An approach to the assessment of medical problem solving : Computerised Case-based Testing

AN APPROACH TO THE ASSESSMENT OF MEDICAL PROBLEM SOLVING:

COMPUTERISED CASE-BASED TESTING

Omslagontwerp:
Guusje van Noorden

ISBN 90-9012263-x

Druk: Datawyse Universitaire Pers Maastricht

AN APPROACH TO THE ASSESSMENT OF MEDICAL PROBLEM SOLVING:

COMPUTERISED CASE-BASED TESTING

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht, op gezag van de Rector Magnificus
Prof. Dr. A. C. Nieuwenhuijzen Kruseman,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen op woensdag
16 december 1998 om 16.00 uur

door

Lambertus Wilhelmus Theodorus Schuwirth
Geboren te Heerlen in 1961

Promotor:

Prof. dr. C.P.M. van der Vleuten

Beoordelingscommissie:

Prof. dr. H.G. Schmidt (voorzitter)
Mw. dr. S. M. Case (National Board of Medical Examiners)
Prof. dr. F. Dochy (Universiteit van Leuven)
Prof. dr. R. Grol
Prof. dr. A. Hasman

CONTENTS

VOORWOORD

Dit proefschrift heeft mijn beeld van de 'wetenschap' sterk gewijzigd. Voordat ik er aan begon had ik de, ietwat romantische, voorstelling van een eenzame wetenschapper, die op zijn kamertje in grijze stoffige manuscripten en oude boeken zijn wijsheid bij elkaar peuzelt. Na jarenlang ploeteren en ontberingen lijden zou hieruit dan een meesterwerk naar boven komen.

De werkelijkheid is anders. Het doen van onderzoek en het schrijven van artikelen is een buitengewoon leuke bezigheid. Het proces waarin je probeert de vage ideeën over een onderzoek om te zetten in een operationaliseerbare en onderzoekbare vraagstelling, om vervolgens na de opzet van het onderzoek en dataverzameling een soort terugkerende pakjesavond te beleven: de data-analyse, is echt opwindend. Het vervolgens zo goed mogelijk verwoorden van wat je allemaal gedaan hebt mag dan wel soms ploeterwerk zijn, het is heerlijk ploeterwerk. Toch is dit niet de grootste 'catharsis' geweest. Wat me het meest is opgevallen, is hoe fout het beeld is van de individualist, de eenling, die zijn werk in strikte afzondering doet met alleen af en toe een overleg met zijn promotor. Voor dit proefschrift geldt dit zeker niet; integendeel, niets is minder waar. Hoewel op de voorkant alleen de naam van de promovendus staat, is het aantal mensen dat verantwoordelijk is voor de totstandkoming ervan zo groot dat dit proefschrift eigenlijk een aftiteling in plaats van voorwoord zou moeten hebben.

Ik vind dan ook dat de lezer van dit voorwoord een inzicht zou moeten kunnen krijgen in de manier waarop al die mensen van invloed zijn geweest. Hierbij moet gedacht worden aan mensen die me telkens weer gemotiveerd of gesteund hebben (Ja,.. het produceren van een proefschrift mag dan wel zo leuk zijn, er zijn natuurlijk momenten waarop het even wat minder is). Vele mensen hebben concrete hulp geboden of zijn rechtstreeks betrokken geweest bij casusproductie, dataverzameling, analyse, schrijven en adviseren. Daarom is het misschien goed ze even aan u, als lezer van dit voorwoord, voor te stellen:

"Kom eens even binnen, Lammie, en ga eens even zitten", "Nee, nee, als zo'n manuscript helemaal rood van mij terugkomt, betekent dat juist dat ik er iets mee ***kan***", *"Fantastisch, maar..... ", "Ik heb een ideetje... "en "Doe maar, probeer maar".*

Cees, ik denk dat dit de meest karakteriserende zinnetjes zijn, die ik van je begeleiding heb onthouden. Als promotor heb je een enorme impact gehad op alles wat met dit proefschrift te maken heeft. Ik ben begonnen met de beroemde 'en toen, en toen'-verhalen bij je in te leveren, en op de een of andere wijze heb je me geleerd om in ieder geval enigszins publicabele producten te schrijven. Het leukste daaraan vind ik, dat je dit gedaan hebt door constant van rol te veranderen tussen die van motivator, corrector en stimulator. Volgens mij

zijn dit juist de essentiële eigenschappen voor een docent (want tenslotte is een promotor eigenlijk een docent in 'wetenschap').

"Het is toch eigenlijk bij de wilde konijnen af", "Zeg, moeten we hier niet eens wat meer achteraan zitten".

"Ja is dat eigenlijk wel zo, weten we dat zeker?", "We moeten wel voorkomen dat de auteur, jij en ik een soort folie à trois vormen"

"Och, dan moeten we maar eens kijken of we dat niet op een andere manier kunnen oplossen", en *"Eigenlijk gaat het daar in deze casus niet om"*

Mijn maatjes in de CCT-ontwikkeling moeten hier echt genoemd worden: Jelle Stoffers, Kees de Kock en Emile Mom. Van Jelle heb ik voornamelijk geleerd vasthoudend en feller te zijn als ik weer eens de neiging had dingen te laten slingeren of een binnenbocht te nemen. Dankzij Kees heb ik juist weer geleerd, mijn zekerheden op het gebied van casus schrijven niet zomaar als waarheden aan te nemen, maar altijd eerst goed te overdenken.
"Emile heeft het goed voor elkaar", zeg ik altijd. Je brede kennis van zaken, je humor en je rust (als ik weer eens 4 afspraken tegelijkertijd had gepland) zijn van onschatbare waarde geweest. Met jullie wil ik ook de contactpersonen van de, aan het CCT-project deelnemende vakgroepen bedanken voor hun tomeloze inzet. Christien Pulles, Thea van der Waart, Bram Kroon, Jim van Os, Rob Schonck en Willem-Jan Gerver, ik besef heel goed dat jullie vaak in een hamer-en-aambeeld positie hebben gezeten; tussen de eisen van het project en de tijdsdruk van jullie collega's. Toch hebben jullie altijd je uiterste best gedaan om de productie op peil te houden

"Zo jongeman, wanneer gaan we weer eens casus maken", "Ken je dit mopje al", "Zo ken ik nog een Limerick".

Voor wie hem niet kent: Frans van den Wildenberg: chirurg, traumatoloog.

There was a young surgeon named Frans
Who treated all patients with his hands
But he learned that instead
He could be using his head
Now all patients are lost without a chance

We hebben heel wat uurtjes samen casus zitten schrijven, doorspreken en bijschaven. Vaak bleven we lang stilstaan bij een casus om maar precies naar boven te krijgen wat de essentie was in dat geval. Nog teer ik hierop (en op je enthousiasme). Ik wil je bij deze dan ook bedanken voor een bijzonder fijne en leerzame tijd.

"Goed zo, doe maar", "Wat ben ik blij voor je"

Dat is duidelijk mijn lieve Iris. Toch kan ik de essentie van je hulp en bijstand niet in een citaatje uitdrukken. Dat was namelijk met name het geduldig toehoren als ik je uitlegde waar ik mee bezig was, bij visite en feestjes doordraafde over mijn werk, en je het zelfde verhaal voor de tiende keer moest aanhoren. Ik weet dat ik soms te veel over mijn werk sprak

(spreek??), doordraaf over statistische technieken die ik net heb begrepen en ze dan vervolgens aan jou ging uitleggen. Van doordraven heb je me echter nooit beschuldigd; geduldig heb je geluisterd en met me meegedacht. Je hebt altijd geprobeerd mijn enthousiasme nog verder aan te wakkeren. Daarvoor wil ik je bedanken. Wie weet kan ik het een keertje terugdoen als het voor jou zover is.

"Kennis is macht", "Doe maar goed je best, jong" en vooral *"Wat is het toch fijn dat je het allemaal zo leuk vindt".*

Het klinkt nog na in mijn oren, pap, mam. Ja het is allemaal heel leuk, en ik heb mijn best gedaan. Ik vind het fantastisch hoe jullie me altijd gemotiveerd hebben om door te gaan, om zaken serieus aan te pakken (lukte met mijn wat speelse natuur niet altijd natuurlijk) en om kansen te benutten. Jullie steun was onbetaalbaar, het is ook een beetje jullie boekje.

"Hoe is het met je proefschrift, jong, waar ben je nu mee bezig?", "Lammie, wie geht's uns denn?".

Mijn lieve schoonouders. Het waren heerlijke momenten om na een goede maaltijd, met een kopje espresso en een lekker digestiefje te praten over waar ik mee bezig was en door te filosoferen over alle mogelijkheden die er zijn. We vinden ook na deze promotie vast en zeker andere onderwerpen om over door te filosoferen; onze gemeenschappelijke liefde voor muziek zal hier zeker voor zorgen.

Mascha Verheggen en Ludo van Etten, mijn maatjes in het Electra project. Wat ben ik blij dat jullie het geploeg in deliverables, bimonthlies, meetings, reports, e.d. van me overgenomen hebben. Zonder jullie had dit proefschrift zeker een milleniumprobleem gehad.

Iets waar Robert Peperkamp, Helma van der Linden en Bart Thomas zeker voor gezorgd hebben dat de programmatuur vrij van is. Jullie hebben fantastisch werk geleverd bij het programmeren van alle modules, die nodig waren om de casus überhaupt af te kunnen nemen bij de studenten. Zonder jullie had ik nooit de data voor dit proefschrift bij elkaar gekregen.

Daniëlle Moonen en Françoise Philipi, bedankt voor jullie ondersteuning bij de organisatie en correspondentie van alles wat met CCT te maken heeft. Alexandra en Tonnie bedankt voor jullie steun in moeilijke uurtjes

Jos op 't Root, Albert Scherpbier, Bas Verhoeven en mijn collega's van de capaciteitsgroep Onderwijsontwikkeling en Onderwijsresearch wil ik hier niet ongenoemd laten. Jullie hulp was altijd welkom en is altijd gewaardeerd.

Met mijn lichte hang naar chaos kan het bijna niet anders zijn, of ik ben iemand vergeten. Zoals ik al gezegd heb; er zijn zoveel mensen die van belang zijn geweest bij de totstandkoming van dit proefschrift. Mocht ik iemand vergeten zijn, dan hoop ik dat diegene het me wil vergeven, en me toestaat hem of haar persoonlijk te bedanken, en samen met haar of hem het glas te heffen.

Het besef dringt zich aan mij op dat ik met deze dissertatie aan een eind ben gekomen. In ieder geval zou je het kunnen beschouwen als het laatste academische examen. En zoals ieder examen (geslaagd of gezakt) blijft het een heerlijke aanleiding of excuus om weer eens een goed feestje te geven.

Op jullie aller gezondheid.

Lambert.

INTRODUCTION

COMPUTERISED CASE-BASED TESTING: AN APPROACH TO THE ASSESSMENT OF MEDICAL PROBLEM SOLVING

When in 1974 the first group of medical students entered the medical school of Maastricht, the educational approach was in its early stages. The educational approach "Problem-Based Learning" (PBL) was adopted from McMaster university in Hamilton, Canada (Barrows, 1984; Woodward, 1987). Although quite revolutionary at that time, many of the essential components of this system could be implemented in the Dutch situation without requiring major revisions.

Concerning the assessment system though, the boundary conditions differed substantially from those at McMaster. Unlike the overall educational approach it was necessary to design and build a totally new examination system for the Maastricht situation. Student populations, for instance, were not comparable. Whereas students at McMaster were selected after having had another academic training and by interviews by the faculty, students in the Netherlands were assigned to the different universities by a national committee directly after secondary education. Students at Maastricht were therefore, in general, younger and more randomly selected than medical students at McMaster.

Another and perhaps more important difference lies in the external certification. In Canada the Medical Council of Canada organises national examinations to which all medical graduates are subjected. In the Netherlands no national examinations exist and every university is autonomously responsible for the quality of their graduates.

