



Pergamon

Studies in Educational Evaluation, Vol. 23, No. 4, pp. 373–398, 1997

© 1997 Elsevier Science Ltd

Printed in Great Britain. All rights reserved

0191-491X/97 \$17.00 + 0.00

S0191-491X(7)00023-0

AN ALTERNATIVE FOR ASSESSING PROBLEM-SOLVING SKILLS: THE OVERALL TEST

Mien S. R. Segers

*School of Economics & Business Administration, Dept. of Educational Development and Research,
University of Maastricht, Maastricht, The Netherlands*

Introduction

One task that credit administrators, controllers, business managers and economists in various professional contexts have in common is that they are expected to solve complex problems regularly and effectively. For Economics degree programs, an important question is: How do graduates deal with the problems they face when starting their professional career? Should we hope and pray the organization that hired them doesn't come tumbling down? Or do we have sufficient evidence that the graduates will be capable of dealing with the informational load that accompanies the problem and that they will use it in a coherent and integrated way to reach a solution?

Three distinct elements provide information about the expert status of the graduate: the content of the economics courses studied (syllabus content), the teaching methods adopted and the methods and results of the assessment used to determine the success of students in solving economics problems.

During the past decades, Economics degree programs have been subject to change in their content and instructional methods. This process of change has seldom been matched with changes in the assessment methods used to determine students' outcomes (Mallier, Morwood, & Old, 1990). This article aims to contribute to the development of appropriate assessment methods in economics education. The rationale of the assessment system implemented is informed by the findings of cognitive research on the constituent cognitive features that underlie expert problem solving and on how experts acquire their expertise. It is based on the cognitive learning theory postulating that all learning involves

thinking. The assessment approach that suits this teaching and learning theory emphasizes the use of a set of measurement tools, integrating conceptual understanding and performance skills to solve authentic problems.

The present contribution describes the results of a number of studies that have tried to find empirical evidence for quality as regards the validity of the instruments adopted in the Maastricht School of Economics and Business Administration. These studies attempt to answer questions such as: Do the measurement tools provide a profile of students' conceptual understanding and problem solving skills? Are they fair, i.e., to what extent can students be expected to meet the goals measured by the test? Knowledge about the extent of overlap between what is tested and what is taught is critical to the interpretation of the test results. This report also presents empirical evidence for some of the basic assumptions of the assessment system adopted.

The Maastricht Assessment System

Rationale

The Maastricht economics curriculum is intended to guide students to become academic professionals: graduates who can identify the problems of different disciplines within the field of economics and who are capable of analyzing and contributing to the solutions of these problems. *Problem* is the key word in this goal definition. The Maastricht School of Economics and Business Administration adopted a problem-based educational approach in designing its curriculum. This approach is significantly influenced by the findings of cognitive psychological research, especially results from expert vs. novice studies. Two general characteristics of expert performance can be identified (Feltovich, Spiro, & Coulson, 1993; Glaser, 1990; Yekovich, 1993):

- Experts' knowledge is coherent. Experts possess a well-structured network of concepts and principles about the domain that accurately represents key phenomena and their interrelationships. Beginners' knowledge is not only patchy, consisting of isolated definitions, it also lacks the principles underlying surface features of a problem presented. In contrast, experts' knowledge is structured and cognizant of underlying principles and patterns.
- Novices often know facts, concepts, principles without knowing the conditions under which this knowledge applies and how it can be used most effectively. "Experts and novices may be equally competent at recalling specific items of information, but the more experienced relate these items to the goals of problem solution and conditions for action" (Glaser, 1990, p. 477). Dochy and Alexander (1995) identify this type of knowledge as *conditional knowledge*; experts are able to use the relevant elements of knowledge in a flexible way in order to describe, analyze and solve novel problems.

This expert profile requires the development of a learning scheme aiming at analyzing, solving and evaluating problems on the basis of a deep understanding of the

subject domain studied. This can be illustrated by an example taken from industrial economics (Lawson, 1992). A student sets out to study the economics of running an airline. First, he will need access to what is known about airline operations and how economists analyze different types of markets, using a range of models of the firm, from wholly competitive firms through to monopolies. Second, the student will have to appreciate the purposes and limitations of the theories for the firm. He will have to be able to assemble the facts concerning the airline's operation. Then he will have to link theories with these assembled facts. For example, describing British Airways as a regulated carrier with considerable yet limited market power within the domestic UK market, would require that the student recognize the appositeness of the regulated industry model to the facts of BA's domestic operations.

This example suggests some principles for the type of instruction that should guide the student who is preparing to deal with such problems. They can be summarized as follows:

- The curriculum should focus on clusters of related concepts. The development of conceptual networks is enhanced when students are actively engaged in the learning process. Students should be encouraged to manipulate and use the knowledge they are acquiring by confronting them with authentic problems. Acquiring knowledge is not the ultimate goal of instruction. A major goal of instruction should be promoting understanding of important conceptual knowledge in such a way that it can be used in analyzing and working with realistic problems (Feltovich, et al., 1993).
- Feltovich et al. (1993) stress that knowledge that will be used in many ways has to be learned, represented, and tried out (in application) in many ways. Therefore, knowledge (including concepts, models, theories) should be interrelated in diverse ways and cases should be addressed in relation to other cases. The use of a variety of cases involving similar concepts, and of similar cases embodying different concepts, helps students to work with novel problems. Cases and knowledge should be "revisited" from different relevant points of view and for the purpose of answering different kinds of questions.

The changes which take place when proficiency develops not only define the criteria for instruction by which competence can be developed but also the criteria by which competence can be assessed. Indeed, instruction and assessment must be linked for at least two reasons. First, student outcomes provide information that can be used in improving educational practice only when the instruments that measure the outcomes match the instructional practice (English, 1992). Second, tests are diagnostic aids only when they identify the extent to which the goals are attained. This means that tests must be sensitive to how well students are able to use knowledge in an interrelated way when analyzing and solving authentic problems.

The instructional principles described above lead to the following assessment principles:

- Assessment instruments should measure the extent to which students possess knowledge that is organized in a way that facilitates fast and correct recognition of patterns. A significant dimension for assessment of competence is the presence of interrelated concepts. In addition, the ability to recognize principles and patterns underlying the problem or task presented is an indication of developing competence that should be assessed (Glaser, 1990).
- The assessment system should substantially focus on measuring the extent to which students are capable of flexibly applying their knowledge to analyze and solve novel problems. These problems should be of a real-world type. Research in the field of mathematics suggests that such problems offer opportunities to develop understanding in context, to develop reasoning in the subject domain and to develop the making of interdisciplinary connections (Blum & Niss, 1991; De Lange, 1992; Lesh & Lamon, 1992).

The Key Features

The Maastricht School of Economics and Business Administration implemented problem-based learning in its curriculum: Students are confronted with authentic problems, i.e., problems they might encounter in real life situations. Because authentic problems are often not solvable within mono-disciplinary constraints, the curriculum is organized on a multidisciplinary basis. This implies that problems are discussed from different points of view (disciplines) such as those of marketing, organization, and micro-economics. The problems are the context within which students study the basic concepts and models within the fields of economics and business administration. Thus students acquire and apply knowledge simultaneously.

