decision to be made (at the end of a course of study) about whether, and to what degree, a student has achieved the aims that were set. Students who measure up to a given standard can proceed further, failure to measure up means that the student must first repeat the course in part or in full before he can proceed, or that he must seek a different course of education.
This form of evaluation is also referred to as 'summative evaluation' (Bloom, Madaus & Hastings, 1981).
Summative evaluation serves to enable decisions to be made about the further educational path that students will undertake. It is clear that, because of the great importance of this form of testing for the student, good instruments (tests) and carefully defined standards must always be applied.

An insight into the learning results is, however, also of importance for the taking of interim measures that affect the academic situation. The purpose of using tests as a diagnostic instrument is to obtain information on the strong and weak points of the individual student, a group of students and/or the academic learning process. In this context the weak points or shortcomings are especially important. The ultimate aim is to remedy them, either by supplementary instruction or by adapting the learning process. This method of collecting information is referred to as 'formative evaluation'. It is only of use if the course of study has not yet finished and the students are able to adapt their study behaviour. In this sense, obtaining information over prior knowledge states and current knowledge states can prove especially worthwhile.

1.6 Tests; construction and choice of form

1.6.1. Construction

A number of factors must be considered before a test question is formulated and a test is constructed. These include the purpose for which the test results will be used, the correlation between the questions and the aims of the test, and the balance between content and level. In addition, the test must satisfy, as far as possible, a number of additional requirements in connection with measurement aspects. All of these considerations form the basis for selecting the type of test, formulating the questions, and constructing the test (Dousma, 1980).

When a test is being constructed the following steps must be followed in order to transform implicit or global aims into concrete topics susceptible to testing.

1. Define the aims.
In reality the aims are known when a course of study commences. The following step is therefore of greater importance for constructing a test.

2. Define the aims to be tested.
It is clear that, ideally, all aims that are being pursued should be tested. Depending on the overall number of aims and the extent of the syllabus to be tested, the available time and the type of test, a choice may have to be made between the aims, with regard to both content and level.

3. Make a detailed list of the topics to be tested, based on the aims.
In this step the aims are transformed into concrete topics derived from the relevant course of study.

4. Place the topics in a specification table and plan the test with regard to content and level.
At this stage, preparation of the test is complete, in so far as guarantees have been included to ensure a balanced distribution of the questions with regard to content and level.

On the basis of the above mentioned criteria, the following general rules of thumb must be borne in mind when the questions are formulated and the test is being constructed:

VALIDITY:

A. Care must be taken that the content of the questions is relevant to what the test is intended to measure.
B. Checks must be made to ensure that the aims of the test are represented as far as possible, with regard to both content and level.

RELIABILITY:

C. The questions should be formulated in consultation with other experts.
D. The questions should be formulated exactly; care should be taken that the test questions are independent of each other.
E. Care should be taken that the test contains an adequate number of questions.

USEFULNESS

F. The questions should be formulated as objectively and concisely as possible; attention should be paid to the layout.
G. A question should include all information that is necessary to enable it to be answered.
H. The length of the test should be adapted to the time available.

If these rules of thumb are borne in mind during the formulation of the test questions and the construction of the test, then most of the criteria will be fulfilled. Only afterwards, when the test results have been analyzed, can conclusions be drawn with regard to the differentiating capacity of the test and its reliability. The estimated degree of difficulty of the test for the group of students affected can also be verified (Dousma, 1980).

1.6.2 Types of tests

A distinction must be drawn between different types of tests. Tests can differ in:
- the type of question used;
- the form in which they are taken (e.g. written or oral);
- the measure of uniformity (is every student given the same questions?);
- the way they are taken (individually/collectively).

If tests are classified according to the type of question used, we can distinguish between two sorts of tests, viz.: tests in which the student himself must formulate the answer (so-called open tests), and tests in which the student must choose between a number of precoded possible answers (so-called closed tests).
We shall restrict ourselves to closed-tests.
The advantage of this type of test is that it can be completed more quickly, and can be marked more quickly and objectively. This type is therefore the most suitable for use with large numbers of students and for integration in electronic education systems.

Closed tests

Closed tests (also known as study tests) are tests that consist of questions accompanied by a number of precoded answers; the student must decide which of the answers is correct. Another characteristic is that they can be marked objectively. In general, a test can be objectively marked if the marker (a person or a system) cannot exert any influence on a student's score (always assuming that the marker adheres to the rules governing the marking of the test). Any given person, or even a computer, can determine a student's score in a closed test. If we list the main advantages and disadvantages of closed tests, we arrive at the following.

Arguments in favour of the use of closed tests:
- Because the required answer is unambiguous and is the same for all students, factors that may have a distorting influence on the test results are excluded.
- An almost complete random test of the material can be made by structuring the test so that it contains a large number of short questions. The measurement is more reliable and the study material can be tested more specifically than in open tests. In addition, each level (knowledge, understanding, application and problem-solving) can be tested separately.
- The student's score is determined objectively.
- When large groups are involved, and when used repeatedly with smaller numbers of students, closed tests soon result in time savings; a relatively large amount of time must be devoted to constructing the tests, whereas corrections can be carried out very quickly (Dousma, 1980).

Arguments against the use of closed tests are:
- Problem-solving questions are not always suitable for closed tests.
- There is a chance of guessing the correct answer. This must be borne in mind when the scores are interpreted and grades awarded.
- The time needed to construct a closed test is relatively long in comparison with that needed for other types of tests (Dousma, 1980).

In particular, two types of closed test are often used, namely:
-yes/no tests;
-multiple-choice tests.

YES/NO TESTS

In yes/no tests the student must indicate whether a statement is true or false (true/false, yes/no).

Example:

The total costs of the sand-shipping company "Leegte" can be described with a continuous and differentiable function. The average total costs of all companies on the sand-shipping sector fall continuously to a price of fl.1- per tonkilometre sand at a production of 10 million tonkilometres sand, and then begin to rise.

true/false the marginal costs of sand production are equal to the average costs at a production of 10 million tonkilometres.

The yes/no question is easy to set. However, it must always be borne in mind that such assertions must be absolutely true or false, without any possibility of doubt.
One disadvantage of this type of question is that there is a 50% chance of guessing the correct answer. If a valid and reasonably reliable test is required, a test of this type must consist of a large number of questions. The actual number of questions depends on the quality of the questions. Wijnen (1973) argues that a test containing 100 questions will generally provide more information than one containing 50 questions. In the same article Wijnen suggests that questions in this type of test be directed not at evident truths but at authors' opinions and the conclusions to be drawn from a study. If questions are taken over literally from text books, there is a danger that the test will place too high a premium on memory.

MULTIPLE-CHOICE TESTS

Multiple-choice tests consist of questions with a number of given alternative answers, of which the best, the most correct, must be chosen.

For example:

The term consumer sovereignty is justified if:
A. consumers can choose freely from the range of goods on offer;
B. the volume and variety of goods produced are ultimately determined by consumer preferences;
C. consumer groups have their own representatives in parliament;
D. consumers receive an income that is sufficient for their needs.

The number of alternative choices per question can vary, with two as the minimum number of alternatives.
Multiple-choice questions with a choice of four possible answers tend to be used most in practice. We shall therefore restrict ourselves to this type.
The advantages of multiple-choice questions over yes/no questions are:
- the chance of guessing the correct answer is lower (0.25 with four alternatives);
- the answer required need not be absolutely true, if only the best alternative has to be chosen.