Due to these differences the assessment system of the Maastricht medical school needed more summative aspects (i.e. assessment to take matriculating or graduating decisions about students) than the McMaster assessment program. In McMaster a near exclusive emphasis could be placed on formative assessment (i.e. assessment for students to steer their own learning). In addition, the medical school in Maastricht felt the necessity to prove the quality

of its graduates, due to the considerable scepticism that existed in the Netherlands about the value of PBL.
For these reasons it was paramount that some 'basic prerequisites' were defined before designing the examination system. This 'basic philosophy of assessment' defined a number of criteria the assessment system should meet. The most important criteria were the following:

- The assessment system must be *congruent* with the rest of the curriculum. This is based on the notion that the nature and the content of assessment strongly influences student learning behaviour (Frederiksen, 1984; Newble & Jaeger, 1983). Since in a PBL system self-responsibility and self-directed learning are considered essential, examinations should not interfere with that. If assessment and education are not congruent, students will be led more strongly by the assessment than by the educational goals.
- The assessment must be *continuous*. Rather than using one final examination to make pass-fail decisions, longitudinal assessment systems should be used. This not only increases the reliability and predictive validity of the assessment, but it also induces a study behaviour that is continuous and regular, preventing peaking and cramming behaviour not uncommon in conventional assessment programs.
- Examinations must not only serve a formal decision-making purpose, but must also provide as much *feedback* as possible. Since self-responsibility and self-directedness are considered essential for PBL, students must be provided with sufficient information about their performance.
- The teacher role and the examiner role should be *separated* where possible. Assessment must not interfere with learning. Therefore, a person serving as a tutor or facilitator must not be an assessor simultaneously. This duality of roles might hinder students from being authentic in a tutorial group and may conceal their incompetencies and insecurities. Apart from that, it can be confusing to teachers too. Their impression during the one task (e.g. tutorial group) may interfere with their performance in another role (e.g. examiner).
- Assessment must be *comprehensive*, and must provide information about the mastery of all important competencies that constitute medical competence. The not uncommon dominance of knowledge assessment should be prevented and complemented by other professional competencies.
- All tests must be monitored by *quality control* mechanisms. Especially committees for test item review and research on the psychometric properties and educational effects of the tests must form an ongoing element in the quality control.

Subsequently the competencies essential to a physician were defined. Based on the prevailing insights with respect to (medical) expertise at that time, competence was subdivided into "knowledge", "skills", "problem-solving ability" and "attitudes". For all four competency domains it was decided that separate assessment instruments should be designed that would match the prerequisite criteria.
The first examination type to be implemented were the block tests. These written tests with (mostly) true-false items are administered at the end of each block period. During a block period a certain topic of medicine is covered in an integrated, multi-disciplinary way. Items in the block test cover the content of the block including all the different disciplines engaged in it. In this way the test is congruent with the content and the multi-disciplinary character of the block. All 'raw' test material is carefully reviewed by a central review committee as a means

of quality control. By using written, structured and centrally reviewed tests, a separation of the teacher and examiner role is realised. Results are reported back to the students using various profile scores to provide optimal feedback. Although in those aspects the block tests met their expectations, using them as the sole (and summative) assessment instrument led to negative side effects. The most important one was that instead of taking self-responsibility for their study, students started to focus on what (they believed) would be in the test and started preparing for the tests strategically. Therefore, another instrument was developed and the block tests became more formative in nature.

This instrument was the progress test. It is a comprehensive 250-item test on relevant medical knowledge administered to all medical students of all classes at the same time. The test is a stratified random selection out of the entire domain of medicine; the stratifications are based on categories (e.g. organ tracti) rather than on disciplines. The test is administered four times per year. Each test passes an extensive review process. The scores of a student on separate tests are combined each academic year to produce a final pass-fail decision. Through this procedure a more continuous assessment of the students' competence is obtained. Since it is not possible for students to determine what will be in the test (because it is a random selection out of the whole domain of medicine), they cannot prepare for it specifically. A regular study pattern is therefore the most efficient one. Detailed reports on their scores are given to the students to provide optimal feedback. As a consequence of the introduction of the progress test, the block tests mainly became formative tests, asking questions on specific knowledge of the block, whereas the progress test was mainly summative with questions aimed at more longitudinal, functional knowledge. The progress test proved to be a valuable addition to the block tests, since it was not only continuous and interdisciplinary, but also because it did not interfere with the self-responsibility and individual learning paths of a student (Van Til, 1998).

For the measurement of skills an Objective Structured Clinical Examination (OSCE) was introduced (Harden & Gleeson, 1979). An OSCE is an observation-based clinical examination during which every student has to enter a number of different rooms. In each room there is a different examiner with a different assignment for the student to perform a particular skill, using simulated patients or models. For every assignment a checklist is constructed that the examiner has to complete when scoring the performance of the student. The content covered by the OSCEs in Maastricht is closely related to the content of a longitudinal skillslab programme. For every class a different OSCE is constructed. Although originally the checklists were very detailed and were mainly used to measure the procedural aspects of skills, this induced unwanted side-effects (e.g. students started memorising old checklists instead of practicing the actual skills). Currently more globalised rating scales are used instead of detailed checklists.

The assessment of attitudes, however important, is still problematic. The character of attitudes is dualistic; it has an internal and an external aspect (Batenburg, 1997). The internal (students' beliefs and feelings) is in fact (still) difficult to measure. Although on many different occasions informal feedback is provided to the students, no formal assessment system has been developed for this aspect so far. The external behavioural side is assessed on various different occasions. In the OSCEs stations on communication skills are included. Judgements of student behaviour towards patients and colleagues are made during the clerkships. In addition to that, a signaling system for attitude problems has recently been introduced. A teacher who encounters a serious attitudinal problem in a student is invited to report this to the

certification committee using a special questionnaire. When multiple reports (from different teachers) are submitted about a single student (s)he is invited for an interview with the certification committee. Based on this interview advice can be given or actions can be taken (up to expulsion).
These components of the assessment program seem to work satisfactorily so far, generally meeting the criteria as were set out. Nevertheless, a deficit still existed in the measurement of problem-solving ability or clinical reasoning. The formal existing assessments were mainly carried out at the end of each clerkship and they varied considerably, ranging from structured orals to written factual knowledge-based tests. Questions have arisen whether these instruments were congruent with the educational goals of the clerkship and whether they provided sufficient feedback. In addition, virtually no inbuilt quality control system was used to ensure the quality of the assessment before it was presented to the students.
In a systematic external review of the Maastricht assessment system, it was concluded that still a lacuna exists (Swanson, 1988). In particular in the testing of problem solving the miscellaneous instruments used were criticised. Apart from some serious flaws with respect to generalisability, validity and influence on student learning behaviour, it was seriously doubted whether the approach to problem-solving as a separate entity is a correct one. It was suggested to structure the assessment of higher order cognitive skills in terms of tasks that physicians perform, rather than hypothetical traits.
As a response to this criticism, it was suggested to design a new method of assessment for the clerkships. This method, Computerised Case-based Testing (CCT), is an assessment method that consists of large numbers of short authentic cases with a limited number of questions asking for essential decisions. These cases and questions are presented to the student using a computer.
The central question of this dissertation pertains to the validity of this method: Does CCT assess medical problem-solving ability? The validity of an assessment method can be approached from different angles. Two mainstream approaches exist in the literature. One of these perceives an examination as a type of psychological test measuring an invisible construct. This implies that validation must be performed using similar approaches as in psychological tests, e.g. by comparing the scores of different test takers or on different tests. This *indirect* validity is strongly advocated by Cronbach (Cronbach, 1983). Others perceive an examination as a set of assignments with a direct and intrinsic meaning, implying that validity must be built into a test, e.g. by careful blueprinting, item writing, etc. This *direct* validity is advocated strongly by Ebel (Ebel, 1983). For both types a large variety of indicators could be studied of which a selection had to be made. It was chosen to focus on two main elements:

- Which cognitive tasks are set by CCT to the candidate?
- What are boundary or enabling conditions with respect to these cognitive tasks?

With respect to the first element the following research questions were addressed.

From a direct validity perspective:
1 What do experts and students judge that the validity of CCT is?
2 Does CCT elicit other cognitive operations than factual knowledge questions?

From an indirect validity perspective:

3 Does performance on CCT increase with increasing expertise?
4 Does CCT yield information about medical competence of students that is not provided by other measures of competence?
5 Is CCT able to detect differences between students of different medical faculties using different approaches to problem solving, better than other tests?

The second element was addressed from a direct validity perspective only.

6 How is CCT embedded in the research on assessment of medical problem solving?
7 What is the influence of question formats on the validity of CCT?
8 What is the influence of the blueprint on the validity of CCT?
9 What are the guidelines for designing valid cases and questions?

Before addressing these questions, however, a brief overview of the research on medical problem solving and of the use of computers in assessment is in order.

MEDICAL PROBLEM SOLVING

During the last half a century considerable change in the thinking about the nature of problem-solving ability has occurred. Some of the original assumptions have been changed more or less dramatically, mainly as a result of psychometric and cognitive psychological research. Some of these changes are:

- *Generic trait versus domain specific state*

An original implicit assumption about the nature of problem-solving ability was that it could be distinguished from factual knowledge as a separate and stable entity. It was viewed as a strategy that, once mastered, could be applied to any applicable situation. Test instruments that were based on this viewpoint typically consisted of complete and integral and sometimes branched patient cases through which the candidate had to work his/her way in order to solve the (patient) problem (Barrows & Tamblyn, 1977; Berner, Hamilton, & Best, 1974; Helfer & Slater, 1971). Psychometric analyses of these tests, however, revealed that intercase correlations were extremely low. In other words, the score a student receives on a given case was apparently a poor predictor for the score on any other case (Elstein, Shulmann, & Sprafka, 1978). Evidence for domain specificity originated also from cognitive psychological research (Chi, Glaser, & Rees, 1982; Glaser & Chi, 1988; Polsen & Jeffries, 1982; Posner, 1988). Problem solving, therefore, must not be viewed as a generic and stable trait but as a highly domain specific characteristic (Anderson, Reder, & Simon, 1996; Norman, Tugwell, Feightner, Muzzin, & Jacoby, 1985; Norman, 1988; Norman et al., 1987; Regehr & Norman, 1996). Large numbers of cases would therefore be needed to reach a generalisable judgement about the competence of the examinees requiring tests of extreme duration (Elstein et al., 1978; Swanson, Norcini, & Grosso, 1987).

- *Algorithmic versus idiosyncratic*

It was assumed that experts solved patient problems using a fixed set of algorithms. Measuring expertise would therefore be based on assessing to what extent the candidate possesses these algorithms. In panels of experts, however, who had to define the correct answer or route through the patient simulations, it proved to be difficult to reach consensus about the optimal pathway or solution to a case, thereby actually challenging the existence of fixed algorithms. Also the weighting of all possible decisions in the case was subject to considerable debate and biases (Swanson et al., 1987). If problem-solving ability really would consist of the possession of a fixed set of algorithms, reaching consensus would have been much easier.

This is true not only for medicine, but even in the research on the problem solving of puzzles the methods applied were found to be at best heuristical (Polsen & Jeffries, 1982). A further demonstration of this idiosyncrasy can be found with Chi et al. who demonstrated that experts are much more idiosyncratic in their determination of essentials of physics problems than novices (Chi et al., 1982). Problem-solving ability must therefore be seen as a quite idiosyncratic ability that is at best heuristical but not algorithmic. In addition, it was found that the way in which contextual information is taken into account by the candidate is an important contributing factor to this idiosyncracy (Hobus, Schmidt, Boshuizen, & Patel, 1987; Regehr & Norman, 1996; Schmidt, Norman, & Boshuizen, 1990).