The assessment system developed follows the organizing principle of the curriculum. The acquisition as well as the application of knowledge is assessed, and for this purpose two instruments are implemented: the Knowledge Test and the OverAll Test.

The Knowledge Test

The Knowledge Test measures primarily the knowledge of facts, the meaning of symbols and the concepts and principles of the four particular fields of study: marketing and organization, micro-economics, macro-economics, and accounting and finance. This type of knowledge is often defined as declarative knowledge (Anderson, 1983; Dochy & Alexander, 1995). The test items require students to reproduce and/or demonstrate understanding of their knowledge about the main subjects studied. It is not sufficient for students to remember or even understand isolated definitions of domain-related concepts. They need to understand the frame of reference which organizes them.

The Knowledge Test covers the domain studied within one *instructional period*.¹ It consists of 100 to 150 multiple-choice true/false format items. To assure relatively even coverage of the domain, an analytic grid is used for the construction of the test.

Figure 1 offers some examples of Knowledge Test items.

Question 1

A very important question is which management principles a manager should use to achieve organizational excellence. During this century several different viewpoints have emerged.

true/?/false According to the contingency viewpoint, managers should analyse and understand situational differences and choose the best solution suited to the firm and the individual in each situation.
(True)

Question 2

The ice cream-company "Magnus" was only a producer of icecream. Today, "Magnus" is producer and seller of icecream.

true/?/false When the ice cream-company "Magnus" combines the producing and the selling of ice cream under the same management, vertical integration takes place.
(True)

Question 3

After a recession, it was observed that employment did not rise at the same time as general economic activity.

true/?/false This can be explained by referring to slack organizational resources.
(True)

Question 4

Suppose that there is inflation, and that the Central Bank changes the growth rate of the money supply so as to equal the long term annual growth rate of production. Suppose also that people believe this money growth rate will continue to equal the growth rate of production. In the following several immediate effects are mentioned

true/?/false An immediate effect would be that the nominal interest rate would fall. (True)

true/?/false An immediate effect would be that actual inflation would temporarily be negative.
(True)

Figure 1: Examples of Knowledge Test Items

These four examples assess conceptual understanding. The first and the second questions require students to be able to recognize the definition of the contingency viewpoint and the definition of vertical integration. The second question is embedded in a simplified authentic situation. It asks for more than merely factual recall. Students have to not only reproduce the definition of the concept of vertical integration, but also apply it to the case of the ice cream company. Since only the relevant variables are mentioned, students do not need to retrieve the relevant information from the case in order to be able

to identify the strategy used as vertical integration. In the third question, in order to be able to give the right answer, students need to build the following frame of reasoning: if economic activity grows, organizational activity will grow. In that case, organizations will first use their slack human resources. For example, by transferring people internally, they are able to meet their increased need for personnel instead of hiring externally. As a result, employment will not rise at the same rate as the general economic activity. Being able to define the concept of *organizational slack resources* is not sufficient. The conditions for the application of these resources and the consequences in macro-economic terms need to be understood.

The fourth question starts from a macro-economics case which, like the second question, presents the critical elements for solving the problem. To answer the questions, students need to understand the various relevant concepts (nominal interest rate, inflation, rate of growth of the money supply, long-run annual growth rate of production). Additionally, they are required to master the interconnections between these concepts.

As is clear from the examples, we introduced the *question mark* option. This option allows the students to "pass". Students who circle the question mark option indicate they have not mastered the subject: This option allows them not to give an answer and therefore to avoid guessing. They are not punished for not knowing: choosing the question mark option gives a score of 0 points. On the other hand, they lose one point (-1) when indicating the wrong answer. Circling the right answer means +1 score. The introduction of this scoring system makes guessing only attractive for students who are reasonably sure of the answer, having mastered an important part of the test items. Therefore, in most cases² choice of the wrong answer reveals that students have misunderstood the objective measured. For the example in Figure 2, the test results revealed that some students had constructed their own interpretation of the meaning of a (un)differentiated marketing strategy. Although the concept was studied during tutorials, the misconception persisted that differentiating has to do with the target market instead of the product. Since quite a large group of students took the risk of indicating the *false* option, they seemed to be quite sure of their answer.

true/?/false Both an undifferentiated and adifferentiated marketing strategy are directed at approximately the whole market.

(true)

Figure 2: An Example of a Knowledge Test Item

The OverAll Test

Figure 3 presents an example of an OverAll Test item to illustrate how it differs from a Knowledge Test item as explained in the previous section.

Case Mexx

The case study presents the history and recent developments in the fashion company Mexx. Main trends within the European clothing industry are described. The Mexx Fashion company is illustrated by its organizational structure, its product profile and market place, its business system, its corporate culture and some current facts and figures.

Question 1

true/false Mexx's corporate culture and philosophy is consistent with the systems viewpoint on management.

(False, it is consistent with behavioural viewpoint)

Question 2

Benneton's and Mexx's corporate strategies are quite different. More specifically, there are two main differences.

- Identify these two main differences in corporate strategies. Illustrate your answer with examples mentioned in the case.
- What are the advantages of Benneton's corporate strategy compared to Mexx's approach?

Figure 3: Examples of OverAll Test Items

The first question is identical to the first sample question of the Knowledge Test: they both refer to the different viewpoints on management. The Knowledge Test item requires from the students to recognize the definition of one of the viewpoints. Memorization of the definition is not sufficient to answer the OverAll Test item. Students have to interpret the case and select the relevant information for this test item. On the basis of a comparison of this information with conceptual knowledge of the different viewpoints on management, they have to deduce the answer. The second OverAll Test question resembles the second Knowledge Test item: they both refer to the concept of vertical integration. However, the OverAll Test item requires students to take more mental steps to reach a solution than does the Knowledge Test item. For the first part of the question (a), these can be schematized as follows:

- Define the concept of corporate strategies
- Select the relevant information for the Mexx company as described in the case study
- Compare it with the definition of the different possible strategies
- Select the relevant information for Benneton as described in the case study
- Compare it with the definitions of the different possible strategies
- Compare the relevant information from both cases with the definition of the strategies
- Define each company's strategy
- Compare both strategies by going back to the definition of the strategies and the relevant information in the case study

For the second part (b), students have to evaluate. Therefore, they have to take some additional mental steps:

- Understand the conditions for efficiency and effectiveness for the different strategies
- Select the relevant information on the conditions for both companies
- Interpret the factual conditions by comparison with those studied in the textbooks

This example illustrates that the OverAll Test measures whether students are able to retrieve the relevant concept (model, principles) for the problem. Furthermore, it measures if they can use these instruments to solve the problem. It measures if the knowledge is usable (Glaser, 1990) or whether students know "when and where" (conditional knowledge). In short, the OverAll Test measures to what extent students are able to analyze problems and contribute to their solution by applying the relevant tools.