A major disadvantage is that the formulation of multiple-choice questions is more demanding than the formulation of other types of question (Dousma, 1980).

1.7. Analyzing tests

Test results are analyzed and assessed with the aim of determining:
- to what degree the student has achieved the set aims;
- whether the test used is valid and reliable;
- whether the education process can be optimized.

The analysis of test results gives us information of value in answering questions about the quality of the test. The degree to which analysis is possible depends on the type of test questions used. In general, the results of open tests are not as easy to analyze as the results of a multiple-choice test.
An item analysis can help in determining whether the tested aims were realized and how the separate items functioned.

The analysis of an item can provide answers to the following questions:

How difficult was the item for this group of students?
How attractive were the false answers for the students?
What is the discriminating capacity of the item?

In order to answer these questions, we need a number of parameters - values of certain characteristics of the item.

A. The degree of difficulty

The first thing to determine is how many students answered each item correctly. The p-value of each item can then be calculated, and is a measure of whether a question was difficult or easy for this group of students.
The p-value gives the proportion of students that answered the question correctly:

$$\text{p-value} = \frac{\text{number of students who gave the correct answer}}{\text{total number of students}}$$

In other words, the higher the p-value, the easier the item.

The significance of the p-value speaks for itself. In this connection, attention must be paid to the following:
- The p-value is an indicator of the difficulty of the item and the level of the students. The item can change as the test is repeated with various groups.
- The p-value is also an indicator of the success or failure of the education, the education method, etc. It is an indication of the degree to which the students have achieved the aims to be measured.

The average p-value of the test as a whole is obtained by adding together the p-values of each individual item and dividing the result by the number of items.
On the one hand, the p-value provides information on the relationship between the capacities of the students and the difficulty of an item; on the other hand it also provides information on the number of differentiations that can be made between the persons in the group. If an item has a p-value of 0 or 1, the item cannot be used to distinguish between students in this specific group who have a good command of the material and those whose command of the material is less good. For selective tests, an attempt should be made to avoid items that produce p-values of 0 or 1 in the test.

B. The attraction value of the decoys

The a-values of the decoys can be determined in the same way as the p-values. For a four-choice question, three a-values and one p-value must be calculated.

$$a = \frac{\text{number of students who chose a decoy}}{\text{total number of students}}$$

If the p-value is high, the a-value will be low. The reverse is also true. The a-value of an alternative must be lower than the p-value of that item.
If this is not the case, then clearly something is wrong with the item, or too little or no attention has been paid to this topic during the relevant education process. An a-value must not be approximately or exactly equal to 0, because this would mean that it was not fulfilling its purpose. In addition, the ideal to be aimed at is that the alternatives for any given item all have almost equal a-values, in which case the alternatives all function as decoys of equivalent value (Dousma, 1980).

C. The differentiating capacity of an item

One aim of a test, and thus of an item, is to enable a differentiation to be drawn between students with certain skills and students who do not command certain skills. After the test has been completed, the total score and the scores per item are available for this purpose.
An item that elicits a correct answer from good students more frequently than from poor students discriminates well - it has differentiating capacity.
In order to determine the differentiating power of an item, the scores for an item are compared with the scores for the whole test. The item-rest correlation

r_ir

is an often used index for expressing this relationship. It gives the correlation between the scores for an item and the scores for the test, excluding the item in question. The measure of correlation between the two is expressed as a correlation coefficient.
A simpler method, requiring less "manual labour" is often used. This involves calculating

r_it

the correlation between the scores for an item and the total score, including the item in question.
This is a slightly less accurate measure, because the total test performance already includes the performance for the one item, and as a result the correlation coefficient is higher than it would otherwise be.

What significance does the r_it value have for us?

In order to be able to answer this question, we must first examine the factors that influence this discrimination index.
1. The degree to which the test and the item measure the same thing.
r_it is an index of the degree of correlation.
2. The quality of the item. The question must be formulated as clearly and unambiguously as possible. The correct answer must be undisputed and the decoys must be incorrect.
3. The spread of the item; the higher the p-value of the item, the smaller the spread.
4. The spread within the total group. If the difference between the students' performances is slight, the discriminating power of the test and the items will be low. For homogeneous groups, therefore, r_it values will be lower than for extremely heterogeneous groups.

What is an acceptable r_it value?

In general, this is difficult to say, but certainly the value of r_it should always be positive, and if possible greater than .20.

A lower or a higher value of r_it may indicate:

an error in the test or the item, an incorrect key, a question that lies outside of the study material, an extremely difficult question, a trick question, or a subjective question.
As well as the r_it values, the r_at values can be calculated or plotted on a graph.
The r_at value is the correlation between a decoy and the score for the whole test. It should always be negative.
If one of the r_at values (a four-choice question always has three r_at values) is positive, then:
the decoy is possibly correct, the decoy is directed at a misunderstanding that is so widespread that even good students choose it, the decoy is a trick, the decoy is too difficult.
The following formula is used to calculate both r_it and r_ir, where X is the item score and Y the total score or the rest score.

$$r_it = \frac{\bar{Y}_g - \bar{Y}_f}{S_y} pq$$

where

$\bar{Y}_g$ = average test score of the students who answered the item correctly

$\bar{Y}_f$ = average score of the students who answered the item wrongly
S_y = standard deviation of the test scores
p = p-value of the item
q = $1 - p$ $(= p_f)$
This is termed point-biserial correlation: the correlation between a dichotomy and a continuous variable.
The item-rest correlation can be determined with the help of the formula:

$$r_it = \frac{r_itS_Y - pq}{(S + pq - 2r_itS_Y pq)}$$

D. The discrimination index method (D-method)

Although the r_it and r_at values can be represented graphically or numerically, this involves a vast amount of calculation. In any case, r_it is only of significance in connection with an item if the group includes more than 100 students.
An alternative indication of discrimination power involves the simple calculation of a D-index. The D-index of an item is determined as follows:

- select two sub-groups, preferably of the same size, out of the total student group; one group must contain the students with the highest test scores, the other must contain the students with the lowest test scores;
- determine the number of students in each group who answered the relevant item correctly;
- convert both numbers to proportions (divide by the total number in the group);
- determine the D-index by deducting the proportion of correct answers in the group with the lowest scores from the proportion of correct answers in the group with the highest scores (Dousma, 1980).

Studies (e.g. Ebel, 1972) have shown that the best distribution of the total group is: 27% of the highest scores and 27% of the lowest scores, so that no account is taken of 46% of the middling scores.
The answers of the two groups to the diverse items for which the D-value is to be calculated are entered in a matrix. When the matrix has been completed, the following values are calculated:

nH and nL: number of candidates from the sub-groups with the highest and the lowest scores respectively.
nHJ and nLJ: number of candidates, from the sub-groups with the highest and the lowest scores respectively, who answered the item correctly.

The D-value is calculated with the help of the formula:

$$D = \frac{nHJ}{nH} - \frac{nLJ}{nL}$$

A D-value can vary from -1 to +1. D = +1 means that the item discriminates perfectly between 'good' and 'poor' students; D = 0 means that the item makes no distinction whatsoever; D = -1 means that the item discriminates completely wrongly between 'good' and 'poor' students (Dousma, 1980).