- *Thoroughness versus efficiency*

Experts were originally assumed to be much more thorough in exploring all the possibilities when solving a problem. In simulation-based examinations in which thoroughness was rewarded, experts were often outperformed by graduates or final year students. This intermediate effect challenged the construct validity of the approach (Schmidt, Boshuizen, & Hobus, 1988)

Expert problem solving, however, is not so much based on thoroughness of information gathering, but instead on more unconscious or automated thought processes including pattern recognition (Norman, 1988; Regehr & Norman, 1996; Schmidt et al., 1990). This finding was again clearly replicated in many other domains (cf. for an overview: Glaser & Chi, 1988; Posner, 1988). Differences in expert and non-expert problem solving would therefore not only consist of differences in the number of correct solutions, but also in differences in the efficiency and the time needed to solve a problem (Regehr & Norman, 1996; Schmidt et al., 1990).

- *More knowledge versus differently organised knowledge*

It has been suggested for a long time that experts mainly possess more knowledge than novices. This is not entirely true, though; the knowledge of experts appears to be organised in a different way too (Chi et al., 1982; Claessen & Boshuizen, 1985; Norman et al., 1985; Regehr & Norman, 1996; Schmidt et al., 1990). The knowledge in experts is more meaningful or more diversely embedded in so-called semantic networks. These networks enable the expert not only to retrieve knowledge faster from his memory, but it enables chunking and consequently faster processing (Chi et al., 1982; Glaser & Chi, 1988; Norman et al., 1985; Posner, 1988; Regehr & Norman, 1996; Schmidt et al., 1990). For this more efficient processing, however, it is necessary that a right cue is given to trigger the semantic networks. In retrieving context-free information, as is required in rote factual knowledge examinations,

experts are often outperformed by novices (Day, Norcini, Webster, Viner, & Chiroco, 1988; Schuwirth, Schrander, & Van der Vleuten, 1993).

- *Question format versus question content*

It has long been thought that the format of the problem or question posed to the candidate determines the sort of competence that is required. In particular on multiple choice questions many studies have been done, looking at the so-called cueing effect (Hettiaratchi, 1978; Hurlburt, 1954; McCarthy, 1966; Newble, Baxter, & Elsmlie, 1979; Ward, 1982). This effect assumes that for answering a multiple choice question, recognition of the correct answer within the distractors suffices to answer the question correctly. In open-ended questions, in which spontaneous generation of the correct answer is needed to complete the question correctly, this cueing effect would not occur. Recognition of a correct answer is not considered to be a higher order cognitive skill, and multiple choice questions were therefore assumed not to be suitable for the testing of problem-solving ability (Hettiaratchi, 1978; Hurlburt, 1954; Newble et al., 1979). None of these studies succeeded, however, in demonstrating a significant effect of question format on the sort of competence measured. A dominant effect of the *content* of the question, though, has been demonstrated repeatedly (Norman et al., 1985; Norman et al., 1987; Ward, 1982).

THE USE OF COMPUTERS IN TESTING OF PROBLEM-SOLVING ABILITY

The logistics involved in high quality assessment are often complicated and the costs are high. Computers could play an important role in increasing the efficiency and quality, by administering examinations, marking the answers, calculating the scores, producing item analyses and enabling item banking. The use of computer technology in assessment has a longer history than is often assumed. The first automated instruction and testing was done during the 1920s (Lumsdaine & Glaser, 1960). The massive enrolment of students into the universities has obviously favoured the logistic advantages of computer use.

The first devices for automated test administration and scoring were highly mechanical devices for both learning and testing. The underlying concept of these devices was closely embedded within the prevailing psychological insights in learning and acquiring expertise. These, mainly behaviouristic psychological insights assumed that higher order cognitive learning was based on the same principles as lower cognitive or subconscious learning, namely frequent and immediate repetition and direct and immediate reinforcement. Therefore these instruments were mainly "practice and drill" devices, probably the most well-known of which is the Skinner machine (Lumsdaine & Glaser, 1960). When the assessment of problem-solving ability became a more popular area of research, computers were used increasingly for simulations and simulated tests.

Large projects have been started to implement long case simulations on a computer (Marshall, 1989; Melnick, 1990; Norcini, J. Keskausas, L. Langdon, & G. Webster, 1986). But as the assumptions about problem-solving ability on which these developments were based were proven to be incorrect (as described above), the developments have not led to a valid

instrument for high stakes assessment yet. In addition, it was difficult to score answers to open-ended questions by computers. This was tackled by using long databases of possible answers with guidance systems that had to lead the students to the correct answer, but these proved either to be too inaccurate or too time consuming. Interpreting and judging natural language by computers on the other hand was, and probably will be for quite some time, too inaccurate to use in a (high stakes) examination (Sabah, 1993).
Developments in the field of informatics have enabled further and broader use of computers in testing. The developments in the internet and databases have made it possible to automatically generate items or cases, to construct tests according to specific psychometric criteria, and to test students on demand at any (remote) site.
A further development lies in the use of computers to increase the authenticity of the test. It is essential that the assignment in the test resembles the assignment in real life as closely as possible (Ebel, 1983). This can be seen as an indispensable part of validity. The presentation of information in a test case should therefore be as congruent as possible with the presentation of the information in real practice. Multimedia offer a relatively inexpensive means to present pictorial and acoustic information. In terms of enhancing the validity of the measure this is not trivial.

CONSEQUENCES FOR CCT

The research and development described above converge in CCT. CCT uses of short cases based on the so-called key-feature approach (Bordage, 1987; Page, Bordage, & Allen, 1995). Different question formats are used, following the rule that each format is adapted to the content of the question and based the number of realistic alternatives in real life. Cases and questions are administered and scored by computer not only to increase the efficiency of the test, but also to enhance authenticity.
The objective of CCT is to provide an instrument to test problem-solving ability of medical students during their clerkship years in a systematic, structured and efficient way. It was intended as a means to fill the gap in the assessment system, i.e. to provide a modernised form of assessment of problem solving in the least structured part of the curriculum, the clerkship years.

STRUCTURE OF THE DISSERTATION

In *chapter 1* the method of CCT is described. A description of the production logistics of case material is presented along with a description of the software developed for storage and administration of cases. Some first general experiences with the method are presented.

In *chapter 2* a study is presented which investigates the extent to which the cueing effect of multiple choice questions interferes with the validity of case-based questions. This chapter further describes the effect the question difficulty on the occurrence of this cueing effect.

In *chapter 3* a comparison is made between an item format specially developed for CCT (computerised long-menu questions) with multiple choice and open-ended questions to assess whether these questions could be an acceptable alternative for machine-scored open-ended questions.

Chapter 4 describes a general study in which the validity of CCT as a measurement of problem solving is assessed. A broad approach to validity is adopted to assess the validity of CCT from different viewpoints reported in the literature (Cronbach, 1983; Ebel, 1983; Flanagan, 1983; Gardner, 1983).

Chapter 5 addresses the validity of CCT in a more direct approach. By using a think aloud protocol methodology a direct comparison between case-based questions and factual knowledge questions is made to evaluate the (differences in) thinking steps used in both types of questions.

Chapter 6 addresses the validity of CCT in a more extrinsic way, by comparing the results of students in a Problem-Based Learning curriculum with those in a traditional one. Since many of these comparisons resulted in a lack of difference, finding a difference would add to the evidence of validity for CCT.

In *chapter 7* item writing guidelines are provided for writing short case-based test items. These guidelines are derived from the experiences of producing cases for CCT during several years.

In *chapter 8* all conclusions are summarised and additional characteristics of CCT (such as reliability, educational impact, costs and acceptability) are discussed. In addition, not only the possibilities and future developments of CCT, but also its limitations and restrictions are discussed.

The research questions were addressed across the different chapters of this dissertation. Research questions 1, 2 (judgments of candidates) and 3 (increase in scores) are addressed in chapter four, whereas research question 4 (comparisons between different tests) is dealt with in chapters four and five.
The question with respect to differences across medical faculties (research question 5) is studied in chapter six. Chapter one describes the embedding of CCT in the research on assessment of problem solving as formulated in research question 6. The next question pertaining to the influence of item format on the validity is studied in chapters 2 and 3. The influence of blueprinting (research question 8) is addressed in chapter four, and the prerequisites for designing a valid case are described in chapter seven.

A final comment on this dissertation must be made. It is based on seven articles most of which have been published or are under editorial review. Every article is written to be read on its own. This inevitably leads to repetitions and overlap across chapters which the busy reader may wish to skip.

REFERENCES

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5 - 11.

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Barrows, H. S., & Tamblyn, R. M. (1977). The Portable Patient Problem Pack: a problem-based learning unit. *Journal of Medical Education, 52*, 1002 - 4.

Batenburg, V. (1997). *Medical students' attitude.* Thesis, University of Utrecht, Utrecht.

Berner, E. S., Hamilton, L. A., & Best, W. R. (1974). A new approach to evaluating problem-solving in medical students. *Journal of Medical Education, 49*, 666 - 72.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7 - 76). Hillsdale NJ: Lawrence Erlbaum Associates.

Claessen, H. F., & Boshuizen, H. P. A. (1985). Recall of medical information by students and doctors. *Medical Education, 19*(1), 61-7.

Cronbach, L. J. (1983). What price simplicity? *Educational Measurement: Issues and Practice, 2*(2), 11-12.

Day, S., Norcini, J., Webster, G., Viner, E., & Chiroco, A. (1988). *The effect of changes in medical knowledge on examination performance at the time of recertification.* Paper presented at the 27 th RIME, Chicago.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7 - 10.

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Flanagan, J. C. (1983). A rational rationale. *Educational Measurement: Issues and Practice, 2*(2), 12.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Gardner, E. F. (1983). Intrinsic rational validity: Necessary but not sufficient. *Educational Measurement: Issues and Practice, 2*(2), 13

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv - xxviii). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*(1), 41-54.

Helfer, R. E., & Slater, C. H. (1971). Measuring the process of solving clinical diagnostic problems. *British Journal of Medical Education, 5*, 48-52.

Hettiaratchi, E. (1978). A comparison of student performance in two parallel physiology tests in multiple choice and short answer forms. *Medical Education, 12*, 290-296.

Hobus, P. P., Schmidt, H. G., Boshuizen, H. P. A., & Patel, V. L. (1987). Contextual factors in the activation of first diagnostic hypotheses: Expert-novice differences. *Medical Education, 21*(6), 471-6.

Hurlburt, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meaning, nouns, verbs, adjectives. *Journal of Educational Research, 47*(8), 561- 76.

Lumsdaine, A., & Glaser, R. (1960). *Teaching machines and programmed learning: A source book*: National Education Association of the USA.

Marshall, J. (1989). Checkup: Computerized home evaluation of clinical knowledge understanding and problem solving. *Teaching and Learning in Medicine, 1*(1), 38-41.

McCarthy, W. H. (1966). An assessment of the influence of cueing items in objective examinations. *Journal of Medical Education, 41*, 263 - 66.

Melnick, D. E. (1990). Computer-based clinical simulation, state of the art. *Evaluation and the Health Professions, 13*(1), 104-120.

Newble, D. I., Baxter, A., & Elsmlie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*, 263-268.

Newble, D. I., & Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education, 17*, 165-171.

Norcini, J. J., J. Keskausas, J., L. Langdon, L., & G. Webster, G. (1986). An evaluation of a computer simulation in the assessment of clinical competence. *Evaluation and the Health Professions, 9*(3), 286-304.

Norman, G., Tugwell, P., Feightner, J., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19*, 344-356.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 86.

Norman, G. R., Smith, E. K. M., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21*, 297-304.

Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194-201.

Polsen, P., & Jeffries, R. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 367 - 411). Hillsdale NJ: Lawrence Erlbaum Associates.

Posner, M. I. (1988). What is it to be an expert? In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xxix - xxxvi). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: Implications for professional education. *Academic Medicine, 71*(9), 988 - 1001.

Sabah, G. (1993). Knowledge representation and natural language understanding. *AICOM, 6*(3/4), 155 - 86.