The OverAll Test is organized within the first year curriculum as follows. After two instructional periods (blocks), the students get two weeks off for self-study. During these weeks, they study on the basis of the study manual they receive at the beginning of this period. This manual presents information about the main goals of the OverAll Test, the parts of the curriculum which are relevant for the study of the material presented in the manual, an example of an elaborated case with test items, some practical (organizational) information, and a set of publications, such as for instance a description of a case relating to innovations in or problems of a national or international firm as published in a newspaper or a journal. Other included articles present the theoretical considerations of a scientist, report research, or comment on a theory or model. During the self-study period the students are expected to apply the knowledge they have acquired over the preceding weeks, with a view to explaining the new, complex problem situations which are presented in the articles. They are asked, while reading these texts, to try to explain spontaneously to themselves (i.e. without being explicitly prompted by a tutor) the ideas/theories described in them by relating them to previously acquired knowledge. This behaviour is often called "self-explanation" (Chi, Feltovich, & Glaser, 1981). In short, the self-study period can be described as an opportunity for students to practice the analysis and synthesis of economics problems as they have learned to do in the tutorial groups. For this purpose the study manual offers them a set of new problems appearing in the set of articles. Figure 4 provides an example.

Article: Schoemaker, P.J.H. (1995). Scenario Planning: a Tool for Strategic Thinking. *Sloan Management Review*, pp. 25-39.

Study Guideline: "...The first part that deserves additional attention is the description of the two applications of scenario planning. On these pages Schoemaker relates his method of scenario planning' to the various statistical techniques you have encountered in the Quantitative Methods blocks. Don't restrict yourself to the role of a passive consumer of his treatment of statistics, but take a more active position by comparing Schoemakers' use and interpretation of statistical concepts with that to be found in our textbook by Wonnacott & Wonnacott (W&W). To give a simple example: what is Schoemaker's definition of a correlation matrix'? And how does this view relate to the interpretation of the concepts covariance (matrix) and correlation (matrix) as found in W&W ? And next, what exactly is the relationship between the correlation matrix (such as the one in Table 3) and the scenario profiles (given in Figure 1)? To phrase the last question differently: if we gave you an arbitrary correlation matrix, could you derive the corresponding scenario profiles?" (OverAll Test I, Information and Study Guidelines, 1995-1996)

Figure 4: An Example of an OverAll Test Study Guideline

After the two weeks of self-study, the OverAll Test is administered. The OverAll Test questions refer to the articles: they assess whether the students are able to interpret and analyze the problems as presented in the articles by applying the concepts, models and tools they have acquired during the tutorials.

Figure 5 displays two questions referring to an article by P.J.H. Schoemaker.

Question 1

In his introduction, Schoemaker compares the method of scenario planning with other approaches such as contingency planning, sensitivity analysis and computer simulations. Hellriegel & Slogum (1996 textbook) offer a similar comparison of three methods: scenarios, the Delphi technique and simulation. They stress that there is an overlap between these approaches, and indeed, it is not difficult to imagine how to use techniques like Delphi and simulation within Schoemaker's framework of scenario planning.

True/?/false The Delphi technique fits better in phase 3 (identifying basic trends) than in phase 9 (develop quantitative models) of the scenario planning.

Question 2

The correlations in Table 3, Part B on p. 31 (Schoemaker, 1995) are nearly all positive, which makes the case rather specific.....Give a new example of scenario planning by solving the following tasks:

- a. Write down a hypothetical correlation matrix of the same size as the one in Table 3, but with the number of entries with '+', '-' and '0' more equally distributed;
- b. Derive a scenario profile (as Figure 1) from this correlation matrix, paying special attention to the existence of both positive and negative correlations. If necessary, make additional assumptions in order to find the profile. Start with one single scenario.
- c. Derive a second scenario profile, assuming this second scenario to be the "reverse" scenario of the first (literally reverse: if the first scenario is something like "recession", then the second is that of "high economic activity");
- d. Give a description in words of the consistency requirements that must be observed in assignments b and c, and
- e. Interpret the outcomes of the scenario profiles you constructed yourself. Schoemaker ends up with one scenario that performs best in all possible regards, whilst the third scenario is the worst one, again in all possible regards. Is the same true for the case you designed?

Figure 5: An Example of an OverAll Test Item

The OverAll Test is administered twice a year. Each OverAll Test assesses the application of knowledge from different disciplines which were studied during the preceding two instructional periods. The Schoemaker item illustrates the integration of knowledge in the field of statistics with the discipline of organization. Knowledge from both disciplines has to be used to tackle the problem of scenario planning.

The OverAll Test is a paper-and-pencil test. The questions are based on the articles studied at home. As is clear from the Schoemaker example, the OverAll Test combines two item formats: true-false questions with the question mark option, and essay- or open-ended questions. The true-false items are mostly intended to measure if students can apply the acquired knowledge in a new situation, if they can use an abstract concept in a specific, quite complex situation that may arise in an economist's work.

In the Schoemaker item the true/false questions ask students to use their knowledge about three approaches (studied during the tutorials) to interpret the method of scenario planning as presented in the article. It is not sufficient for students to memorize the techniques described in their textbook. They are required to know the interconnections between these approaches and how they can be effectively used within the distinct phases of scenario planning. These kinds of multiple choice questions in the OverAll Test are set in the context of authentic problems and they focus on the use of knowledge in a new problem situation. Where the test doesn't require from the students to elaborate on the relevance of the Delphi technique for scenario planning, the multiple choice format is considered to be appropriate. In contrast, the open-ended question asks for elaboration, something that cannot be accomplished with a multiple choice format. Students are asked to analyze a new problem, i.e., deriving two scenario profiles from a correlation matrix and to evaluate the outcomes of the two scenario profiles. The essay subtest and the true-false subtest have the same weight. The OverAll Test consists of seven to twelve cases or articles describing one or more related economic problem. The choice of this number of cases is based on the finding that because the sampling breadth is limited, the generalizability of scores may be poor due to content specificity (Swanson, Case, & Van der Vleuten, 1991). These findings were confirmed by the results of a pilot-study with the OverAll Test (Segers et al., 1991, 1992). Most variability was explained by the interaction effect of persons and cases (35.41% for the essay subtest and 65.48% for the true-false subtest). This means that students who perform better for one case are not necessarily the ones who perform well on other cases. It implies that one case has a low predictive value for the other cases. The findings suggest that for an OverAll Test containing 12 cases, the generalizability coefficient is 0.67.

Since it is the faculty's intention to simulate a real-world situation in the assessment system, the OverAll Test is not only based on authentic cases but also has an open-book character. This means that students are allowed to bring with them the study material they think they will need. Much like in the real world, resource materials are available. Thus, to begin with, students have to be able to select the proper resource materials and equipment related to the test. If they cannot use them in an interpretative way, they will not be able to analyze and solve the problem posed (Feller, 1994).