1.8. Giving grades

We have seen above how the answers to a test can be processed. Test analysis and item analysis give an insight into the quality of the test as a whole, into the value that can be placed on students' test scores, and into the quality of separate items.
In most cases there is one further step - the assignment of grades (Os, 1987). Grades are assigned as a method of expressing a value judgement on the test scores.
In the case of knowledge state tests, this will not occur very often. The coupling of the test with academic aims or domain-specific knowledge gives the student direct information on his knowledge state profile or his current situation. The comparison with the situation towards which he is striving gives him the input needed to undertake direct academic activities or to formulate specifically targeted academic tasks (Koper, 1989).

2. The value-added between different knowledge state tests.

"Value added," a term frequently used to describe students' gains on tests of knowledge and skill, had attracted a great deal of attention in order to demonstrate that universities are educating their undergraduates.

The growing popularity of this value added approach comes partly from the public's and legislatures' desire for evidence that students really learn enough in college to warrant the expense of higher education, and partly from colleges' own desires to evaluate their students' gains in (as opposed to absolute levels of) knowledge and skill.

In America, particularly, attention has been given to the measurement of the "value-added" to an individual as a result of undergoing education. Astin (1982) writes:
The basic argument underlying the value-added approach is that true quality resides in the institution's ability to affect its students favourably, to make a positive difference in their intellectual and personal development.'

The importance of the ability to measure value-added lies in the information which it could provide in respect of the relationship between inputs and outputs, i.e. the efficiency of the education process. The more efficient institutions produce more value-added at the same or a lower cost. In this sense the efficiency of one institution relative to any other can be crudely measured by the ratio of average value-added to average cost. The higher the ratio the more efficient the institution.
Measuring value-added is also important because it should allow the nature of the returns to scale in teaching to be explored more fully. In order to be able to allocate resources efficiently information is required on how outputs change in response to marginal changes in inputs. At present we do not have a measure of the output of the teaching process and this might, in principle, be provided by the value-added measure.

What is Value Added? The concept originated in economics and refers to the value added at each stage of the processing of a raw material or the production an distribution of a commodity. It is usually calculated as the difference between the cost of raw materials, energy, etc. used to produce a product and the price of the product (Greenwald, 1983). This general idea is now being applied to higher education, with the "product" being the students and the "value" being the knowledge and skill students possess or their "knowledge states". The value (or knowledge and skills of students at the beginning of their college education) and their knowledge and skills at a later time are compared. The difference between knowledge states is held to be the value added by higher education.
However, this simple economic analogy makes many academics nervous. They object to the notion of thinking of students as products, and are particularly sceptical about attempts to quantify educational growth. As Fincher (1985) notes, they are not disposed even to think in such terms:" . . . their values, preferences and incentives have all been channelled in other directions." They are also given pause by the stated or implicit rationale that value-added information is basically

for external groups that want to hold the institution accountable rather than for the internal improvement of teaching and learning. Can value added be used in a more constructive way? The answer to this question lies in its definition and the way this definition is implemented.

Between two knowledge states (the moments of entry and exit), the student experiences a learning process. The intention underlying that process is to modify the student as received at the point of entry. Clearly, there are different viewpoints as to the kind of modifications in the student the process of higher education should bring about. Nevertheless, something takes place between the states of entry and exit. Something is going on in the "black box" which is intended to change the student in various ways.

In practice, the meaning of value added varies considerably, from the general, such as "the institution's ability to affect its students favourably, to make a difference in their intellectual and personal development" (Astin, 1982), to rather specific blueprints that specify tests and the scores students need to obtain (Northeast Missouri State University, 1984). Taylor (1985) describes the purpose of the value-added assessment at NMSU as to "measure the gains in knowledge, skills and personal development within each individual". He identifies two driving forces behind its introduction. First, where the funding of universities in the USA has been based upon

measures of throughput of students, the incentive for universities to do well against these quantitative measures has detracted from what should be an emphasis on quality. Second, there has been growing dissatisfaction with the traditional methods of assessing student performance and a greater public desire to measure the extent to which education results in an improvement in individuals in broader terms.
However, the common these definitions is the idea of "gain," which is usually assessed by administering a test at one point, typically as students enter university, and the same or a related criterion test at a later point. Although value added could refer to any criterion of interest, including personality, values, or creativity, it most commonly refers to academic achievement (e.g., increased scores on a test of mathematics or better scores than expected on knowledge of a field, such as economics. This latter and most popular conception leads to several kinds of issues.

The Maze of Measurement.
Value added assumes that we can measure change. The measurement of change is a very tricky and difficult issue, involving problems of both measurement and statistical design (Goldstein, 1983; Wood, 1986). The first problem is to find measures that will be reliable for both the initial measurement and the follow up measurement, but which will yet be sensitive to students" growth in knowledge and skill (Carver, 1974; Kessler and Greenberg, 1981). Tests need to be reliable to provide accurate estimates of students' knowledge and skills.

That is, a measure should give approximately the same estimate of a student's level of academic skills from one day to the next. However, the key is that the measure must also be sensitive to gain and to the influence of the learning process on the magnitude of gain. For example, students' height is an extremely reliable measure. To use gain in height as a criterion to assess the "value added" of a university's cathering services would be absurd, even though height can be measured reliably and its measurement is sensitive to gain.

This hypothetical example illustrates one of the dilemmas in measurement of student characteristics. The more tests assess
general characteristics, the less sensitive they are to change due to educational programs. That is, the tests become so general as to assess relatively stable characteristics of students. In the cognitive area, the more general tests border on measures of general intelligence. For example, the Watson-Glaser Critical Thinking Appraisal correlates with some measures of intelligence as highly as the two forms of the test correlate with each other, which leads to the possibility that it is measuring intelligence rather than something distinct.
The challenge, then, is how to use or develop measures that are specific enough to be sensitive to students' in knowledge states gain but general enough to rise above the trivial. Therefore, domainspecific or even contents specific tests seem necessary. Note, too, that the tests need to be reliable at each point they are administered.

The Vicissitudes of Change: Statistical Design Issues

The statistical issues involved in assessing change have been debated for many years (e.g., Harris, 1963; Nuttall, 1986). Recommendations have ranged from simple change scores (final score minus initial score) to the use of latent trait models (see Traub and Wolfe, 1981). The problem, as described by Fincher (1985), is that the initial and the later test scores include both random error and test specific variance, as well as a shared variance. In other words, the two tests measure both the feature common to both tests, and different features unique to each test. The test scores also reflect the random, unsystematic aspects of students' responses to the test. As shown in Figure 0, when difference scores are used in value-added studies, part of the common variance is subtracted out, leaving a large share of the difference to be determined by the unique variances and error, as well as real change in the student.