Schmidt, H. G., Boshuizen, H. P. A., & Hobus, P. P. M. (1988). Transitory stages in the development of medical expertise: The "intermediate effect" in clinical case representation studies, *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 139 - 45). Montreal, Canada: Lawrence Erlbaum Associates.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611 - 22.

Schuwirth, L. W. T., Schrander, J. J. P., & Van der Vleuten, C. P. M. (1993). Assistententoets kindergeneeskunde: een beschrijving van de psychometrische eigenschappen [The national examination for residents in pediatrics: a description of its psychometric properties]. *Bulletin Medisch Onderwijs, 12*(4), 146-51.

Swanson, D. B. (1988). *Review of the assessment system used by the University of Limburg medical school* (PES nr 88 - 22). Maastricht: University of Maastricht.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van Til, C. (1998). *Voortgang in Voortgangstoetsing. [Progress in Progess Testing.]* , University of Maastricht, Maastricht.

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

Woodward, C. (1987). Monitoring an innovation in medical education: The McMaster experience. In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status* . New York: Springer Publishing Company.

CHAPTER 1

COMPUTERISED CASE-BASED TESTING: A MODERN METHOD TO ASSESS CLINICAL DECISION MAKING

SUMMARY

This chapter describes an assessment-system that has been developed to focus on application of knowledge. Its two major features are case-based testing and the use of multimedia and computer tools. The case-material for this testing type is based on the key-features concept, reporting the most relevant characteristics of a case and asking a limited number of questions, each aimed at essential decisions. These cases are produced in daily practice using real patients. Subsequently an extensive review process is used to check for flaws in description, phrasing or answer keys. Cases are stored in an item bank out of which an automated random stratified selection can be drawn, according to a pre-specified blueprint. Cases are then presented to the students by way of a specially developed interface using multimedia when indicated. The chapter further describes advantages, disadvantages and experiences.

INTRODUCTION

Assessment of clinical reasoning has attracted major attention as one of the components of measuring clinical competence (Van der Vleuten & Newble, 1995). In the 1960s, partly resulting from concern about the knowledge orientation of most examinations and the widespread use of multiple choice questions, instruments were proposed to measure clinical reasoning skills (Helfer & Slater, 1971; McGuire & Babbott, 1967; Rimoldi, 1961). These measures typically confront students with patient information, require the examinee to collect further data and to take diagnostic and management decisions. The quality of the decisions and the pathway through the clinical problem were used to judge the clinical reasoning skills of the examinee. The most prominent exponent was the Patient Management Problem (PMP) (McGuire & Babbott, 1967; McGuire & Solomon, 1976) which became popular in many medical schools and certifying agencies. Some types of PMPs used complex branching pathways, rather than merely allowing a fixed or linear route through the problem. Ingenious technical devices (e.g., latent imaging techniques to reveal invisible ink) were used to withhold data until selected by the examinee. Later, alternative approaches such as Sequential Management Problems (Berner, Hamilton, & Best, 1974); Modified Essay Questions and computer simulations (Norcini, J. Keskausas, L. Langdon, & G. Webster, 1986; Williams, Nu Viet Vu, Barrows, & Verhulst, 1984) were proposed. The common denominator in all these instruments is the use of a patient simulation presenting an examinee with a realistic clinical problem. The examinee is required to work through the entire simulation in order to solve the problem.

Research data and practical experience revealed a number of problems with this type of assessment (Swanson, Norcini, & Grosso, 1987). First, the scoring was a concern. Apart from managing the complexity of some of the scoring systems, it appeared to be quite difficult to reach consensus on appropriate pathways, correctness of diagnoses, management decisions, and weighting schemes for rewarding proficiency, thoroughness, and overall competence. Secondly, it appeared that a score on one simulation was not predictive - did not correlate - with a score on another simulation (Norman, Tugwell, Feightner, Muzzin, & Jacoby, 1985). Therefore, a problem of reliability occurred. This low correlation implies that if an examinee were submitted to a random set of different but parallel problems, a different score would be obtained. Consequently, many problems would be required to achieve a reproducible or reliable score, yielding an impractically long test. This was called the content-specificity problem of clinical reasoning. Finally, some findings cast serious doubts on the validity of these methods. The fact that in some studies medical students outperformed more experienced doctors (Goran, Williamson, & Gonella, 1973; Marshall, 1977; Newble, Hoare, & Baxter, 1982) seriously questions the construct validity. Furthermore, scores on clinical reasoning measures, if reliably measured, were found to correlate very highly with regular multiple choice questions (Norcini, Swanson, Webster, & Grosso, 1983). This raises the question whether these measures really tested a different trait than simple factual knowledge.

In recent years, evidence from cognitive psychological research on the development of medical expertise has generated new insights into the nature of clinical reasoning (Schmidt, Norman, & Boshuizen, 1990). The ability to reason is no longer considered a generic trait that generalises across situations, problems or time, with growth characterised as a monotonous process. Rather, expertise develops as a transition from a conceptually rich and rational

knowledge base (acquired from educational experiences) to a non-analytical ability to recognise and handle situations efficiently (acquired from extensive clinical experience). The expert uses a sophisticated form of pattern recognition characterised by speed and efficient use of information. These cognitive processes are highly dependent on the clinical content. Expertise is not a trait but a 'state', depending on the specific problem and personal experience.
These findings had a number of implications for assessing clinical reasoning (Bordage, 1987; Case & Swanson, 1993; McGuire, 1987; Newble, Vleuten, & Norman, 1995; Swanson et al., 1987; Van der Vleuten et al., 1994):

- To overcome the problem of content specificity more content can be sampled if the problem given to the examinee is focussed on critical elements only. Instead of working through each patient case extensively, only the key aspects are focussed on instead and other aspects are skipped, yielding more time to test more problems. This has been called the 'Cambridge-case' or 'key-feature' approach (Bordage, 1987).
- What is being measured is more dependent on the task given to the examinee than on the format used; i.e., the content of the test is more important than the characteristics of the measurement method. In this respect, a distinction must be made between stimulus (the task posed) and response format (MCQ, essay, etc.). The validity of the stimulus will primarily determine the validity of the test; the response format is of secondary importance.
- The stimulus format should be 'conceptually and clinically rich'. The problems given to the examinees should approximate the real world as closely as possible. Clinical details and patient context information (i.e., the way the person looks or walks) provides important triggers for approaching clinical problems.

In this theoretical context, the medical school of the Maastricht University developed an assessment method using these implications as guidelines. Furthermore, the introduction of the method was motivated by the school's wishes to reward higher order cognitive skills more in the assessment system of the problem-based curriculum and the desire to introduce more standardisation in assessment during the clinical rotations in the last two years of the (6 year) program (Van der Vleuten & D. Swanson, 1990). The method has been called Computerised Case-based Testing (CCT). In essence, CCT represents nothing new nor anything spectacular, but it merely synthesises a number of the previous developments in case-based testing. This chapter describes the method in more detail.

COMPUTERISED CASE-BASED TESTING

CCT is a testing procedure using a set of programs implemented on a network of personal computers. It consists of a part that allows the students to log on to the system and to be tested in a particular discipline (student interface) and a part that, in an interactive way, allows the teachers to convert written test material into computer tests (teacher interface).
The system presents to the (authorised) student a test consisting of a pre-specified number of patient cases based upon a discipline-specific blueprint. This blueprint specifies the number of

cases within relevant content areas. Since the cases are randomly selected within these content areas, each student is given a different sample of cases.
The set of programs uses an MS-Windows environment. The case text and graphical or pictorial information (photographs, X-rays, CAT-scans) are revealed in separate windows. A mouse click on an "OK button" opens a third window with the question. Cases may be accompanied by additional multimedia possibilities: pictures, moving images and sounds.
After the question is answered, another question may follow or new case information may be added. When a case is completed, the student cannot return to a previous case, since in randomly generated tests one case may reveal the answer to another case. All items are immediately scored. Case and item scores are filed, expressed as a proportion of correct scores (in %). Immediately after completion of the test, the total score is presented in combination with feedback on group performance. Other feedback (i.e., correct answers, study suggestions), although technically possible has not yet been realised since the primary purpose of the present test is summative.

THE STIMULUS FORMAT

Type of cases

Cases used in CCT do not represent a real time or a branched simulation but are instead, in the key-feature format (Bordage, 1987).

Table 1.1: An example of a case with two questions

Mrs. Van Horn, a 25 year old waitress in a well-known restaurant consults you because of abdominal complaints. She has had a piercing and cramping pain in the upper abdominal region for about two days. It radiates to her back. She also has a serious headache. She does not feel ill, has not had a fever. She is afraid to eat. Although she worked until 12.15 the night before she still slept very badly.
She would like you to prescribe something for her stomach because she has to go to a wine tasting course in two days.
Further history taking reveals that she has never had such complaints before, she has neither heartburn nor eructation. She is somewhat nauseated but has not vomited. Her complaints do not correlate with the type of food she takes. There is neither diarrhea nor constipation.
On physical examination you find nothing except extreme tenderness on palpation in the upper abdomen below the xyphoid .

Question 1: What lab-tests would you order?

You ordered the liver function tests plus amylase. Amylase was increased.

Question 2: This patient must be referred to an internist. True/false

You decide to refer her to an internist.

Question 3: Indicate the most appropriate level of urgency with which this patient must be seen by an internist.

a today
b this week
c this month

A short description of the patient is given and only relevant characteristics, signs and symptoms are reported. A limited number of questions are asked, focussing on essential decisions that have to be made in that particular case. This makes it possible to present many different cases per test, thus covering the domain optimally. An example of a case with a few questions is presented in table 1.1.
All cases are based on real life patients in real practice. This guarantees that the cases are as realistic as possible. A case should have an adequate description allowing the examinee to answer the question correctly. This means not only that the amount of information given

Table 1.2: Some guidelines for making case-based test material.

- Use real life patients as a base for your cases
- Do not make your cases too long or too short. Always check if all the information needed to answer the question(s) is available; eliminate redundant information.
- Make sure that what you write down is a correct representation of all your findings. For instance, do not leave out important negative findings (like "no rebound tenderness")
- Do not make cases easier than real life. Use the most appropriate tool to present findings (e.g. use audio to present a heart murmur, instead of a written description)
- Do not make cases more difficult than real life. For instance, take into account that in real life most corrections can be done even after some days (like a forgotten lab test can still be ordered after several days), whereas in a test situation no correction is possible after completion.
- Use extra resources (pictures, sounds, moving images, etc.) only when indicated. The use of these tools when they are not needed creates an unwanted distraction.
- Show all your material to colleagues for criticism, or reread your material after a "latency" period.

  Questions should be aimed at decisions that have to be taken in this case. It is a waste of the authors' and students' time and energy to use a case description for the testing of general knowledge.
- An examinee should not be able to answer the question, without having read the case. This helps to ensure that the question is really about decisions in the case. Furthermore, the questions should be such that students could be allowed to use textbooks.
- The question must be phrased correctly:
  - do not use semi-quantitative terminology (often, sometimes, usually, etc.)
  - use short sentences
  - focus on one issue per question
  - avoid absolute terminology (never, always, etc.)
- The correct answer must be defensibly correct; a false answer must be defensifly incorrect.
- Choose the question type resembling real life the most (e.g. a multiple choice type of question for labtests, an open-ended question for a simple diagnosis, etc.)

should be optimal, but also that this information should be stated unambiguously. Other possible sources of information, like graphics, sounds and even short films, should be used only on indication. More detailed case development guidelines used are summarised in table 1.2.

Production of cases

All participating (practising) physicians first received a training in which the purpose of CCT was explained, case-writing techniques were trained and common pitfalls of case and item writing were discussed. Subsequently they started writing patient cases they normally encounter during their clinical consultations. This case material is submitted to a two-step review procedure. In the first step, a small committee (2 people) reviews the test material on content and formulation. Critical comments are sent back to the authors with suggestions for alterations. This cycle may be repeated several times before the material is considered satisfactory. In a second step, the material is presented to a larger group (6-8 people) for a further review of the content of cases, questions and answers. Only after having passed this second review successfully is the case stored in an item bank.