Main Concerns About the Assessment Practice

During the last five years the faculty has gained experience with the described assessment system. Although this generated a lot of enthusiasm, empirical evidence to interpret the effectiveness of the system implemented was lacking. A set of questions emerged. In this paper I will elaborate upon three of them.

- Are the assessment instruments fair? To what extent are the scores on the OverAll Test and on the Knowledge Test influenced by the match between instruction and test? For the OverAll Test, students' evaluations³ indicate they experienced difficulties because they had not gained enough experience in applying the acquired knowledge within a diverse set of realistic situations. However, the OverAll Test is based on the faculty objectives as operationalized within the students' study materials and the tutors' guidelines. The question arises as to whether there is a lack of match between the formal and the operational curriculum. If this is so, students might be expected to have serious difficulties answering the test (Birenbaum, 1996; Pelgrum, 1989). Within a problem-based curriculum where tutors are only the guide the students have for generating learning issues and self-study, it is especially important to obtain information on this issue (Dolmans, 1994). This leads to two interrelated questions. First, is there a match between the formal and the operational curriculum? Second, to what extent does the test measure the formal and operational curricular objectives?
- Does the OverAll Test measure the extent to which the students are able to use a conceptual network to analyze authentic problems? Or is it just another instrument for measuring factual recall?
- What is the use of a Knowledge Test as compared to a more traditional assessment instrument? Do the Knowledge Test scores provide additional and indispensable information about the students' level of expertise?

These questions were addressed in three studies. For each study, the theoretical framework, research method and results will be described.

Study 1: Are the Assessment Instruments Fair?

This study examines the curricular and instructional validity of the faculty assessment instruments.

Rationale

Is it fair to expect the students to answer the test questions? If the test is valid or, in other words, if the knowledge assessed is part of the curriculum, the answer should be yes. This means the test content matches the formal curriculum, i.e. the curricular objectives and the curriculum material. To check this match is the most common method test constructors use to establish test validity. In so doing, they assume that the objectives are actually taught. Many studies indicate that this assumption may be questioned (Calfee, 1983; De Haan, 1992; English, 1992; Leinhardt & Seewald, 1981; Pelgrum, 1990). The operational curriculum – what is actually taught in the classrooms – can significantly differ from the formal curriculum as described in textbooks and syllabi. McClung (1979) introduced the term *instructional validity* to describe the match between the operational curriculum and what is tested. The overlap between the test content and the formal

curriculum is called *curricular validity*. A mismatch between the formal and the operational curriculum may have considerable consequences. On the basis of the assessment results, the faculty makes inferences on the extent to which the faculty objectives are reached. They are one source of input for the evaluation of faculty practice. If the test does not measure what has been taught, no inferences can be made about the quality of the teaching process (English, 1992). In a summative context, when the tests are used as selection instruments, the faculty only expects the students with a certain profile to pass the tests. This profile is defined in congruence with the faculty objectives. When the assessment instrument lacks instructional validity, how can the students be described in terms of knowledge and skills? To what extent can the assessment results indicate if first year students will be able to follow the second year courses which build on knowledge presumably acquired in the first year?

Although instructional validity is a concern for all types of curricula, what about problem-based curricula? In a Problem-Based Learning-setting, more than in a conventional one, students are expected to take responsibility for their learning process. In a conventional curriculum, teaching is the central process. The tutor defines the objectives, the content of the courses, the ways to reach the objectives. In most cases, the teacher directly "delivers" the information to the students through lecturing. In the case of a local test (no national standardized test), he constructs the test on the basis of his notes for the lectures. In Problem-based Learning, the path from the objectives to the test is longer and the students have more freedom to choose their own way.

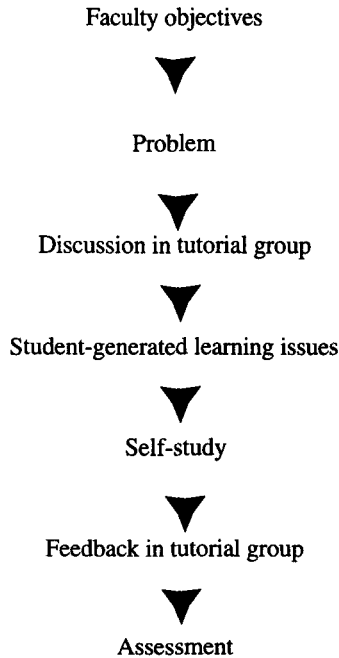


Figure 6: The Learning Process in a Problem-based Learning Setting

Faculty objectives



Lectures/Syllabus



Self study



Assessment

Figure 7: The Learning Process in a Tutor-centred Instructional Process

The faculty objectives are operationalized in a set of tasks. These tasks present the problem situation which students have to analyze and try to solve. Students work on the problem in small tutorial groups. The result is a list of learning objectives considered to be relevant for analyzing and solving the problem at hand. Since in many cases the problems are ill-structured (i.e., complex, as real problems mostly are), there may be differences in learning objectives. The learning objectives are the starting point for students' self-study. They look for the relevant information to achieve their learning objectives. In discussion with their peers and the tutor, they check the relevance of the information for the problem and build the relevant theoretical framework. At the end of the instructional period, a test is administered to measure the extent to which they have mastered the basic knowledge of that instructional period. How sure can we be the test is fair to various students who may employ different paths for analyzing and solving the single problem posed? Various studies have tried to gain insight into the relationship between the formal and the operational curriculum within a Problem Based Learning setting (Coulson & Osborne, 1984; Dolmans, 1994; Shahabudin, 1987; Tans, Schmidt, Schade-Hoogveen, & Gijsselaers, 1986). They conclude that there is a significant overlap between both curricula. Additionally, Dolmans (1994) investigated the relationship between the time spent in tutorial groups on the core concepts of the instructional period and test scores on items referring to these concepts. The correlation seems to be weak ($r=.22$, $p<0.5$, $n=94$). Probably it is the quality, more than the quantity, of the time spent on problems that affects test scores.

Research Method

Procedure

The formal curriculum was described by analyzing the textbooks, syllabi and tutorial manuals. The analysis resulted in a list containing more than 500 detailed topics for each period. This extended list was screened by domain specialists to get a workable

list. They constructed a hierarchical schema of the list of topics. The highest hierarchical levels of the networks of subjects are included in the final version; for example the concepts of *entry strategies*, *export*, *licensing* and *joint ventures*. They were all included in the draft version. In the final version only the concept of *entry strategies* was included. Thus the list of central concepts was reduced to 147 topics for the Marketing and Organization period and 136 topics for the Macro-economics period. The curricular validity was examined by comparing the formal curriculum with the test of the first instructional period. The list of concepts was compared with the list of objectives of the Knowledge Test and OverAll Test.