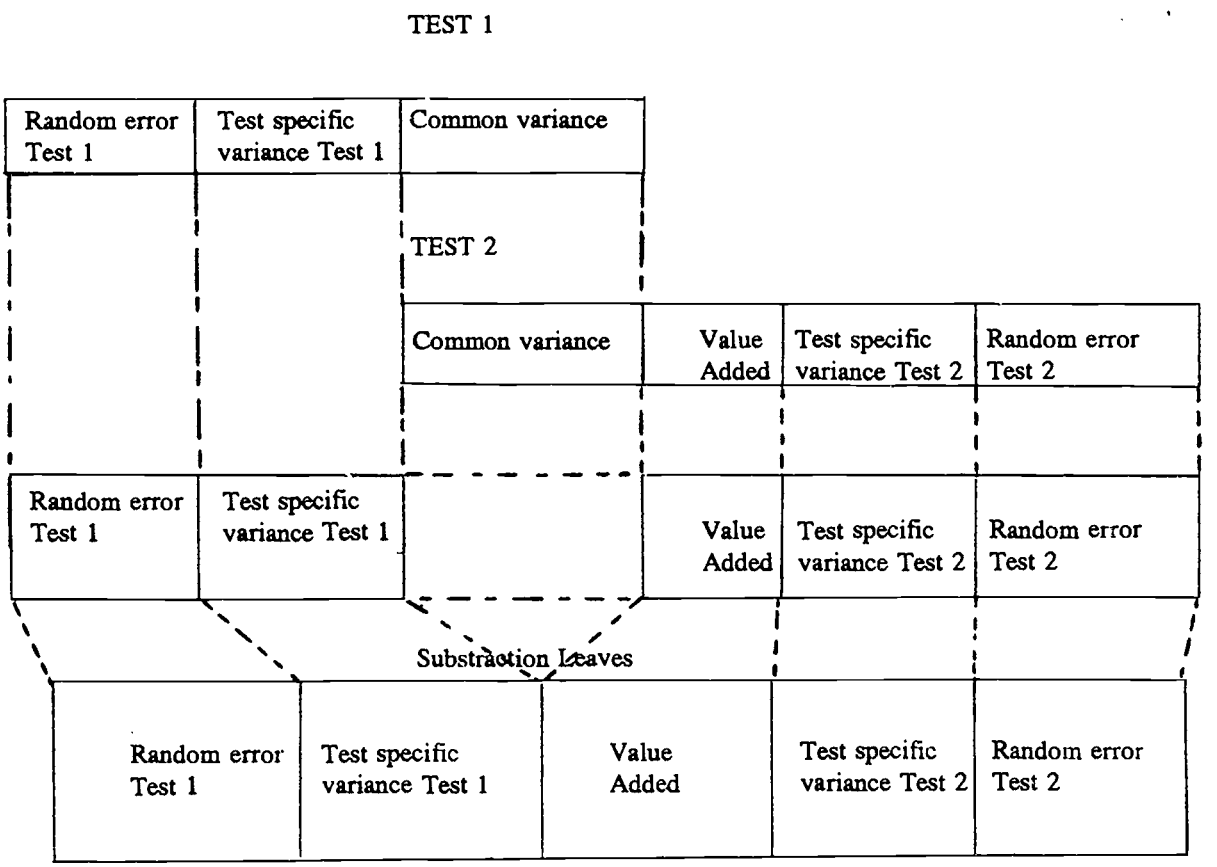
For example, let us say that beginning first year students take a test concerning accountancy. Then, at the end of the year, they take an alternate form of the same test. The scores on the first test are subtracted from those on the second test to provide a measure of "gain." Consider that any test score includes error that is unrelated to the variable being assessed (e.g., some items may be rather poorly written, testing conditions may be less than ideal, some students may be tired, distracted or unmotivated, etc.).

All of these add to the randomness of the score. Likewise, each test samples a part of behaviour or knowledge, so that its coverage of the content area in question will vary. Therefore, scores will vary from the first test to the second, simply because each test is, in fact, somewhat different. When the scores on the first test are subtracted from the second to yield a "gain" score, what is left over is partly due to real changes in students', knowledge states on accountancy, partly due to expected errors in the tests, and partly due to the fact that the tests are not, in fact, exactly the same. Thus, it is hard to know how much faith to place in the change score.

Residual Scores

These difficulties of interpretation are increased when the first and second tests are different. For example, a domain-economy specific test as the initial test, and a accountancy test as the second one. Since the tests are clearly different, a frequently used method is to use the correlation between the two tests to produce an expected or predicted score for the second test, based on performance on the first test. For example, using hypothetical data, and based on a hypothetical correlation, a student scoring 450 on the first test might be predicted to obtain a score of 150 on the second; a student scoring 600 on the first might be expected to score 170 on the second. The predicted score for students is subtracted from the actual score. The difference is known as the "residual score." If the student scores higher than predicted, the residual is positive; if the student scores lower than predicted, the residual is negative. The average residuals for students with different curricular experiences could be compared to see if students with those experiences do better or worse than expected. For example, the average residual scores of students majoring in general economics could be compared with those of students in business economics. any difference in residual scores would be considered differences in the value added in the curriculum of those academic divisions.

The potential problems of simple change scores are compounded when this procedure is employed. The tests are different; and their unique characteristics are even more important than in the case of simple change scores. The common variance is less, and their specific variances considerably larger. Predictions based on the residual method are also imperfect; even a correlation of .60 will yield many predictions that are far off the mark. The average residual, then, can be questioned in terms of reliability-- that is, there is a possibility the residual could be due to chance. Residual scores have also been criticized because they are not measures of change per se, and because performance is judged on deviations from an average prediction. Therefore, as many students will usually be below average as above, and if one program appears to be more effective, another will appear to be less effective, even if both are doing a decent job.



Random error = unsystematic aspects of students' responses to the test

Specific = test specific variance

Common = shared variance

Fig. 0: The nature of value added

An additional consideration in analyses of change is that test scores can be influenced by factors other than performance on the variables of concern. For example, Astin and Panos (1969) found that performance on tests of mastery of the humanities, natural sciences, and social sciences was predicted best by the National Merit Scholarship Qualifying Test, but was also predicted by background characteristics (e.g., gender, initial career choice, parents' social class, etc.). The question for those analyzing predicted versus actual scores is whether to use all the variables that predict performance on the second test. To do so would no doubt increase the power of prediction. But some of those variables do not help us explain students' real gains; and others might be politically too delicate for an institution to include in its analyses.

Ceiling Effects and Non-completion rates

Another troublesome conceptual and technical issue is the lack of independence of students' performance on the first test and the later test. If this correlation is positive, it suggests that students with higher scores are gaining more than students with lower scores. If it is negative, it suggests that lower-scoring students gain more than higher-scoring students. Either result may be disturbing to an institution's faculty and officials. The former raises the possibility that the institution's programs are short-changing its less able students, who may not score high for a variety of personal and social reasons. The latter raises the possibility that the institution is short-changing its more able students, a considerable concern in times when institutions are searching for quality. This latter possibility is related to an artifact of many assessment designs: that when a criterion test is relatively easy, there is a "ceiling effect," which limits the amount of "growth" that can be shown.

Students who often gain the most on the tests are those who have the most to gain (i.e. , the academically ill-prepared). Thus, if an institute or program wished to show the most value added, it should admit the most poorly prepared students (i.e., those with poor high school grades, poor admissions test scores, and poor course preparation), as has been empirically demonstrated by Banta et al. (1987).

In analyzing change in the learning of university populations there is the additional problem of non-completion. That is, it may be that only the students who do, in fact, gain, will still be attending the institution when the final assessment is made. The institution will then appear to be effective, simply because the students who would supply counter-evidence are no longer there.

Another very difficult problem is the attribution of gain to the institution's programs, when it may be due to maturation, the general college environment, or simply to the fact of college attendance. In fact, it appears that students "gain" at about the same rate, wherever they are attending (Pascarella, 1985). This problem of attribution is especially vexing when the subject of assessment is "general education."

The problem is smaller for domain-specific matters.

In this same context, the level of difficulty should also be examined carefully, for if the test is too difficult or too easy, the difference in scores for students with different educational experiences will be due more to chance than if the tests were of appropriate difficulty.

A related issue is the degree of specificity of the assessment. An assessment that provides detailed information about each learning objective will be much more useful than a global assessment.