THE RESPONSE FORMAT

Questions linked to the cases are directed towards decisions rather than to (factual) knowledge, and are phrased in such a way that no disagreement about their meaning may arise.

So far, twelve different types of questions have been implemented. This enables the author to choose a question type most closely resembling real practice and to use optimally the potentials of the computer. Question types can be divided into multiple choice types of questions, open-ended type of questions and 'compound' questions (using unrolling cases). Four types of multiple choice questions are used (normal single-best answer multiple choice, MCQ; multiple choice allowing more correct options to be chosen, MMCQ ; and true-false with and without a 'don't know' option). All these types are scored in a usual way. In the MMCQ a point is given for each correct decision (either to choose the option or to leave it open) The total points obtained are then divided by the number of options.

A special type of multiple choice is the Multiple Probability Estimate (MPE) (Van Rossum, Briët, Bender, & Meinders, 1990), in which a number of options are presented to the examinee. For each of the options the examinee has to judge the probability on a seven-point scale. To score performance an expert-panel's score is used as a gold standard. Student scores are calculated based on the distance of the examinee's answer to the mean expert rating. An example of a case with an MPE is given in figure 1.1.

Three types of open-ended questions are used. Two types use a dialogue box in which the examinee has to enter text only. The efficiency of the algorithm for scoring the answer, though, is too limited to be used in high stakes examinations. Most algorithms for understanding natural language suffer from the same problem (Sabah, 1993; Schuwirth, Van der Vleuten, & Donkers, 1991). Based on the work of Veloski (Veloski, Rabinowitz, & Robeson, 1993), two computerised forms of long-menu questions have been created, one for single answers and one for multiple answers. Both use the same principle: students type in their (short) answers in a dialogue box, and the program searches for 'hits' in an extremely long list (over 2500 alternatives). While typing, the examinee sees whether the answer has

been recognised by the program. More information of these questions, including some of the characteristics and some psychometric qualities of these question types have been reported elsewhere (Schuwirth, Van der Vleuten, & Donkers, 1995).

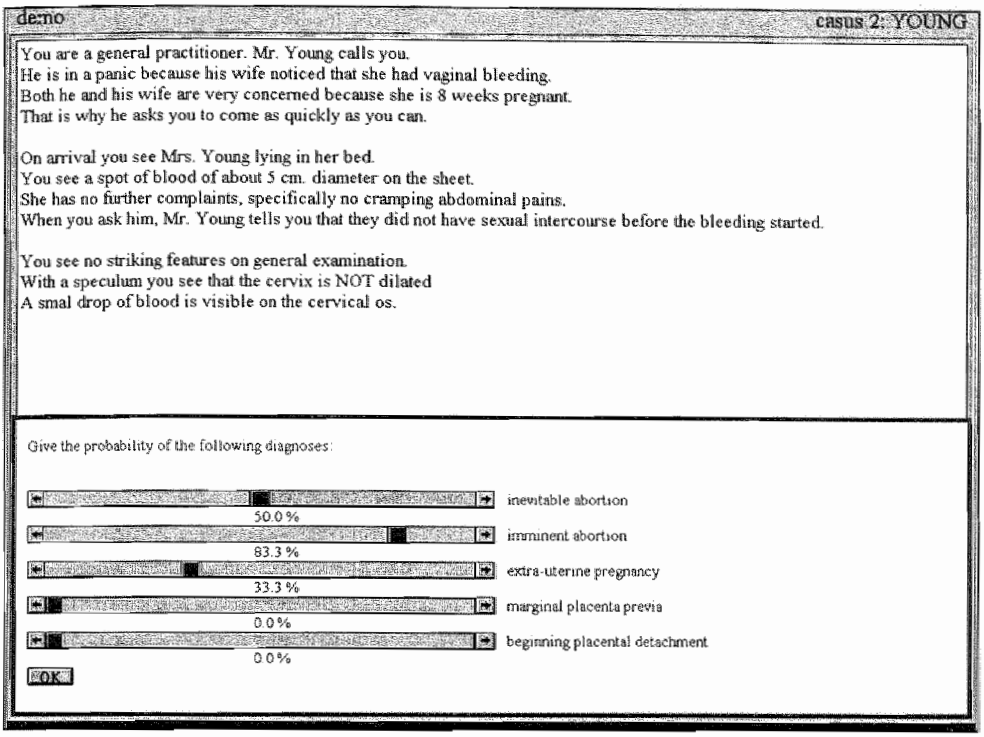


Figure 1.1: An example of a Multiple Probability Estimate Question. The student uses the mouse to drag the button to the appropriate probability of each of the alternatives

Finally three compound types of question types have been implemented. They all have in common that information to the examinee is provided in steps, and the same question is repeatedly asked. In one of these types, for example, the first part of a case reports the initial complaints. A question then asks for the options that should be included in the differential diagnosis. Further segments of the case then reveal the history, physical examination and further investigations. After each part the same question is repeated, but the number of correct options diminishes. Scoring in these question types is based on the proportion of correct answers given divided by the total number of options that had to be considered during the case. Other scoring options are currently under investigation.

THE COMPUTER PROGRAMS

Apart from the interface for the students presenting the cases and questions and recording answers, which has been described above, the set of programs contains an interface for the

teacher and an item bank managed by a database management system (DBMS) for storing and retrieving test material and students' scores. The teacher interface allows the teacher, in an interactive way, to convert all written test material to files that can be used by the computer. This involves application of the specially developed syntax, linking graphics to cases, etc. The student and teacher interfaces gave been implemented, using Borland Pascal and Borland C++ for Windows. The item bank is managed by Sybase SQL Server 10.0.1.

Why the use of computers

The decision to use the computer in CCT was made on the basis of a number of perceived advantages.

Many of the resources normally used in examinations may be saved by using a computerised system for administration and scoring of tests. Particularly when used in combination with automated storing and retrieving of test material. Ideally, these resources will be redirected towards maintaining or improving the quality of test material.

The use of multimedia tools can help to improve the fidelity of the cases. In written tests only pictures can be used, the reproduction of which is quite expensive. The computer may not only use moving images and sounds but can also reproduce images repeatedly without generating further costs.

Furthermore some typical advantages of computers can be used. Firstly, cases providing information in steps, revealing the correct answer to the previous question prior to asking the next may make both questions independent. In all types of written examinations this is not possible. Using computers, this is simply realised by preventing students from going back to a previous screen. Secondly, other outcome variables than the usual proficiency scores can be assessed including response time or number of corrections. This allows exploration of other measures of problem solving as suggested by Norman (Norman, 1988). Thirdly, by the random selection of tests every student receives a different test, and cases (and scores) remain within the item bank. This diminishes security threats.

Finally, the computer enables sequential testing. This is a testing procedure in which a limited number of cases are administered to all examinees. For those examinees whose scores are within a chosen interval from the cutoff score, an additional set of cases is presented. The others are excused from further testing. The number of cases, i.e., the length of the test, varies according to the ability of the candidate. The procedure optimises reliability and efficiency.

On the other hand, computer illiteracy of some of the students, the initial costs for software and hardware, the difficulty (or even impossibility) to score (long-answer) open-ended questions and the fact that taking an examination from a computer screen is more fatiguing as opposed to written examinations can be considered to be the main disadvantages.

THE EXPERIENCES SO FAR

Until now in only one discipline (general practice) have sufficient cases (about 500) been produced to enable the selection of a sample large enough to adequately cover the domain. In two other disciplines (gynaecology and internal medicine), the production of cases was started, but progresses only slowly. All authors indicate that the amount of time associated with production of high quality cases is burdening. Moreover, the number of corrections

suggested by the review committee is substantial and requires considerable additional time. As experience of the case writers increases, however, the production time decreases to some extent. Production of good test material remains nevertheless a time consuming activity. Despite this, case authors and other clinicians involved are very positive about CCT. They consider it a valuable additional tool to other, more knowledge-oriented, testing methods.
The limited psychometric information available indicates that scores from students at the end of a clinical rotation are well above chance level. The number of cases per hour of testing time is approximately 45. The reliability was modest (0.32 per hour of testing time), but may have been suppressed by the homogeneous nature of the group of students tested so far.
Student evaluations using an evaluation form with open-ended questions indicate that nearly all students like CCT. They perceive the content of CCT highly relevant and adequately reflecting the practice in the rotation. Negative comments were related to the impossibility to correct an answer on previous cases and the fact that working on computer screens is fatiguing.
So far, we are content with CCT. With further developments more research will be carried out. Not only should methods be found to improve reliability of the test scores (e.g., by using sequential testing methods) but also validity issues have to be addressed to assess whether using this method really has advantages over using more knowledge-based tests.

REFERENCES

Berner, E. S., Hamilton, L. A., & Best, W. R. (1974). A new approach to evaluating problem-solving in medical students. *Journal of Medical Education, 49*, 666 - 72.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Case, S. M., & Swanson, D. B. (1993). Extended-macthing items: a practical alternative to free response questions. *Teaching and Learning in Medicine, 5*(2), 107 - 115.

Goran, M. J., Williamson, J. W., & Gonella, J. S. (1973). The validity of Patient Management Problems. *Journal of Medical Education, 48*, 171 - 7.

Helfer, R. E., & Slater, C. H. (1971). Measuring the process of solving clinical diagnostic problems. *British Journal of Medical Education, 5*, 48-52.

Marshall, J. (1977). Assessment of problem-solving ability. *Medical Education, 11*(329 - 34).

McGuire, C. (1987). Written methods for assessing clinical competence. In I. R. Hart & R. M. Harden (Eds.), *Further Developments in Assessing Clinical Competence* (pp. 46-58). Montreal: Can-heal Publications.

McGuire, C. H., & Babbott, D. (1967). Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement, 4*, 1 - 10.

McGuire, C. H., & Solomon, C. M. (1976). *Construction and use of written simulations.* Chicago: The Psychologic Corporation.

Newble, D., Hoare, J., & Baxter, A. (1982). Patient Management Problems, issues of validity. *Medical Education, 16*, 137 - 42.

Newble, D. I., Vleuten, C. P. M. v. d., & Norman, G. R. (1995). Assessing clinical problem solving. In J. Higgs & M. Jones (Eds.), *Clinical reasoning in the health professions* (pp. 168-78). Oxford: Butterworth-Heinemann Ltd.

Norcini, J. J., J. Keskausas, J., L. Langdon, L., & G. Webster, G. (1986). An evaluation of a computer simulation in the assessment of clinical competence. *Evaluation and the Health Professions, 9*(3), 286-304.

Norcini, J. J., Swanson, D. B., Webster, G. D., & Grosso, L. J. (1983). *A comparison of several methods of scoring patient management problems.* Paper presented at the 22nd Annual Conference of Research in Medical Education, Washington.

Norman, G., Tugwell, P., Feightner, J., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19*, 344-356.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 86.

Rimoldi, H. J. A. (1961). The test of diagnostic skills. *Journal of Medical Education, 36*, 73 - 9.

Sabah, G. (1993). Knowledge respresentation and natural language understanding. *AICOM, 6*(3/4), 155 - 86.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611- 22.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1991). Het gebruik van open en gesloten vragen: Effecten van cueing en nauwkeurigheid van computergestuurde scoring [The use of open-ended and multiple-choice questions: cuing effects and accuracy of computerized scoring systems.]. In C. P. M. Van der Vleuten, A. J. J. A. Scherpbier, & M. C. Pollemans (Eds.), *Proceedings van het eerste Gezond Onderwijs Congres* (pp. 312 - 8). Houten/Zaventem, Nederland: Bohn, Stafleu en Van Lochum.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1995). Computerized long-menu questions, an acceptable un-cue-version. In A. I. Rothman & R. Cohen (Eds.), *The sixth Ottawa Conference on Medical Education* (pp. 178 - 81). Toronto: University of Toronto Bookstore Custom Publishing, Toronto, Canada.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van der Vleuten, C. P. M., & D. Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58 - 76.