To examine the instructional validity, two questionnaires were developed on the basis of the lists of concepts. The questionnaires are a modified version of the Dolmans Topic Checklist (1994). The first Topic Checklist (TOC 1) consist of 147 topics and eight main themes in the disciplines of Marketing and Organization. An example of the TOC 1 is presented in Figure 3. The first column contains some examples of the 147 topics of TOC 1. The upper row presents some examples of the eight main themes. Its relevance will be explained under Study 3.

| Topics | Organization + Systems | Marketing Mix | Consumers' Behavior |
|----------------------------|---------------------------|---------------|---------------------|
| Structure follows strategy | 1 | 2 | 3 |
| Product attributes | 1 | 2 | 3 |
| Giffen good | 1 | 2 | 3 |
| Lorenz curve | 1 | 2 | 3 |

Figure 8: Example of Topics and Main Themes in the Topic Checklist 1

Students were asked to indicate whether the topic was discussed in their tutorial groups or not, by marking the topic or not. In order to gain some insight into the quality of the time spent on the topic, the second Topic Checklist (TOC 2) on Macro-Economics, consisted of two additional questions. Students had to indicate the level of comprehension they believed they had reached. For every respondent, the number of topics they had mastered on each of the three levels of comprehension was counted. These levels were defined as the *level of definition*, the *level of comprehension* and the *level of analysis*. Mastery on the level of definition indicates that the student is (only) able to reproduce the meaning of the concept as formulated in the textbooks. Comprehension of the topic implies that the student is able to define the concept in his own words, describe its relevance and its relation to other concepts. To master a topic on the level of analysis would require the student to be able to apply the concepts when being presented with a problem to be analyzed. The staff members who developed the course were asked to indicate for each topic the intended level of comprehension. Finally,

students were asked if a topic had received much, moderate or not much attention during the tutorial meetings.

Sample

The sampling procedure employed in the study was that of the quota sample. The group of first year students was, for organizational reasons, divided into four groups. Two groups had their meetings in the morning, two groups in the afternoon. Students were equally selected from these four groups. For the TOC 1, 34 student volunteers participated, for TOC 2, 45 students.

Results

As the results in Table 1 indicate, there is significant overlap between topics as planned for study by the staff and the topics indicated by the students as being subject of discussion and study during the instructional period. Table 1 indicates that on average 87% of the topics of TOC 1 and 77.4% of the topics of TOC 2 have been subject of study (RT). Other studies investigating the match between the formal and the operational curriculum in a Problem Based Learning setting (Dolmans, 1994) showed an overlap of 64.2% ($s=26.7$).

Students perceived they had mastered on average 47% of the topics of TOC 2 on the level of comprehension, i.e., that they were able to explain in their own words the meaning of the topics, their relevance and their relation to other concepts. For, on average, 31% of the topics, students stated they were able to use these topics for the analysis of problems (level of analysis). For 22% of the topics, on average, students indicated they had mastered them on the level of definition, i.e., that they were able "only" to reproduce the definition. The correspondence with the aims of the staff is considerable.

Table 1: The Degree of Overlap Between the Formal and the Operational Curriculum

| Variables | Mean | Standard Deviation | n |
|---------------|----------------------------------|--------------------|----|
| RT1* | 87% | 17.33 | 34 |
| NRT1* | 12% | 15.67 | 34 |
| RT2 | 77.4% | 12.64 | 45 |
| NRT2 | 22.6% | 25.67 | 45 |
| Definition | 22.1% (student) 20.6% (staff) | 21.24 | 45 |
| Comprehension | 47% (student) 40.4% (staff) | 22.64 | 45 |
| Analysis | 30.9% (student) 39% (staff) | 16.58 | 45 |

* RT: Recognized Topics (1=TOC 1, 2=TOC 2)

* NRT: Not Recognized Topics

Comparing topics that had either been or not been discussed (RT/NRT) with test items content, none of the topics which were indicated as not having been subject of discussion by more than 29% of the students (percentile 25) were part of the tests. This result suggests high instructional validity of the Knowledge Test as well as the OverAll Test.

Additionally for TOC 2, the more topics students marked as having "received much attention during the meetings", the higher their OverAll Test score ($r = .40^*$). On the other hand, the more topics students regarded as having "received moderate attention during the meetings", the lower the OverAll Test scores were ($r = -.32^*$). Probably, students acquired partial knowledge by informal exchanges they had about the topic. This partial knowledge might impede instead of enhance successful problem analysis. There was only a very weak correlation between topics which received not much attention and the test scores ($r = .01$).

Study 2: Criterion Validity of the OverAll Test

The second study focuses on the evaluation of the OverAll Test criterion validity: To what extent do the test scores reflect students' ability to analyze and solve economics problems?

Rationale

Despite the general enthusiasm for alternative forms of assessment, stressing complex cognitive processes and analogy with the actual conduct of problem solving, this type of assessment poses some challenging problems. As compared to traditional tests, authentic assessment instruments that measure problem solving are often thought to be better reflections of the criterion performances that are of importance in the students' future professional careers (Linn & Burton, 1994; Magone, Cai, Silver, & Wang, 1994). Until now, there are only a few examples of studies offering empirical evidence for this assumption (Burger & Burger, 1994; Magone et al., 1994). Magone et al., (1994) report on the QUASUAR project (Quantitative Understanding: Amplifying Student Achievement and Reasoning) and more precisely on the QUASAR Cognitive Assessment Instrument. It consists of a set of open-ended assessment tasks, asking students not only to select or produce answers but also to show their work or to justify or explain their solutions. Magone et al. used different sources of logical and empirical evidence for judging the validity of the assessment instrument: well-defined tasks specifications, systematic internal and external reviews of each task, and qualitative analysis of students' responses. This quantitative analysis focused on the processes underlying task performance: Does the analysis of the students' responses indicate their conceptual understanding and their ability to use basic concepts to solve a problem? According to Magone et al. (1994), the results support the validity of the instrument. They suggest that the tasks require high-level thinking and reasoning processes.

Another source of information for the validity of a test is the relation of the test scores to an external criterion. Shepard (1992) describes the empirical evidence of relations to external criteria as an integral part of today's definition of validity. Tests always involve simplifications of what we intend to measure. Therefore, it is important to determine if and to what extent test scores reflect other abilities than those intended. Writing skills, for example, might confound open-ended assessment of the analysis of economics problems. Criterion-related validity is especially important in practice for selection and placement decisions. If the test is used to select students for graduation or for entering postgraduate courses, a practically significant statistical relationship should be evident between test score and relevant criterion. Burger and Burger (1994) compared three instruments, two performance-based assessment instruments measuring writing and reading skills and a norm-referenced test series designed to measure achievement in basic skills taught throughout the nation. This study provides some "encouraging" (p. 14) evidence for the validity of the performance assessment instruments.

The purpose of the study presented in this article is to use Burger and Burger's approach to determine the criterion validity of the OverAll Test: To what extent does the OverAll Test measure the ability of students to define, analyze and solve economics problems?

Research Method

Procedure

Student performances on the OverAll Test and on a set of economics problems were compared. Four problems dealing with real-life situations, were formulated by experts in the field of macro-economics and finance. The construction and review of the problems were guided by a set of criteria for case writing (Leenders & Erskine, 1989; Vilsteren van, Heijden van der, & Arts, 1993). The described economic problems vary in length from 25 to 100 lines. Each problem starts with an introduction, presenting information about the company (context information), and the position of the student. The specific problem situation is described next. The problem description ends with a set of no more than three analysis tasks. They refer to the analysis of the problem presented as well as to the analysis of reasons (Messick, 1989).