The research reviewed by Pascarella (1985) and Nucci and Pascarella (1987) leads to reservations about the use of value added at the institutional level. It is not clear whether a gain in test scores would be attributable to students' maturation, the experience of attending college, the overall college experience or the particular course or program the students had taken. It seems plausible, however, that if an institution or program has very explicit educational goals, the gain could be attributed to the institution or program.

Thus, although a value-added assessment strategy may have some utility as a way of examining the educational effectiveness of programs for the purpose of internally generated improvements, it must be done very carefully, keeping the points discussed in this chapter in mind.

The diagnostic potential of the programme was illustrated at NMSU in 1979. McClain et al. (1986) approaches which could be taken to improve the mathematical skills of their students. The recomm ndation became effective in 1979/1980 with the result that in each of the subsequent years,test results have shown improvement.

Taylor (1985) believes that there have been two important side effects of the introduction of the value-added programme. Faculties now take the view that 'students come first'. The emphasis has moved from attention to quantity to attention to both quantity and quality. It has allowed demonstrate to funding authorities and the general public that the education process is contributing value to individuals.
If executed thoughtfully, value added-assistment has some potential for the improvement of instruction at the program level. It is much less appropriate or useful at the institutional level of analysis.

3. Knowledge state tests

A search of the literature was made, to determine how prior knowledge is measured. It is often measured by means of tests which have not been specifically developed for this purpose (Dochy, 1988). In addition, different sorts of prior knowledge have been measured. In developing our tests, we have tried to make a clear distinction between the different sorts of prior knowledge, and have developed a test for each individual sort: subject-specific knowledge state tests, an optimal requisite knowledge test (OR test) and a domain-specific knowledge state test.

3.1 Description of the PKS tests

A. Subject-specific prior knowledge state tests

The term 'subject-specific prior knowledge state test' is used in the study to identify a test which is of direct relevance for the material to be studied (in this case the study modules or blocks to be studied for the Economics and Finance course), or which is concerned with specific knowledge that is required, such as mathematics.
Four experts (economists) from the University of Limburg screened the items from the block tests of the UL to ascertain whether they were of direct relevance with regard to the aims of blocks 3 and 4 of the Economics and Finance course. In addition, the relevance of the selected items was assessed once more by an expert from the OU. Finally, a representative number of these items were collected together to form a test (in agreement with the projected aims).
The study was made with subject-specific prior knowledge state tests for the modules "The supply of goods: the costs" and "The supply of goods: producer behaviour and market forms" These tests each consisted of 12 items of the four-choice type. The choice was partially influenced by the possibility of converting the existing material (tests) to similar prior knowledge state tests.
Obviously the subject-specific prior knowledge state test for economics examines what the student is already capable of and what he knows with regard to the course to be studied.

In addition, a prior knowledge state test for mathematics was constructed at VWO (pre-university science education) level. This test comprises 28 multiple-choice questions and is based on the self-test at beginners level in mathematics, the validity and reliability of which were studied by Dyck (1976).

B. Optimal requisite knowledge test

Prior knowledge, however, is much broader than knowledge of the content of a subject in the narrow sense. It also includes optimal requisite prior knowledge. This is the knowledge that a student must possess if he is to commence his course of study in optimal circumstances. The OR-PKS test was constructed by asking all of the general economists from the Economics Product Group and ten economists from the Economics Faculty of the University of Limburg to describe the optimal prior knowledge needed to study the above mentioned modules. In addition, they were asked to present concrete themes on this knowledge and to name articles and books where it is treated. On the basis of the answers, a test with 8 multiple-choice questions was constructed.

C. Domain-specific prior knowledge

It can be assumed that the learning process is also influenced by prior knowledge that is broader than strict subject-specific prior knowledge. For this reason, a domain-specific test was developed to cover the whole scientific field or study-material field. In this case with which we are concerned, the domain is economics. This test, which is aimed at the whole study-material domain, is set at the level which should be attained by the end

of the general first year university course. The heterogeneity of the test population (or student population) is so great that a test at beginners level would not be able to bring to light all of the differences between the students. After all, it can be assumed that students with years of experience in, for instance, the financial sector, or students who have obtained other WO (higher education) or HBO (higher vocational education) diplomas will have advanced further than the beginners level in certain areas, and may achieve a score approximating to the final 'economist' level.

In other words, because some students have already gained a great deal of experience in a work environment or have already attained a relatively high educational level, a test set at beginners level (final VWO level) would not be able to measure some of the prior knowledge state.

The University of Limburg possesses a wealth of experience in constructing tests, especially tests associated with attainment targets. Studies into the importance of these attainment target tests or progress tests for the first study year at the University of Limburg have been carried out by Imbos (1982, 1989) and Wijnen (1984). Wijnen constructed a first year attainment target test based on the p-values of the test questions. If an item had a relatively high p-value, it was considered to be suitable for a first year test. The lower limit was set at .40 (PES bulletin no. 68).

A representative random test was constructed of items selected from the item bank of the Economics Faculty of the University of Limburg, which is continually updated with validity and reliability data. The items in the data bank are classified in 9 subject areas, viz.:

a - reporting
b - financing
c - organization
d - marketing
e - macroeconomics
f - microeconomics
g - public finances
h - international economic affairs
i - behavioral and social sciences

The total random test comprised 154 items divided between the various subject areas; the distribution was as follows:

a:	18 items
b:	18 items
c:	18 items
d:	18 items
e:	25 items
f:	25 items
g:	11 items
h:	11 items
i:	10 items

These categories correspond to an equal number of basic disciplines in economics. Each category contains 4 types of questions with different codes. The codes have the following meanings:

. A question number accompanied by an asterisk (*) indicates general economics or business economics at first year level.

. An underlined question number accompanied by an asterisk indicates a question with a quantitative economic accent at first year final level.

. An underlined question number without any accompanying symbol indicates a question with a quantitative economic accent at second year final level.

. A question number without any further accompaniment indicates a general economics or business economics question at second year level.

The test consists of simple and multiple questions. Each question or each part of a multiple question is preceded by a question number. Only that which comes after the question number must be assessed. The correctness of any text that precedes the question number (the so-called stem of the question) need not be assessed; it can be assumed that the information given in this text is correct within the context of the particular question. The question numbers for which each stem is relevant are indicated in the stem, in parentheses. Furthermore, the information in the stems from one and the same case is cumulatively valid.

An example from the domain-specific prior knowledge state test will make this clear:

The company currently uses four capital units.

1. y/?/false The total short-term variable-costs comparison is:
$$F_v = 0.0025\ Q^2 + 4$$
where F_v represents the total variable costs.

2. y/?false Based on the reported total variable-costs comparison (F_v = ...), the company will supply 400 units of the end product at an end product price of 2.

The result of the test is calculated by subtracting the number of incorrect answers from the number of correct answers, whereby question marks count as zero. In other words:
correct = +1, incorrect = -1, ? = 0.

C. Validity, usefulness and reliability of knowledge state tests

In order to obtain as practical, valid and thus reliable a test as possible, attention must first be paid to:
- validity and usefulness;
- reliability and feasibility aspects with regard to ecologically valid study and further implementation of the tests.

Validity and usefulness

For the content of a test to be valid, the test must comprise a number of questions related to clearly delineated aims and to a specific subject or domain. The more representative a random sample of the items is of the total number of possible items, the more valid is the test.
Because the item bank of the University of Limburg is based on a thorough analysis of the programme of economics education, and has been built up by a large number of experts, a representative random sample of items from this data bank can certainly be claimed to be valid with regard to content. This means that the construction of the subject-specific PKS is balanced, and is constructed of items related to the aims contained in the study material.