Van der Vleuten, C. P. M., & Newble, D. I. (1995). How can we test clinical reasoning? *The Lancet, 345*, 1032-1034.

Van der Vleuten, C. P. M., Newble, D. I., Case, S. M., Holsgrove, G., McCann, B., McGrae, C., & Saunders, N. (1994). Methods of Assessment in Certification. In D. I. Newble, B. Jolly, & R. Wakeford (Eds.), *The certification and recertification of doctors, issues in the assessment of clinical competence* (pp. 105 - 25). Cambridge: Cambridge University Press.

Van Rossum, H. J. M., Briët, E., Bender, W., & Meinders, A. E. (1990). *The transfer effect of one single patient demonstration on diagnostic judgement of medical students: for better and worse.* Paper presented at the Teaching and Assessing Clinical Competence, Groningen.

Veloski, J., Rabinowitz, H., & Robeson, M. (1993). A solution to the cueing effects in multiple choice questions: the Un-Q-format. *Medical Education, 27*, 371-375.

Williams, R. G., Nu Viet Vu, C., Barrows, H. S., & Verhulst, S. (1984). Profile of the clinical reasoning test. In H. Schmidt & M. De Volder (Eds.), *Tutorials in problem-based learning*. Assen: Van Gorcum.

CHAPTER 2

A CLOSER LOOK AT CUEING EFFECTS IN MULTIPLE CHOICE QUESTIONS

SUMMARY

This study investigates the cueing effect occurring in multiple choice questions. Two parallel tests with matching contents were administered. By means of a computer program examinees of different training levels and professional expertise were presented the same set of 35 cases (derived from patient problems in general practice) twice. The first time the cases were linked to open-ended questions; the second time they were linked to multiple choice questions. The examinees comprised 75 medical students from three different years of training, 25 residents in training for general practice and 25 experienced general practitioners. Across groups, total test scores reflected a difference in mean scores on both formats, and a high intertest correlation. Within each level of expertise differences in mean scores and high correlations were also found. The data are further explored per group of examinees. Two types of cueing effects were found, positive cueing (examinees were cued towards the correct answer) and negative cueing (examinees were cued towards an incorrect answer). These effects were found at all levels of expertise and in almost all items. Both effects, however, decline with increasing level of expertise. Positive cueing mainly occurs in difficult items, whereas negative cueing mainly occurs in easy items.

INTRODUCTION

Pass-fail decisions in medical examinations have major implications for both the candidates and society. This requires examinations to be highly reliable. The aim for high reliability has been pursued mainly by adding more structure to test items. But, adding more structure to test items tends to diminish their reality similitude or fidelity. According to some authors, this is particularly the case if the objective is to assess higher order cognitive skills (Newble, Baxter, & Elsmlie, 1979). One reason for this is that multiple choice questions were mostly used to test simple factual knowledge. Nevertheless, since then researchers have come to realise that it was possible to test "understanding" by means of multiple choice questions, and many of these questions have been developed and used since (Elstein, 1993). Another reason, though, why multiple choice questions were considered unsuitable to test higher order cognitive skills is that recognition of the right answer suffices to answer the question correctly. In open-ended questions, on the contrary, spontaneous generation of the right answer is a requisite for answering the question correctly. The recognition of the correct answer, on seeing it in a row of distractors, is called the cueing effect. It is also because of this cueing effect that multiple choice questions are believed to test a lower level of cognitive skills than open-ended questions (Elstein, 1993; Newble et al., 1979). Attempts to assess the amount of cueing occurring in multiple choice questions are complicated by the fact that cueing may be confounded with guessing. In studies focussing on outcome rather than an introspection of the examinees, it is virtually impossible to disentangle both effects. However, the problem of guessing applies, in part, to open-ended questions too; correct answers may be based on any problem-solving strategy ranging from sheer guessing through educated guessing to a fully competence-based deduction, the amount of which cannot be established. Other disadvantages of open-ended questions have been reported, mainly in terms of ambiguity and lower reliability (Case & Swanson, 1993).

Many studies have been performed comparing scores on a multiple choice test and an open-ended test. Most of these studies used parallel tests (Newble et al., 1979; Norman et al., 1987; Page, Bordage, Harasym, Bowmer, & Swanson, 1990; Stalenhoef-Halling, Van der Vleuten, Jaspers, & Fiolet, 1990), but identical subtests have also been used (Veloski, Rabinowitz, & Robeson, 1993). The results of practically all these studies indicated a higher mean score on the multiple choice subtest (Case & Swanson, 1993; Newble et al., 1979; Page et al., 1990; Veloski, Rabinowitz, & Robeson, 1986), although the opposite was found in one study (Stalenhoef-Halling et al., 1990). The latter may be explained by the scoring system used in that study; it included subtraction of incorrect responses from the total test scores.

There are two consistent findings that contrast with the assertion that the format of the question limits the order of cognitive skills that can be measured. Firstly, intertest correlations are invariably high (Case & Swanson, 1993; Maatsch & Huang, 1986; Maatsch, Huang, Downing, & Barker, 1983). Secondly, the content of the question appears to influence the amount of cueing occurring. Questions asking for diagnostic skills generate a more prominent cueing effect than questions asking about laboratory or management skills (Swanson, Norcini, & Grosso, 1987). The first finding raises the question whether or not open-ended questions add unique information to that already obtained by the multiple choice questions. The second may indicate that apart from the format of the question, its content may influence cueing.

The consistent direction of the mean effect suggests that cueing predominantly triggers towards the correct answer in the multiple choice question. The opposite, however, could theoretically occur also. Participants giving the correct answer on an open-ended question could be pointed to a wrong answer by the distractors in the multiple choice question. To make the distinction between both types of cueing, we will introduce the terms *positive* and *negative* cueing. The former means cueing in favour of the multiple choice question, the latter means cueing in favour of the open-ended question.
The first aim of this study is to investigate the amount of positive and negative cueing occurring in multiple choice questions. The second aim is to assess the influence of level of expertise and the difficulty of the question on the magnitude of the cueing effects. We expected that positive cueing would primarily occur in more difficult items and in test scores of examinees of a lower level of ability. Conversely, for easy items and at a higher level of expertise the (relative) amount of negative cueing was expected to be higher.

METHOD

In this study two identical tests, each containing 35 patient cases, were used. Each case was followed by one question. Both tests were presented to the examinees by the use of an interactive computer interface. This interface presented the cases and questions on the screen. Questions could be answered by clicking a mouse button on an alternative in the multiple choice question or by typing in the answer in a dialogue box when an open-ended question was asked. The computer prevented the examinee to go back to a previous case.
In the first test the examinees had to complete the open-ended questions (OEQs) linked to each case and asking for the most probable diagnosis. These cases were in a fixed order. Immediately after that, the second test was presented, but now linked to multiple choice questions (MCQs) prompting for the most probable diagnosis. These MCQs provided 4-8 possible answers. The cases in this test were in a fixed, though different, order too. The computer program prevented the examinees from returning to a previous screen to alter answers already given.
All cases were derived from real patients from general practice. Their transcriptions to paper cases were based on concepts of the key-feature approach (Bordage, 1987). This implied a description of all relevant signs, symptoms and findings of a patient, with questions prompting for the essential elements of a case only. Prior to the administration, the cases, answering keys and multiple choice alternatives were judged by three experienced general practitioners, and based on their comments some minor adjustments were made. Questions asking for diagnosis only were chosen because cueing appears to be most prominent in this area (Swanson et al., 1987).
The examinees were requested to give one best answer in the deepest level of detail possible (e.g., pneumonia, but if possible bacterial pneumonia or even pneumococcus pneumonia). In addition to that, they were explicitly instructed to consider all the alternatives given in the multiple choice questions, even when they recognised the case from the open-ended set and remembered the answer given there. The test was experimental only, and had no (educational) consequences for the participants. Examinees were encouraged, though, to use the same

strategy as they would use in a real test. All participants were volunteers and received a financial compensation for their efforts.
Table 2.1 shows an example of one of the cases, including the two possible question formats.

Table 2.1: An example of a case with the two question formats.

Peter Johnson (3 years old) suddenly had a fever of about 38,7 °C and developed a sudden rash.
It started, according to his mother, with little red spots that grew into little fluid-filled bullae.
On physical examination you see, indeed, the little red spots and the little fluid-filled bullae which have a greyish colour.
On the chest the exanthema is more severe than on the arms and legs.

OEQ: What is the most probable diagnosis ?

MCQ: Which of the following diagnoses is the most probable one ?

1. rubella	4. varicella
2. scarlatina	5. urtica
3. exanthema subitum	6. exanthema infectiosum

One hundred and twenty-five examinees were used with five different levels of expertise, 25 randomly selected second year, 25 fourth year and 25 sixth year medical students (from a six year medical programme). Also, a group of 25 residents in training for general practice (after 1 year of training in a two-year training programme) and a group of 25 experienced general practitioners (mean duration of active practice 11.3 years) were used, all randomly selected. All students came from a problem-based medical school and were used to being confronted with patient cases for problem-solving exercises. Therefore a higher ability to solve cases could be expected in students having completed more years of their studies.
All answers were recorded and filed by the computer. The open-ended questions were hand-scored afterwards according to a previously fixed answering key. All test scores were expressed as percentages of correct scores.
Potential bias effects that may influence the answers to the cases in the second test, like activation of prior knowledge, may be present, but their presence or magnitude could not be determined in this design.

RESULTS

In table 2.2 descriptive statistics of both tests are provided. There is an approximately equal increase in mean scores on both the OEQ test and the MCQ test with increasing level of expertise, except for the general practitioners. A question format within subjects by level of training between subjects two-way ANOVA was used to establish the significance of both effects. Both main effects were statistically significant: level of training (F=134.09, df=4,

$p<.0001$), and question format ($F=110.46$, df=1, $p<.0001$). The interaction effect level of training by question format was also significant ($F=3.26$, df=4, $p=0.014$).
Reliabilities of the MCQs are lower than those of the OEQs in the student groups, whereas the opposite occurs in the residents and general practitioner groups. What is striking, is the sudden drop in reliabilities in the last two groups.
Due to the unreliability of both tests the observed correlations reported here, reflect an underestimation of the real association. Therefore, in test format comparisons using correlations, Norman et al. (Norman, Swanson, & Case, 1996) suggest a correction for attenuation to estimate the true correlations. To apply this correction here is only partly correct. Since both formats tested similar content, attenuation can only be caused by general error and not by content differences. The true correlations, therefore, will reflect an overestimation of the real correlation. To obtain an estimation of the magnitude of this overcorrection a rough and ready method was used. Both tests were randomly divided into two subtests, yielding four separate tests. This made it possible to calculate three types of correlations: between subtests with similar content and different format (r_{xy}: .77 and .89, p<0001,), between subtests with different content and equal format (r_{xy}: .78 and .80, p<0001) and finally between subtests of which both content and format are different (r_{xy}: .73 and .81, p<.0001). Of all correlations reported here the true correlations were 1.0 except for one (r_{xy}: .73 R_{xy}: .96). Although this method only marginally reflects a true split-half design it seems to indicate that overcorrection of the correlations is not substantial.

Table 2.2: Descriptive statistics of both tests and correaltions

	OEQ		MCQ		Correlation	
Level of training	mean (SD)	Alpha	mean (SD)	Alpha	observed	true
2	34.17 (10.4)	.59	43.89 (10.2)	.51	.60	1.00
4	59.43 (10,4)	.60	70.74 (8.5)	.41	.41	0.83
6	71.54 (10.5)	.64	75.89 (9.0)	.50	.65	1.00
GP trainee's	80.57 (6.2)	.14	86.40 (6.1)	.33	.67	1.00
GPs	77.14 (6.6)	.14	83.54 (6.9)	.36	.64	1.00
Total	64.57 (19.1)	.88	72.09 (17.3)	.87	.90	1.00

To determine the percentage in which positive and negative cueing occurred a cross-tabulation was made for each item were made of all four possible combinations of answers. Subsequently these results were aggregated across all 35 cases and examinees per group. Table 2.3 provides these percentages, in addition to the total amount of cueing appearing (positive plus negative cueing) and the net cueing effect remaining (positive minus negative cueing).