In order to analyze the processes underlying the problem analysis, the method of think-aloud protocols is used. The participating students read the problem description aloud. Then they are asked to think aloud as they analyze the problem (Messick, 1989). In order to analyze the route students follow in their analysis, they are asked to mention if they return to a previous section of the problem description. Immediately after thinking-aloud problem analysis, students are asked to write down their response to the problem.

The analysis of student (written and oral) responses focuses on the knowledge structures that are used during problem solving. It does not only look at the points of decision between alternatives, but also attempts to map the whole process from the formulation of an hypothesis to the reaching of a solution to the problem, and considers the nature of the knowledge and the cognitive operations used to reach the solution (Patel & Arocha, 1995) The schemes for the analysis of the responses were based on a

detailed model for problem analysis by the expert-constructors. If necessary, the schemes were expanded and modified as a sample of actual responses was reviewed and coded. Central criteria for the coding were the amount of correct concepts, relationships between the concepts used for problem analysis, and the correctness of the product (solution of the problem). For the latter criterion, three categories were used: correct answers, partially correct answers and wrong answers. In the analysis of student responses to the problems, two more categories were examined: the length of the reasoning process and the degree to which students went straight to the aspects of the problem relevant to the analysis (Flaherty, 1974). Finally, comparisons were made of the results of the protocol-analysis for the three groups of students.

Sample

The results of the analysis of the four problems were obtained for fifteen first-year students. The sampling procedure used in the study is a quota sample. From the 45 participants of the first study (TOC 2), 15 were selected on the basis of their scores on the OverAll Tests. We divided the 37 participants into three groups: the group of students with the 27% highest OverAll Test scores, the group of students with the 27% lowest OverAll Test scores and the group in between. Five students were equally selected from these three groups.

Results⁴

In general, the students with a high score on the OverAll Test (high achievers) performed better on the problem tasks. They identified more relevant concepts and clusters of concepts (interrelated concepts). As presented in Table 2, the high achievers identified 63% of the relevant concepts, the low achievers 37% and the moderate achievers 38%. Additionally, the amount of correct answers to the analysis tasks formulated for each problem (decisions in the problem solving process), was significantly higher for the high achievers (6.2) than for the low achievers (2.5) and moderate achievers (3.8).

Table 2: Average Amount of Concepts Used; Average Amount of Correct Answers; Average Amount of Partly Correct Answers, and Average Amount of Wrong Answers on a Set of Cases

| Results | High achievers | Moderate achievers | Low achievers |
|---|----------------|--------------------|---------------|
| Amount of concepts | 63% (3.0) | 38% (18.7) | 37% (16.7) |
| Amount of correct answers (max 12) | 6.2 (1.3) | 2.6 (2.6) | 2.5 (3.1) |
| Amount of partly correct answers (max 12) | 0 (0.0) | 1.4 (1.1) | 0.2 (0.5) |
| Amount of wrong answers (max 12) | 5.8 (1.3) | 8.0 (2.9) | 9.2 (2.8) |

The differences between the low achievers and the moderate achievers for the amount of concepts as well as for the correctness of the answers were negligible. Comparison between the three groups of students on the amount of partially correct answers yielded important differences. The moderate achievers especially take partly correct decisions (1.4). The low achievers take the most wrong decisions (9.2), although the difference between low achievers and the moderate achievers is small. Table 2 presents the descriptive statistics for the two protocol-analysis criteria: the average amount of concepts used during problem analysis and the correctness of the decisions made.

It can be concluded that the preliminary findings of this study provide some evidence for the criterion-related validity of the OverAll Test.

Study 3: The Influence of Knowledge on Problem-solving

What are the merits of assessing students' knowledge structures when the main goal of instruction is successful problem-solving? To what extent can student knowledge profiles serve as feedback for their problem-solving abilities? The third study investigates the influence of students' knowledge structure on their performance on problem-solving tasks as measured with the OverAll Test.

Rationale

Research into the differences between experts and novices in performance on problem-solving tasks resulted in a profile of successful problem-solvers (Glaser & Chi, 1988; Yekovich, 1993). Smith (1991) summarized the internal factors affecting problem-solving performance. Successful problem solving is enhanced by

- affective variables, including self-confidence, motivation, beliefs, etc.
- the length of prior successful problem-solving experience
- knowledge of the domain from which the problem is drawn (factual, conceptual, procedural)
- knowledge of general problem-solving procedures such as means-ends analysis, trial-and-error, etc.
- knowledge which is adequate, organized, accessible, integrated and accurate (misconception free)
- other personal characteristics such as cognitive development, personality, etc.

The importance of adequate, well-organized and easily accessible conceptual knowledge of the relevant content domain is confirmed by studies of Chi et al. (1981) and Perkins, Schwartz, and Simmons (1988). The knowledge base serves as the basis for the representation, analysis and solving of the problem presented. In addition to this conceptual understanding, the successful problem solver knows what to do, as well as how and when to do it. Problem solving requires procedural knowledge (Smith, 1991). The present study focuses on the influence of students' declarative knowledge on their performance on problem-solving tasks in the domain of economics. If the study provides empirical evidence for the relevance of an organized knowledge structure for successful

problem-solving, examining students' knowledge profile is shown to be a relevant instrument in the regular instructional process aiming at successful problem-solving as well as for remedial purposes.

Research Method

Procedure

The procedure used is sorting concepts. According to Chi et al. (1981) and Shavelson (1974) this method is a valid way to try to provide an answer to the question concerning the degree to which a student's knowledge is structured. The respondents are asked to sort the concepts presented in TOC 1 within the presented eight main themes (see Study 1). Students' results from the sorting task are compared with their performance on the OverAll Test. For TOC 2, the students were not asked to classify but to indicate the level of competency they acquired for each of the presented concepts (see Study 1). This variable is correlated with students' score on the OverAll Test. Finally, students' scores on the Knowledge test, covering the same content domain, are compared with the OverAll Test scores.

Sample

The same sampling procedure is used as in Study 1.

Results

The results of this study confirm previous research results on the influence of an organized knowledge base on problem-solving performance. Correlation coefficients (see Table 3) indicate that the better students are able to classify the concepts of the domain of marketing and organization (TOC 1), the better they are able to analyze and solve problems within this domain.

Table 3: Pearson's Correlation Coefficients Between Students' Sorting Performance and the OverAll Test Scores

| | OverAll Test score (Total %) |
|----------------|---------------------------------|
| CST | 0.49* |
| WST | -0.1338 |
| NST | -0.2452 |
| KT score (C-I) | 0.69** |

CST = correctly sorted topics

WST = wrongly sorted topics

NST = not sorted topics

KT score (C-I) = Knowledge test score (correct-minus-incorrect score)

* statistically significant with a confidence level of 95%

** statistically significant with a confidence level of 99%

The more concepts are wrongly classified, the lower the students' performance on the OverAll Test.