As regards usefulness, it has already been described how care is given to the formulation and lay-out of the tests. When tests are being constructed, attention is paid to criteria such as efficiency, fairness and the time available.

Reliability and feasibility aspects of ecologically valid studies and possible further implementation of knowledge state tests

During the planning stage, great attention must be paid to aspects such as objectivity and test length, with a view to attaining as high a degree of reliability as possible.

In addition, the construction of the test will not be complete unless care is devoted to the distribution of true and false items. Here too, chance successes must be avoided. Earlier studies (Ebel, 1972; Grosse and Wright, 1985; Verwijnen, Imbos, Van Hessen and Wijnen, 1987) have demonstrated that questions with the answer key 'true' are answered correctly more often than questions with the answer key 'false'. This is of importance for the comparison of parallel tests. A significant difference in the distribution of true and false items will result in an built-in bias to one direction or the other (Imbos, 1989). There is no significant difference in the distribution of both sorts of items in our test.

In addition, the consistency of the test itself is only shown up by the analysis. We shall deal with this in more detail when we describe the analysis of the test.

With a view to the feasibility of using PKS tests in practical education situations, test construction manuals (De Groote and Van Naerssen, 1969; Ebel, 1972) draw attention to the sort of items and the number of items in a test.

If test analysis is to be quick and the method of answering the questions is to be simple, multiple-choice or true/false questions must be used.
In addition, it is recommended that the extent of the test be kept to a minimum. Only the extensive domain-specific knowledge state test would present problems in this connection, if it were to be generally introduced. Although there are sufficient good arguments to justify the length of this test, practical considerations may require it to be shortened. There are, however, sufficient possibilities for doing so.
If the representativeness of the selected items is not to be distorted, a well thought out method must be found. The random selection method, the method based on the actual answer percentages of students (Wijnen, 1984) or another method of selection based on expert opinions (Imbos, 1989) are some of the possibilities. From the point of view of study methodology, the method involving a choice based on content is conspicuous on account of the time consuming and laborious labour it involves. The p-value or empiric method is relatively easy, because a large student population is much easier to find than a large group of experts.
In this context, Imbos (1989) has compared the degree of difficulty of items selected on the basis of content and items selected by means of the empirical method. The criterion applied was that the number of selected items with high p-values must be equal to the number of questions selected by the experts.

In this comparative study, the average p-value that was taken as the lower limit in the first year was .17.
According to Imbos (1989, personal communication), there is no significant difference between the two methods. The empirical method can therefore be used; furthermore it is to be recommended if the initial selection method was based on content (in our case the representative random sample from the content-valid item bank). In any case, Imbos's study showed that the degree of difficulty of items selected solely on the basis of the empirical method was intermediate between item groups selected by means of both methods and items that would be selected on the basis of content.

3.2. The domain-specific prior knowledge state of first year economics students

In October 1988, 994 economics students at the University of Limburg took the domain-specific test we had constructed. The results of the 536 first year students were analyzed in more detail.
If we plot the global percentage scores of all year groups (corrected for incorrect scores), it can be seen that the prior knowledge state is represented by a rising line (Fig.1). The more difficult specialized items tend to depress the results of the third and fourth year students. If the specialized items are left out of consideration, the results of the starred and non-starred items can be plotted as a rising straight line (Fig.2).

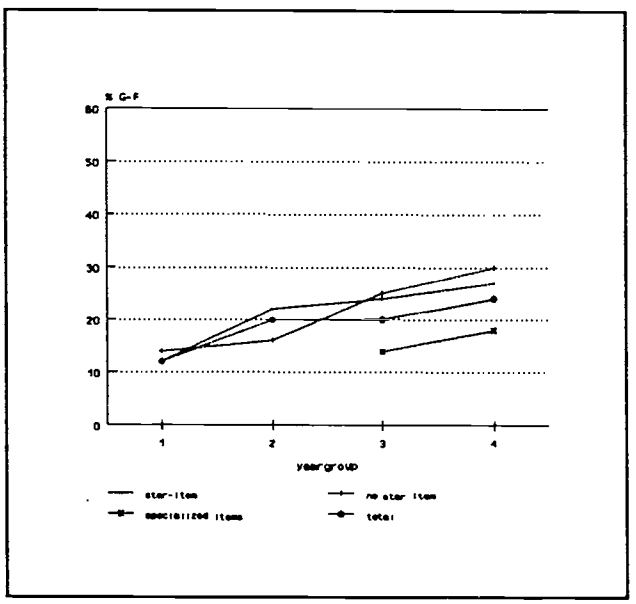


Fig.1: Domain-specific prior knowledge state test for all year groups (with specialized items)

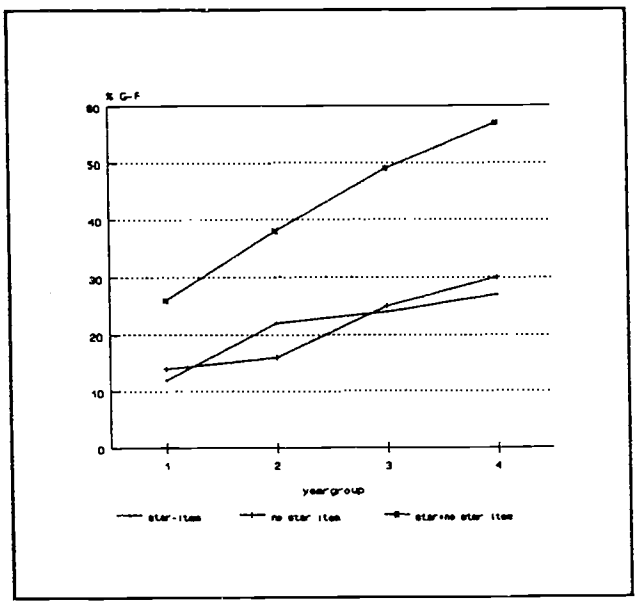


Fig.2: Domain-specific prior knowledge state test for all year groups (excluding specialized items)

The item analysis of the domain-specific prior knowledge state test (see Appendix 1) gives information about the test and about the level of the students with regard to the whole domain and to sub-domains.
The discrimination index indicates that all items differentiate to a lesser or moderate degree (mostly between .10 and .40) The $R_{i}t$ values (R-tot) are fairly low, which may indicate that the group was relatively homogeneous. The reliability of the test as a whole is .71, which is acceptable. Consideration must be given to the question of whether to scrap items that exhibit negative correlation with the total, certainly if they are found to give the same results with mature students.

The student population of the Open University can be expected to be more heterogeneous, so that higher R-tot values should be obtained. In addition, consideration must be given to the question of whether items that exhibit negative correlation should be retained.

The domain-specific knowledge state of the 536 first year students is represented in Fig.3. The students scored higher on non-starred items than on starred items, except in the sub-domains microeconomics, macroeconomics, and behavioral and social sciences.