Table 2.3: Percentages of OEQ and MCQ answer combinations per level of expertise.

Level of training	Both answers correct	Both answers wrong	Positive cueing	Negative cueing	Total cueing	Net cueing
2	25.7	47.7	18.2	8.5	26.7	9.7
4	54.4	24.2	16.3	5.0	21.3	11.3
6	64.7	17.3	11.2	6.9	18.1	4.3
GP Trainee's	76.0	9.0	10.4	4.6	15.0	5.8
G Ps	71.4	10.7	12.1	5.7	17.8	6.4
Total	58.4	21.8	13.6	6.1	19.9	7.5

As table 2.3 shows, positive as well as negative cueing occurs at all levels of expertise. The amount of cueing covaries with expertise rather closely. With more expertise, cueing diminishes, except for the residents and the general practitioners (GPs). However, judged by their test scores, the residents had more expertise than the GPs. The positive cueing is a more prominent effect than the negative cueing. Across groups the total cueing effect is rather large (approximately 20 %). The net effect of cueing (i.e. the remaining cueing when negative cueing is subtracted from positive cueing) is, however, considerably smaller (approximately 7 %). The net cueing effect will produce differences in the total scores on both tests.

At the item level, similar comparisons (not reported) revealed positive as well as negative cueing in nearly all items. In two items there was positive cueing only, there were no items in which only negative cueing occurred. Calculating the net cueing effect per item showed that in 28 items a positive net cueing and in 6 items a negative net cueing effect remained. In one item the amount of positive cueing was equal to the amount of negative cueing.

As it was expected that item difficulty influences the amounts of cueing, items were divided into two categories, difficult and easy. In each expertise group the mean p-score on the open-ended questions was used as the cutoff score. Results are shown in table 2.4.

Several aspects may be of interest in this table. Again the difference in net and total cueing is striking, indicating that using parallel tests would underestimate the magnitude of the overall cueing effect considerably, this could be regarded as hidden cueing. To obtain an impression of the magnitude of this hidden cueing, the net cueing is divided by the total amount of cueing and recalculated to a percentage. Subsequently this percentage is subtracted from 100%. The resulting percentage, reported in the table as "$P_{\text{hidden cueing}}$", represents therefore the proportion of the total cueing that can not be seen in parallel test comparisons.

Table 2.4: Percentages of different types of cueing in easy and difficult items.

Level of training	Difficulty	Number of Items	Positive cueing	Negative cueing	Total cueing	Net cueing	$P_{hidden\ cueing}$
2	Difficult	19	20.8	4.6	25.4	16.2	36.3
	Easy	16	15.0	13.0	28.0	2.0	92.9
	$P_{e/(e+d)}$				*52.4*	*11.0*	
4	Difficult	18	25.3	4.2	29.5	21.1	28.5
	Easy	17	6.8	5.9	12.7	0.9	92.2
	$P_{e/(e+d)}$				*30.1*	*4.1*	
6	Difficult	14	18.3	7.7	26.0	10.6	59.2
	Easy	21	6.5	6.3	12.8	0.2	98.4
	$P_{e/(e+d)}$				*33.0*	*1.9*	
GP trainee's	Difficult	15	18.9	5.6	24.5	13.3	45.7
	Easy	20	4.0	3.8	7.8	0.2	97.4
	$P_{e/(e+d)}$				*24.1*	*1.5*	
GPs	Difficult	13	23.4	3.1	26.5	20.3	23.4
	Easy	22	5.5	7.3	12.8	-1.8	14.1*
	$P_{e/(e+d)}$				*32.6*	*8.1**	

* These values were calculated disregarding the negative sign.

It is seen in both difficult and easy items and in all expertise levels. The large difference between hidden cueing in difficult items and in easy items indicates that much more cueing remains hidden in easy items compared to difficult items. So parallel tests comparisons may well overestimate the influence of item difficulty on the amount of cueing occurring. A more direct way of showing this was used by calculating the proportion of cueing that can be attributed to the easy items. This was calculated by dividing the percentage of cueing in easy items by the sum of cueing in easy and difficult items. In the table these figures are in the rows $P_{e/(e+d)}$ and indicated using italics. Even in this tighter representation the amount of overestimation of the influence of item difficulty on cueing effect is clearly visible.
The net effect in the easy items of the GP-group was negative. In the calculation of $P_{hidden\ cueing}$ and $P_{e/(e+d)}$ the negative sign was disregarded, because the outcome still represents a visible portion of the cueing effect.
Neither in the overestimation of the influence of item difficulty nor in the amount of hidden cueing a clear tendency could be found with regard to the level of training.

DISCUSSION

The results of previous studies investigating the cueing effect were reproduced in this study: differences in mean scores on both question formats have been found, and correlations between scores on both formats were relatively high.
A more detailed exploration of the nature of the cueing effect yielded a bidirectional effect, which sometimes inflates scores on the MCQs (positive cueing) and sometimes deflates the MCQ scores (negative cueing). The positive cueing effect is considered as a disadvantage of the MCQ and has been documented consistently. However, MCQs apparently also cue in an opposite direction; they lead the examinee to choose the wrong answer.
The overall cueing effect is quite sizable. In approximately 20% of all answers given cueing plays a role, either positive or negative. So in 20% of the cases items disagree as to the ability of the examinees. This amount of disagreement is only marginally reflected in the total scores where both effects are partly neutralised. Here, a net effect of approximately 7% remained.
There was a clear relation between cueing and difficulty of items. Easy items tend to show more negative cueing; difficult items elicit more positive cueing. This has lead to a striking difference in total and net cueing appearing in easy and difficult items. The result of this would be that using parallel tests severely underestimates the amount of cueing occurring, and overestimates the influence the item difficulty has on the magnitude of the cueing effect.
The negative cueing effect as such is a rather peculiar effect as it is normally expected that it would lead to higher scores on the multiple choice part than the open-ended part. We assume that sometimes with an OEQ question the cases described may be perceived as rather straightforward, while with the MCQ the distractors may lead the examinee to believe that the case is more difficult, therefore tempting him or her to select a more unlikely answer.
The overall cueing effect tends to diminish as level of expertise increases supports this interpretation. This effect may explain the loss in reliability which was found for the MCQ in relation to the OEQ. The cueing effect introduces more "noise" in the measurement. The fact that in the resident and GP groups the MCQs proved more reliable than the OEQs, however, is not in accordance with this explanation.
Although the validity of both formats was not the prime interest of this study, a few remarks can be made. Both formats seemed to discriminate the levels of expertise equally effective. Nevertheless, a few subtleties can be noted. In both tests the average score of the residents exceeded the GP scores. This has been seen and documented before (Day, Norcini, Webster, Viner, & Chiroco, 1988; Grant & Marsden, 1988; Van Leeuwen, Pollemans, Düsman, Van der Vleuten, & Grol, 1993). The phenomenon may be explained by findings in cognitive psychological research of clinical expertise. A consistent finding here is the "intermediate effect" (Schmidt, 1990; Schmidt & Boshuizen, 1993). This intermediate effect supposes a shift in the storage and retrieval of knowledge in the process of becoming an expert. This shift implies a change from a more fragmented to more encapsulated or compiled pathway in storing and retrieving knowledge, thus leading to a different approach to solving patient problems. Accepting this intermediate effect as a logical and valid phase of expertise, the OEQs in this experiment were slightly better at differentiating the GPs from residents. Again, the noise introduced by the cueing effect may be responsible for this.

An equally interesting phenomenon was the sudden decrease in reliability, parallelled in both tests, in the high expertise groups. A ceiling effect as a possible explanation for this, could not be demonstrated. Cognitive psychological research, however, indicates an increasingly individually differentiated knowledge base, usually as a result of a accumulation of idiosyncrasies due to clinical exposure. The phenomenon of "content specificity", i.e. the variability of examinee performance across cases, is well documented in the medical assessment literature (Case & Swanson, 1993; Elstein, 1993; Van der Vleuten & D. Swanson, 1990; Van der Vleuten et al., 1994). It would be logical for content specificity to increase with level of expertise. The decrease of reliabilities could be a reflection of this. Unfortunately the design used here does not allow separation of examinee by case interaction from general error, in order to check this explanation.
Some practical implications may come from this study. As cueing was found to be a bidirectional effect, dependent not only on the level of expertise but also on the difficulty of the question, and, as literature shows, the content of the question (Swanson et al., 1987), one cannot simply assume that all multiple choice questions are easier than (parallel) open-ended questions. Yet, with regard to accuracy of measurement these data could indicate that OEQs are to be preferred over MCQs. MCQs provide to some extent a biassed estimate of the ability due to cueing effects. Yet, the net effect is not dramatic and the amount of common information, i.e. their intercorrelation, is considerable. This should be weighed against the increase in resource requirements when OEQs are used (correction time, testing time needed to cover domain). The decision about which question format to use may vary from situation to situation.

REFERENCES

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Case, S. M., & Swanson, D. B. (1993). Extended-macthing items: a practical alternative to free response questions. *Teaching and Learning in Medicine, 5*(2), 107 - 115.

Day, S., Norcini, J., Webster, G., Viner, E., & Chiroco, A. (1988). *The effect of changes in medical knowledge on examination performance at the time of recertification.* Paper presented at the 27 th RIME, Chicago.

Elstein, A. E. (1993). Beyond multiple choice questions and essays: The need for a new way to assess clinical competence. *Academic Medicine, 68*(4), 244-48.

Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education, 22*, 173 - 79.

Maatsch, J., & Huang, R. (1986). *An evaluation of the construct validity of four alternative theories of clinical competence.* Paper presented at the Proceedings of the 25th annual RIME conference, Chicago.

Maatsch, J. L., Huang, R., Downing, S. M., & Barker, D. (1983). *Predictive validity of medical specialty examinations* (HS 02038-04): Michigan State University.

Newble, D. I., Baxter, A., & Elsmlie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*, 263-268.

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teaching and Learning in Medicine, 8*(4), 208-216.

Norman, G. R., Smith, E. K. M., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21*, 297-304.

Page, G., Bordage, G., Harasym, P., Bowmer, I., & Swanson, D. B. (1990). A new approach to assessing clinical problem-solving skills by written examination: Conceptual basis and initial pilot test results. In W. Bender, R. J. Hiemstra, A. Scherpbier, & R. J. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence, Proceedings of the fourth Ottawa conference* (pp. 403 - 7). Groningen: Boekwerk Publications, Groningen, The Netherlands.

Schmidt, H. G. (1990). Innovative and conventional curricula compared: what can be said about their effects? In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in Medical Education: An Evaluation of Its Present Status* (pp. 1-7). New York: Springer Publishing Company.

Schmidt, H. G., & Boshuizen, H. P. (1993). On acquiring expertise in medicine. Special Issue: European educational psychology. *Educational Psychology Review, 5*(3), 205-221.

Stalenhoef-Halling, B. F., Van der Vleuten, C. P. M., Jaspers, T. A. M., & Fiolet, J. B. F. M. (1990). A new approach to assessing clinical problem-solving skills by written examination: Conceptual basis and initial pilot test results. In W. Bender, R. J. Hiemstra, A. Scherpbier, & R. J. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence, Proceedings of the fourth Ottawa Conference* (pp. 552 - 7). Groningen: Boekwerk Publications, Groningen, The Netherlands.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van der Vleuten, C. P. M., & D. Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58 - 76.