Correlation of student's score on the Knowledge Test with the OverAll Test score is even more convincing. For the TOC 2, students were asked to indicate the level of competency they perceived to have achieved for each of the concepts. Correlation of this variable with students' scores on the OverAll Test indicates that the more concepts are mastered on the level of analysis, the higher the scores on the OverAll Test (see Table 4).

Table 4: Pearson's Correlation Coefficients Between Students' Perception of the Level of Comprehension and the OverAll Test Scores

| Level of Competency | OverAll Test Score (Total %) | OverAll Test Scores/Open- ended Questions |
|---------------------|---------------------------------|--|
| Definition | -0.43 | -0.12 |
| Comprehension | 0.06 | -0.11 |
| Analysis | 0.29* | 0.37* |
| KT-score (C-I) | 0.45** | 0.69** |

KT score: Students' score (correct-minus-incorrect score) on the Knowledge Test

OverAll Test Score: Students' total score (open-ended and close questions) on the OverAll Test

* statistically significant with a confidence level of 95%

** statistically significant with a confidence level of 99%

In summary, a well-organized knowledge base seems to affect successful problem-solving as measured by the OverAll Test. There is some empirical evidence that students' perception of mastering the concepts on the level of analysis relates to their performance on the OverAll Test.

Conclusions

Contemporary cognitive psychology has suggested several changes for instruction and assessment (Calfee, 1995). For example, the importance of knowledge application instead of knowledge consumption and additionally, assessment and instruction must be contextualized, reflective and social. On the basis of these ideas, a lot of schools are looking for and experimenting with alternative ways to develop their curricula. The Maastricht School of Economics and Business Administration introduced a problem-based curriculum, intending to educate competent problem-solvers. Like many schools, the Maastricht school struggled with the choice and the implementation of a congruent assessment system. We chose for two assessment instruments: a Knowledge Test and an OverAll Test. The present article aimed to describe the case of the Maastricht assessment system as an example of assessment within an innovative curriculum. One of the main concerns of the faculty was to gain empirical evidence for the quality of the assessment system in its broad sense. In this way, the article presented a second case: a research methodology that looks for empirical evidence for the quality of assessment innovations.

The three studies presented try to contribute to the discussion about the feasibility of alternatives in instruction and assessment.

I addressed three questions. When introducing student-centred programs, there is a lot of concern about student outcomes. Do students in settings such as problem-based programs actually conduct learning activities that correspond with those intended by the faculty (Dolmans, 1994)? Do the students work on the topics the faculty describes as essential for a competent professional in the field? If not, is it fair to assess students on the basis of the formal goals? To investigate these questions, a Topic Checklist was designed as a map of the formal curriculum. This map was presented to students in order to describe the instructional practice. This map was also used as a blueprint to analyze the assessment goals. The present study suggests there is an important degree of overlap between the formal and the operational curriculum, in terms of concepts studied as well as in terms of the level of mastery intended and achieved. Although learning in the problem-based curriculum is highly self-directed, students do address the issues the faculty describes as essential. Additionally, there is a sufficient congruence between the assessment practices in terms of goals assessed and the formal and operational curriculum. This implies that, even when the students have a considerable amount of freedom within the program, it seems to be possible to make assessment instruments that are fair to the student. Moreover, because of the match between the curriculum and the assessment practices, student outcomes are a relevant source of information about the teaching practices.

The second question concerns one of the main issues associated with performance-based assessment instruments. Even when faculty develops case-based assessment instruments, it remains the question whether a student's performance on the cases has anything to do with professional problem-solving. The second study addressed the criterion validity of the OverAll Test. Are high achievers successful problem-solvers? The preliminary results of the analysis of the think-aloud protocols suggest there is some confirmatory empirical evidence to this question. It seems that it is possible to assess students' problem-solving with assessment instruments based on a set of authentic cases with analysis tasks.

Finally, one of the basic assumptions of the Maastricht assessment practices was addressed: the influence of a student's knowledge profile on his performance on the OverAll Test. A student's performance on a concept-mapping task seemed to be related to his performance on the OverAll Test. Students' performance on the Knowledge Test indicated the same relation between knowledge profiles and OverAll Test performances. These findings confirm the results of research in the field of cognitive psychology: a well-organized knowledge base is important for successful problem-solving.

Considering the findings, some implications for assessment as well as instruction can be formulated. The so-called innovative assessment movement has led to a growing interest in new forms of assessment, e.g., open-book exams, take-away exams, projects, real life tasks, simulation exercises, self and peer assessment. Assessment instruments aiming to measure students' conceptual understanding do not seem to fit in these ideas. They are often condemned as traditional instruments measuring a low cognitive level such as conceptual understanding. However, the results presented here indicate the importance of the measurement of conceptual understanding. They suggest that the breadth of the student's knowledge base and degree of structure are relevant dimensions of assessment.

Although the assessment of problem solving skills is the ultimate goal, we should not relinquish the traditional assessment techniques. Alternative assessment techniques such as the OverAll Test should not replace the Knowledge Test. The use of both instruments enables a triangulation based on a wide range of evidence, thus increasing the quality and the validity of the inferences drawn on the basis of the assessment (Birenbaum, 1996). If diagnosis of the sources of poor problem-solving performance is one goal of assessment, then the assessment should permit identification of the nature and the extent of a student's knowledge. If assessment can uncover more precise deficits in students' knowledge bases, then more specific guidelines for instructional remediation can be made for individuals and groups with similar strengths and weaknesses. Knowledge about the processes and products of successful reasoners coupled with the same knowledge about less successful ones provides some instructional guidance regarding "what to teach" (Brown, Bransford, Ferrara, & Campione, 1983).

For instruction, our results imply that when problem-solving is a main goal, learning environments should be designed with a view to enabling students to acquire a knowledge base which constitutes a sufficient basis to identify, define, analyze and solve authentic problems. The extent to which students reach this goal is an important indicator for the design as well as the review of the learning environment.

For assessment, the findings suggest that feedback should involve two dimensions: the breadth and the depth of a student's knowledge profile and the extent to which this knowledge is usable. No single assessment technique can satisfy both assessment dimensions without presenting a distorted view of student's capabilities (Birenbaum, 1996). Therefore, a variety of assessment tools is preferable to a single tool.

Notes

1. The first year comprises four instructional periods, called *blocks*, each lasting for eight weeks.
2. If the psychometric data (item-test correlation coefficients) do not indicate insufficient quality of the test item itself.
3. After each OverAll Test administration, students filled in a questionnaire asking for their study strategies, the match between instruction and the test, and the difficulties they experienced.
4. At the time of publication, the protocol analysis was not yet finished. Therefore, only preliminary results are presented.

References

- Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum, & F.J.R.C. Dochy, *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-30). Boston: Kluwer.