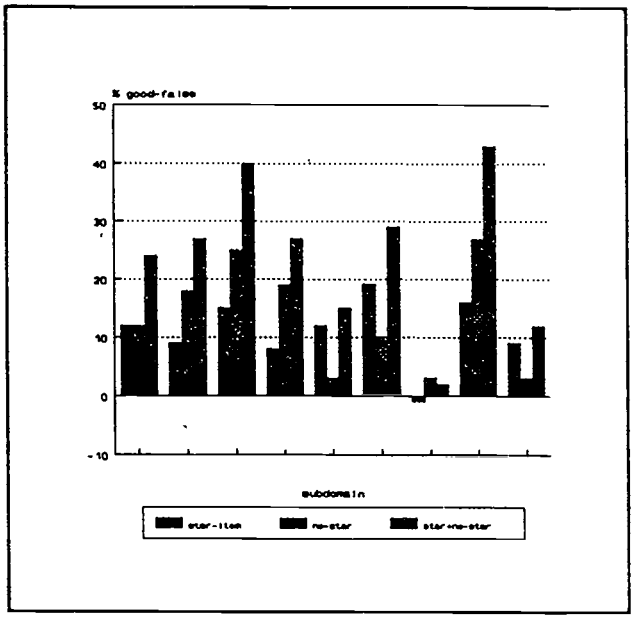


Fig.3: The domain-specific knowledge state of first year students, according to starred and non-starred items

The domain-specific prior knowledge state profile of the average student is represented in Fig.4. This can be used to determine individual deviations.

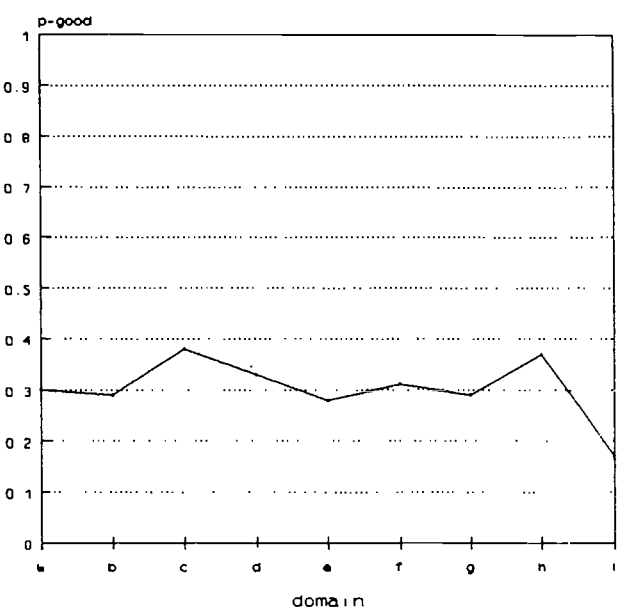


Fig.4: Domain-specific prior knowledge state profile of the average student

Furthermore, the average prior knowledge state profiles for each sub-domain give an indication of the students' behaviour (Fig.5). The knowledge state profiles for the other sub-domains are given in the Appendix. These profiles enable comparisons to be made between groups of students. Later studies can determine whether the profiles of students with good academic results differ from those of other (poorer) students.
In addition, comparison with these profiles can give an indication of students' behaviour in the sub-domains.

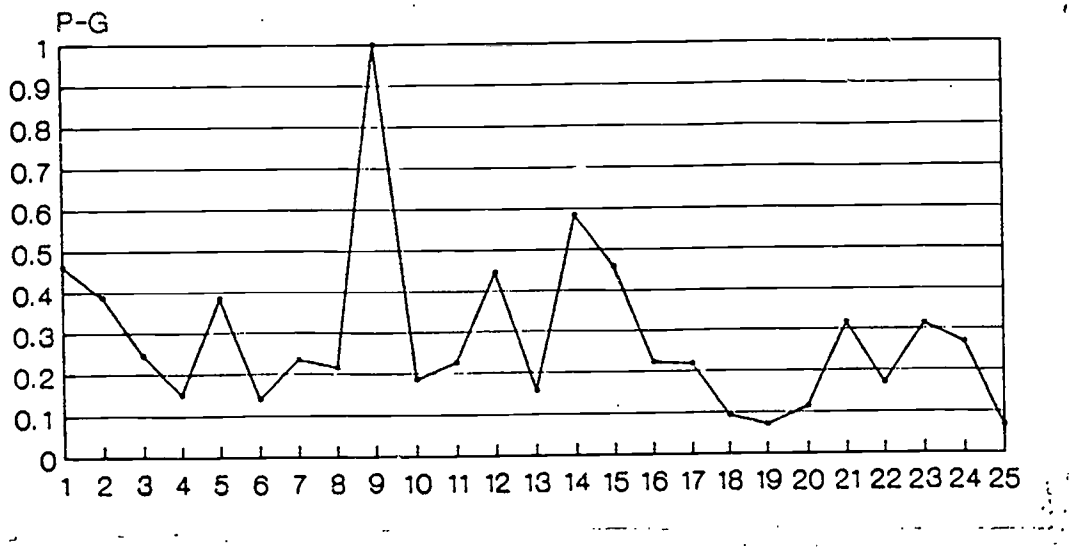


Fig.5: Average prior knowledge state profile for the sub-domain 'macroeconomics'

3.3 Further use of domain-specific prior knowledge state tests and knowledge-state profiles

A future study will be carried out to determine the domain-specific knowledge state and the knowledge state profile of Open University students and to compare them with the states and profiles already obtained. The question of whether the knowledge state and the profiles can yield more information on education and its results with regard to students in an open, modular education system. Finally, attention will be directed towards the consistency of the tests and the comparability of the items. The validity of the tests has been considered primarily in this report.

References.

Astin, A.W., Panos, R.J., (1969). *The Educational and Vocational Development of College Students.* Washington D.C.: American Council on Education.

Astin, A.W., 1982). Why not Try Some New Ways of Measuring Quality. *Educational Record,* vol. 63. pp. 10-15.

Banta, T.W. et al., (1987). Estimated Student Score Gains on the ACT COMP Exam: Valid Tool for Institutional Assessment? *Review of Higher Education,* vol. 27 (1987), pp. 195-217.

Block, A. de, (1970). Voorstel van taxonomie. *Onderwijs en media,* (2), 58-61.

Block, A. de, (1975). *Taxonomie van leerdoelen.* Amsterdam, Standaard.

Bloom, B.S., Madans, G.F., Hastings, J.T., (1981). *Evaluation to improve learning.* New York: Mc Graw-Hill.

Camstra, B., (1977). Orientatie op evaluatie. *Onderzoek van onderwijs,* 6, 6-11.

Camstra, B., (1981). *Bouwstenen voor onderwijs.* Utrecht/Antwerpen, Het Spectrum.

Carmines, E.G., Zeller, R.A., (1979). *Reliability and validity assessment.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-017. Beverly Hills and London: Sage Publications.

Carver, R.C., (1974). Two Dimensions of Tests: Psychometric and Edumetric. *American Psychologist,* vol. 29 (1974), pp. 512-18.

Corte, E. de, Geerligs, C.T., Lagerweij, N.A.J., Peters, J.J., Vandenberghe, R., (1981). *Beknopte didaxologie.* Groningen: Wolters-Noordhoff.

Corte, E. de, et al., (1976). *Beknopte didaxologie.* Groningen, Wolters-Noordhoff.

Dochy, F.J.R.C., Luijk, S.J. van, (1987). *Handboek voor vaardigheidsonderwijs.* Lisse, Swets en Zeitlinger.

Dochy, F.J., (1988). *Theories and research into the effect of the Prior Knowledge State on Learning.* Educational Technology Innovation Centre, OTIC Research Report 1. Heerlen, Open University.

Dousma, T., Horsten, A., (1980). *Tentamineren.* Utrecht/Antwerpen, Het Spectrum.

Dyck, W.E., (1976). *Geschiktheid en selectie in het universitair onderwijs.* Proefschrift. Universitaire instellling Antwerpen.

Ebel, R.L., (1969). *Encyclopedia of educational research.* London, Collier-Macmillan.

Ebel, R.L., (1972). *Essentials of educational measurement.* Englewood Cliffs N.J.:Prentice Hall.

Fincher, C., (1985). What is Value-Added Education? *Research in Higher Education,* vol. 22, 395-398.

Gagné, R.M., (1977). *The conditions of learning.* New-York, Rinehart and Winston.

Goldstein, H., (1983). Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement,* vol. 20 (1983), pp. 396-78.

Greenwald, D., (eds.), (1983). *The McGraw Hill Dictionary of Modern Economics,* (3rd ed.) New York: McGraw-Hill.

Groot, A.D. de, Naerssen, R.F. van, (1969). *Studietoetsen, construeren, afnemen en analyseren.* Den Haag: Mouton.

Grosse, M.E. & Wright, B.D., (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement,* 45, 1-14.

Harris, C.W., (ed.) (1963). *Problems in Measuring Change.* Madison: University of Wisconsin Press.

Imbos, Tj., (1982). *De betekenis van voortgangstoetsen en bloktoetsen in het kader van een selektieve propedeuse.* PES-bulletin nr.2 Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Imbos, Tj., (1989). *Het gebruik van einddoeltoetsen bij aanvang van de studie.* Proefschrift ter verkrijging van de graad van doctor aan de Rijksuniversiteit Limburg te Maastricht.

Kessler, R.C., Greenberg, D.F., (1981). *Linear Panel Analysis: Models of Quantitative Change.* New York: Academic Press.

Koper, E.J.R., (1989). *Inscript een scripttaal voor het systematisch ontwerp van interaktieve leersystemen.* Onderwijstechnologisch Innovatiecentrum-OTIC research Rapport 4. Heerlen: Open Universiteit.

Koper, E.J.R., (1989). *Leertaken onder de loep: een conceptueel model voor onderzoek naar leertaken in relatie tot knowledge acquisition support systems (KASS).* Onderwijstechnologisch innovatiecentrum-OTIC research Rapport 10. Heerlen: Open Universiteit.

Mc Clain C., (1986). Northeast Missouri State University's Value-Added Assessment Program. A Model for Educational Accountability. *Journal of Institutional Management in Higher Education,* Vol. 10, No. 3, pp 252-271

Meerling, (1981). *Methoden en technieken van psychologisch onderzoek.* deel 1: Model, observatie en beslissing. Meppel, Boom.

Mellenbergh, G.J., (1986). Twintig jaar onderwijsmeetkunde. In: W.J. van der Linden (Ed.), *Moderne methoden van toetsconstructie en gebruik.*3 Lisse, Swets en Zeitlinger.

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

Northeast Missouri State University. *In Pursuit of Degrees With Integrity: A Value-Added Approach to Undergraduate Assessment.* Washington, D.C.: American Association of State Colleges and Universities, 1984.

Nucci, L., Pascarella, E.T., (1987). The Influence of College on Moral Development. In: J. Smart (ed.), *Higher Education: Handbook of Theory and Research,* Vol. III, New York: Agathon Press, 1987, pp. 271-326.

Nuttall, D.L., (1986). Problems in the Measurement of Change. In: D.L. Nuttall. (ed.), *Assessing Educational Achievement.* Philadelphia: Falmer Press, pp. 153-167.

Os, W. van, (1987). *Evaluatie in het hoger onderwijs.* Groningen, Wolters-Noordhoff.

Pascarella, E., (1985). College Environmental Influences on Learning and Cognitive Development: A Critical Review and Synthesis. In: J. Smart (ed.), *Higher Education: Handbook of Theory and Research.* New York: Agathon Press, pp. 1-61.

Pascarella, E.T., (1986). *Are Value-Added Analyses Valuable?* Paper Presented at the 1986 ETS Invitational Conference, New York.

Taylor, T., (1985). A Value Added Student Assessment Model: Northeast Missouri State University. *Assessment and Evaluation in Higher Education,* Vol. 10, No. 3, pp 190-202.

Traub, R.E., Wolfe, R.G., (1981). Latent Trait Theories and the Assessment of Educational Achievement. *Review of Research in Education,* vol. 9 (1981). pp. 377-435.

Verwijnen, M., Imbos, Tj., Van Hessen, P., Wijnen, W., (1987).
Jaarverslag over het academisch jaar 1984/85 van de Voortgangstoets-beoordelingscommissie. PES-bulletin nr.117 Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Wijnen, W.H.F.W., (1971). *Onder of boven de maat: een methode voor het bepalen van de grens onvoldoende/voldoende bij studietoetsen.* Amsterdam, Swets en Zeitlinger.

Wijnen, W.H.F.W., (1984). *Frequentieverdeling van de goede antwoorden bij eerstejaars studenten. PES-bulletin nr.68 Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.*

Wijnen, W.H.F.W., Vleuten, C.P.M. van der, (1985). *Toetsing: hordenloop of voortgangskontrole?* Universiteit en Hogeschool, 31, 270-279.

Wood, R. The Agenda for Educational Measurement. In: *Nutall,* pp. 185-204.

Previous english reports published in this series.

The 'Prior Knowledge State' of students
and its facilitating effect on learning
OTIC research report 1.2
F.J.R.C. Dochy, 1988

Variables influencing the indexation of the
'Prior Knowledge State' concept and a
conceptual model for research
OTIC research report 2.2
F.J.R.C. Dochy, 1988

Students' views on Prior Knowledge
OTIC research report 3.2
F.J.R.C. Dochy, W.H.L. Steenbakkers, 1988

Modularisation and student learning in modular
instruction in relation with prior knowledge
OTIC research report 8
F.J.R.C. Dochy, L.J.J.M. Wagemans, H.C. de Wolf, 1989

The didactics of open education: Background, analysis
and approaches
OTIC research report 9
W.J.G. van den Boom, K.H.L.A. Schlusmans, 1989

Practical objectives at the Open University of the
Netherlands
OTIC research report 13.2
P.A. Kirschner, M.A.M. Meester, E. Middelbeek, 1989

Schema theories as a base for the structural
representation of the knowledge state
OTIC research report 18
F.J.R.C. Dochy, M.R.J. Bouwens, 1990

Practicals and the acquisition of academic skills
OTIC research report 19
P.A. Kirschner, 1990

Learning objectives for practicals in institutes
of higher education in the Netherlands:
A descriptive study
OTIC research report 21.2
P.A. Kirschner, M.A.M. Meester, E. Middelbeek,
H. Hermans, 1990

Studies on the multi-functional nature of courses
in economics and the role of domain specific
expertise. Ex post facto research 1
OTIC research report 22
F.J.R.C. Dochy, M.R.J. Bouwens, 1990

An object-oriented hypertext system for learning
OTIC research report 23
B. van Ginderen, 1990

The role of subject-oriented expertise. A study of
the impact of personal and contextual variables on
success in an economics course as indicators
of expertise. Ex post facto research 2.
OTIC research report 25
F.J.R.C. Dochy, M.R.J. Bouwens, L.J.J.M. Wagemans,
D.W. Niestadt, 1991

Analysis of the quality and impact of expertise
in economics
OTIC research report 26
F.J.R.C. Dochy, M.M.A. Valcke, L.J.J.M. Wagemans, 1991