- Blum, W., & Niss, M. (1991). Applied mathematical problem solving, modelling, applications and links to other subjects. State, trends and issues in mathematics instruction. *Educational Studies*, 22 (1), 7-68.
- Brown, A.L., Bransford, J.D., Ferrara, R.A., & Campione, J.C. (1983). Learning, remembering and understanding. In J. H. Flavell, & E. M. Markman (Eds.), *Carmichaels's manual of child psychology* (Vol. 1, pp. 77-166). New York: Wiley.
- Burger, S., & Burger, D. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practices, Spring 1994*, 9-15.
- Calfee, R. (1983). Establishing instructional validity for minimum competence programs. In G.F. Madaus, *The courts, validity, and minimum competence testing* (pp. 95-114). Boston: Kluwer-Nijhoff.
- Calfee, R. (1995). Implications of cognitive psychology for authentic assessment and instruction. In T. Oakland & R.K. Hambleton (Eds.), *Academic assessment*. Boston: Kluwer.
- Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Chi, M.T.H., & Van Lehn, K.A. (1991). The content of physics self-explanation. *Journal of the Learning Sciences*, 1, 69-105.
- Coulson, R.L., & Osborne, C.E. (1984). Insuring curricular content in a student-directed problem-based learning program. In H.G. Schmidt, & M.L. De Volder (Eds.), *Tutorial in problem-based learning. A new direction in teaching the health professions* (pp. 225-229). The Netherlands: Van Gorcum.
- De Haan, D.M. (1992). *Measuring test-curriculum overlap*. Enschede: Febo.
- De Lange, J. (1992). Assessing mathematical skills, understanding and thinking. In R. Lesh, & S. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 195-214). Washington, DC: American Association for the Advancement of Science.
- Dochy, F. J. R. C., & Alexander, P. A. (1995). Mapping prior knowledge: A framework for discussion among researchers. *European Journal for Psychology of Education*, X, (3), 225-242.
- Dolmans, D. (1994). *How students learn in a problem-based curriculum*. Maastricht: Universitaire Pers.
- English, F.W. (1992). *Deciding what to teach and test*. Newbury Park California: Sage.
- Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, 20, 235-238.
- Feltovich, P.J., Spiro, R.J., & Coulson, R.L. (1993). Learning, teaching, and testing for complex conceptual understanding. In N. Frederiksen, R.J. Mislevy & I.I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

- Flaherty, E.G. (1974). The thinking aloud technique and problem solving ability. *Journal of Educational Research*, 68, 223-225.
- Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research*, 14, 475-483.
- Glaser, R., & Chi, M.H.T. (1988). Overview. In M.H.T. Chi, R. Glaser & M.J. Farr (Eds.), *The nature of expertise* (XV-XXVIII). Hillsdale, New Jersey: Erlbaum.
- Lawson, C. (1992). On the relation between course structure, teaching methods and evaluation procedures in economics. *Assessment and Evaluation in Higher Education*, 17, (1), 1-10.
- Leenders, M.R., & Erskine, J.A. (1989). *Case research: The case writing process*. London, Ontario: University of Western Ontario.
- Leinhardt, G., & Seewald, A.M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement*, 18 (2), 85-95.
- Lesh, R., & Lamon, S. (1992). *Assessment of authentic performance in school mathematics*. Washington, DC: American Association for the Advancement of Science.
- Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, Spring 1994, 5-15.
- Magone, M.E., Cai, J., Silver, E.A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research*, 21 (4), 317-340.
- Mallier, T., Morwood, S., & Old, J. (1990). Assessment methods and economics degrees. *Assessment and Evaluation in Higher Education*, 15 (1), 22-44.
- McClung, M.S. (1979). Competency testing programs: Legal and educational issues. *Fordham Law Review*, 47, 6511-712.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-104). New York: Macmillan.
- Patel, V.L., & Arocha, J.F. (1995). Methods in the study of clinical reasoning. In J. Higgs & J. Mark (Eds.), *Clinical reasoning in the health professions* (pp. 35-48). Oxford: Butterworth-Heinemann
- Pelgrum, W.J. (1990). *Educational assessment: Monitoring, evaluation and the curriculum*. Enschede: Febo.
- Perkins, D.N., Schwartz, S., & Simmons, R. (1988). *Toward a unified theory of problem-solving; A view from programming*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Schoemaker, P.J.H. (1995). Scenario planning: A tool for strategic thinking. *Sloan Management Review*, 25-39.

Segers, M.S.R., Tempelaar, D., Keizer, P., Schijns, J., Vaessen, E., & Van Mourik, A. (1991). *De OverAll Toets: Een eerste experiment met een nieuwe toetsvorm*. [The OverAll Test: A first experiment]. Maastricht: University of Limburg.

Segers, M.S.R., Tempelaar, D., Keizer, P., Schijns, J., Vaessen, E., & Van Mourik, A. (1992). *De OverAll Toets: Een tweede experiment met een nieuwe toetsvorm*. [The OverAll Test: A second experiment]. Maastricht: University of Limburg.

Shahabudin, S.H. (1987). Content coverage in problem-based learning. *Medical Education*, 21, 31-313.

Shavelson, R.J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, 11, 231-249.

Shepard, L.A. (1992). Evaluating test validity. *Review of Research in Education*, 19, 405-450.

Smith, M.U. (1991). A view from biology. In M.U. Smith (Ed.), *Toward a unified theory of problem solving* (pp. 1-19). Hillsdale, New Jersey: Erlbaum.

Spiro, R.J., Coulson, R.L., Feltovich, P.J., & Anderson, D.K. (1988). Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains. In *The tenth annual conference of the Cognitive Science Society* (pp. 375-383). Hillsdale, NJ: Erlbaum.

Swanson, D.B., Case, S.N., & Vleuten, C.P.M. van der (1991). Strategies for student assessment. In D. Boud & G. Feletti, *The challenge of problem-based learning* (pp. 260-274). London: Kogan Page.

Tans, R.W., Schmidt, H.G., Schade-Hoogeveen, B.E.J., & Gijsselaers, W.H. (1986). Sturing van het onderwijsleerproces door middel van problemen: Een veldexperiment. [Directing the learning process by means of problems: A field experiment]. *Tijdschrift voor Onderwijsresearch*, 11 (1), 35-46.

Vilsteren, P.P.M. van, Heijden, M.P. van der, & Arts, A.R.M. (1993). *Het gebruik van casussen in cursussen van de Open Universiteit* [The use of cases in Open University courses]. COP-reeks 9301. Heerlen: Open Universiteit.

Yekovich, F.R. (1993). *A theoretical view of the development of expertise in credit administration*. Paper presented at the 1993 Annual Meeting of the American Educational Research association, Atlanta, Georgia.

The Author

MIEN SEGERS is Associate Professor of Assessment and Evaluation at the Department of Educational Development and Research, School of Economics and Business Administration, Universiteit of Maastricht, The Netherlands. She received her PhD in the field of quality assurance in higher education. Her current research activities focus on the implementation of innovative assessment practices within problem-based curricula.