Van der Vleuten, C. P. M., Newble, D. I., Case, S. M., Holsgrove, G., McCann, B., McGrae, C., & Saunders, N. (1994). Methods of Assessment in Certification. In D. I. Newble, B. Jolly, & R. Wakeford (Eds.), *The certification and recertification of doctors, issues in the assessment of clinical competence* (pp. 105 - 25). Cambridge: Cambridge University Press.

Van Leeuwen, Y. D., Pollemans, M. C., Düsman, H., Van der Vleuten, C. P. M., & Grol, R. P. T. M. (1993). *De huisartsgeneeskundige kennistoets, construct validiteit en betrouwbaarheid: welke maten meten wat? [A knowledge-based test for general practitioners, construct validity and reliability, which measures measure what?].* Paper presented at the Tweede Gezond Onderwijs Congres [Second National Conference on Medical Education], Veldhoven.

Veloski, J., Rabinowitz, H., & Robeson, M. (1986). *Cueing in multiple choice questions: a reliable, valid and economical solution.* Paper presented at the 27 th annual RIME conference, Chicago.

Veloski, J., Rabinowitz, H., & Robeson, M. (1993). A solution to the cueing effects in multiple choice questions: the Un-Q-format. *Medical Education, 27*, 371-375.

CHAPTER 3

COMPUTERISED LONG-MENU QUESTIONS AS AN ALTERNATIVE TO OPEN-ENDED QUESTIONS IN COMPUTERISED ASSESSMENT

SUMMARY

To optimally avoid cueing effects and computer scoring problems in computerised examinations a computerised long-menu question (CLM) was developed. This question type was compared to open-ended questions in one treatment group and to multiple choice questions in another treatment group. Also, scores were compared to self-perceived anxiety of the participants. CLMs yield comparable scores to open-ended questions, but the scores differ significantly from those on multiple choice tests. Correlations in the first comparison (CLMs with open-ended questions) were higher than those in the second comparison (CLMs with multiple choice questions). The amount of positive and negative cueing was considerably higher in the first than in the second comparison. Response times of CLMs were higher than those of multiple choice questions and open-ended questions, differing significantly from both. Computer anxiety did not influence the mean scores in either comparison. Therefore, in computerised testing, CLMs seem to offer an acceptable replacement of open-ended questions.

INTRODUCTION

From their earliest uses multiple choice questions have been used extensively in all kinds of examinations (McCall, 1920). The reasons for this may be that they yield more reliable scores, and the low labour intensity in answering and scoring (Swanson, Norcini, & Grosso, 1987). Two major disadvantages have also been reported, the relatively high possibility of a correct answer by sheer guessing and the cueing effect.

To discourage sheer guessing, several methods have been described. The effects of these different methods to oppose guessing and the guessing itself have been the object of many studies (Harden, Brown, Biran, Dallas Ross, & Wakeford, 1976; Lord, 1963; Mattson, 1965; Naerssen, 1961; Ruch & Stoddard, 1925; Votaw, 1936; West, 1923). In summary, the results show that correction for guessing may often be necessary, but it has negative psychometric effects, probably through the introduction of more random error.

The cueing effect, an effect suggesting that examinees on seeing the correct answer will recognise this as such, was first suggested in the 1950s in the domain of knowledge of word meaning (Hurlburt, 1954). Twelve years later a first study was conducted to assess the influence of cueing in medical examinations (McCarthy, 1966). Especially in tests aimed at assessing higher order cognitive skills, like problem solving or clinical reasoning cueing was perceived to be disadvantageous, because recognition of a correct answer was not considered a higher order cognitive skill (Newble, Baxter, & Elsmlie, 1979).

Many studies have been performed in which different item formats were compared by using parallel tests (Hettiaratchi, 1978; McCarthy, 1966; Newble et al., 1979; Norman et al., 1987; Page, Bordage, Harasym, Bowmer, & Swanson, 1990; Schuwirth, Jean, Van der Vleuten, & Van Santen, 1993; Schuwirth, Van der Vleuten, & Donkers, 1992; Stalenhoef-Halling, Van der Vleuten, Jaspers, & Fiolet, 1990; Veloski, Rabinowitz, & Robeson, 1986). In almost all of these studies mean scores on the multiple choice questions were higher than those on the open-ended questions (Hettiaratchi, 1978; McCarthy, 1966; Newble et al., 1979; Norman et al., 1987; Page et al., 1990; Schuwirth et al., 1993; Schuwirth et al., 1992; Stalenhoef-Halling et al., 1990; Veloski et al., 1986), but the opposite has also been found (Schuwirth et al., 1993; Schuwirth et al., 1992; Stalenhoef-Halling et al., 1990).

Intertest correlations, however, proved to be high in most comparative studies, especially when corrected for attenuation (by which an estimate of the correlation is given when both tests would have had an ideal reliability) (Maatsch & Huang, 1986; Norman et al., 1987; Schuwirth et al., 1992; Stalenhoef-Halling et al., 1990). One study reports low to moderate correlations (Hettiaratchi, 1978), but in view of the length of the tests used (the author does not report reliabilities) it may be assumed that disattenuated correlations would be much higher.

In one of our previous studies the same test was administered to the examinees twice using a computer (Schuwirth et al., 1992). The design enabled cross tabulations per item of the possible answer combinations: both answers correct, both answers incorrect and two combinations of one correct and one incorrect answer. Apart from the (expected) finding that cueing in favour of the multiple choice question occurred in all items, an opposite effect was also found. In nearly all items a substantial percentage of the examinees answered the open-ended question correctly and the parallel multiple choice question incorrectly. We referred to the expected effect as 'positive cueing' and to the opposite as 'negative cueing'.

Generalisations from these studies as to the size of the cueing effect or to one of the question formats being intrinsically superior must be made with extreme caution. Correlations are quite high in most studies. Furthermore, it is impossible to disentangle cueing from sheer guessing with the methodology used. Norman et al. (Norman, Swanson, & Case, 1996) have described several setbacks of the methodology used. First, and most important, differences in mean scores do not indicate that a different trait is measured. Second, even when low correlations are found, they would probably be suppressed by the unreliability of the tests used. Finally, cueing appears to be an unpredictable effect that does not only act in two directions, but its size seems to vary with the level of expertise of the examinees (Newble et al., 1979), the item difficulty (Schuwirth et al., 1992) and the content of the question (Swanson et al., 1987).

The easiest way to avoid cueing would therefore be the sole use of open-ended questions, but their labour intensity in answering and scoring has encouraged the development of alternative question types that avoid cueing (Case & Swanson, 1993; Veloski et al., 1986). In one of these a so-called long-menu is used (Veloski et al., 1986). This are booklets with an alphabetically ordered long list of possible answers (over 500) each with its specific code. Examinees are then supposed to generate the correct answer, look it up in the booklet and to transfer the code to the answer sheet. Although it appeared to diminish the cueing effect substantially, the method was time consuming and mistakes could be made in the process of transferring the code from the booklet to the answer sheet.

Computer scoring of open-ended questions with algorithms or natural language understanding appears far too inefficient yet to be used in high stakes examinations (Sabah, 1993).

Therefore, we developed a computerised version of a long-menu question (CLM) in which a student may find an answer in the list by typing it into a dialogue box. The computer then searches a long (over 2500 alternatives) list for 'hits'. The alternatives found are reported back to the examinee immediately so he can check whether the retrieved option is the desired one.

In this present study the CLM was compared to normal (6-8 options) multiple choice (MCQ) in one condition and to normal open-ended questions (OEQ) in a second condition.

Furthermore all participants were given evaluation forms to indicate their familiarity with or anxiety for computers.

The following research questions were asked:

1 *Do CLMs resemble (psychometrically) more the OEQs or the MCQs?*
2 Does the amount of experience with handling computers influence proficiency scores and response times?

As parameters for the first question, mean scores, correlations, cross tabulations, and mean response times were used. As some studies indicate that the level of expertise may influence the amount of cueing occurring (Newble et al., 1979; Schuwirth et al., 1992), we also investigated whether possible effects are influenced by the level of expertise of the examinee.

The second question has mainly emerged from the statements of some students, indicating that they are relatively unfamiliar with the use of computers. Norcini et al. (Norcini, J. Keskausas, L. Langdon, & G. Webster, 1986), on the other hand, have indicated that the level of self-perceived 'computer illiteracy' is not reflected in proficiency-scores.

METHOD

Material

Thirty written cases were used in a computerised test. All cases were transcriptions of real patients seen in a general practice setting, using a so-called key-feature approach (Bordage, 1987). In this approach the case descriptions are kept short, only reporting relevant characteristics, signs and symptoms, and questions are directed towards essential decisions. Since literature indicates that questions prompting for diagnosis show the most prominent cueing effect (Swanson et al., 1987) each case was linked to one question asking for the most probable diagnosis. In order to assess their fidelity all cases were presented to three physicians first. The computer presentation uses a Windows environment in which students are presented the case first. They then have to generate the most probable diagnosis, click an "OK" button to activate the question and answer it. The time between this activation of the question and the confirmation of the answer is recorded by the computer.

Design

Two conditions were created. In one condition the examinees were presented the set of cases linked to CLMs first and then the same set of cases linked to MCQs. In the second condition the first set consisted of OEQs and the second of CLMs. Therefore, based on the question format and the condition four subsets can be distinguished. These will be addressed further by a three-letter abbreviation and a number. CLM1 would thus be the subset containing the CLMs in condition 1. Examinees were randomly assigned to either condition. Also all examinees received an evaluation form, presenting several items (in a five-point Likert format) about this kind of testing. Two items explicitly asked about computer-illiteracy or anxiety.

Subjects

Thirty medical student from each of the second, fourth and fifth year (of a 6-year problem-based medical curriculum) participated in this study. Although the second and fourth year are pre-clinical years, students are regularly presented paper cases to solve. An increase in their ability to solve cases could therefore be expected. Fifth year students all were in the last week of their clerkship in general practice (12 weeks). Some of these students, however, already had some clinical experience from previous clerkships, as the order of the rotations is not fixed. No academic credits could be gained by participation, but all examinees received a financial compensation for their efforts

All examinees were instructed about the impossibility to return to a previous case. For the second set in each condition they were asked to consider each case as a new one, i.e., to reread the text and to reconsider the answer.

Scoring

Answers to the CLMs and MCQs were scored and filed by the computer; answers to the OEQs were filed and hand scored afterwards. For this a previously fixed answer key was used. All response times were filed. Scores were expressed as percentages of correct scores; response times were expressed in minutes.

Analysis

Descriptive statistics were calculated for scores and response times for every subset, including an estimate of the reliability of the scores using Cronbach's alpha. Within both conditions a repeated measures analysis of variance was used to test the effect of question-format on mean scores and response times. Mean score differences between CLM1 and CLM2 were tested using an independent T-test. Within both methods correlations between the question-formats were calculated. Cross tabulations of all four possible answer combinations were made per item and were then summed over all 30 cases, to detect the magnitude of the positive and negative cueing effect. Positive cueing was defined here as cueing favouring the scores on the question-format resembling the MCQ most strongly (MCQ1 and CLM2). To test the effect of computer-illiteracy or anxiety on mean scores and mean response times, a factor was calculated using the sum of the answers given to the items on the evaluation form. In both conditions a repeated measures analysis of variance was performed testing the mean score against this factor ("computer anxiety"). This analysis was repeated for response times.

RESULTS

The scores and the response times are given in tables 3.1 and 3.2.

In all subsets an increase in mean scores is found with increasing level of training. Differences between the mean scores in condition 1 are larger than in condition 2. Analysis of variance shows a significant effect of question format in condition 1 ($F(1,44) = 52.21$, $p < 0.0001$).

In condition 2 this effect is not significant ($F(1,44) = 0.46$, $p = 0.5$). The effect of level of

Table 3.1: Mean scores and reliabilities of all four subsets.

	Comparison condition 1						Comparison condition 2					
	CLM 1			MCQ 1			OEQ 2			CLM 2		
Year of training	mean	S.D	alpha	mean	S.D.	alpha	mean	S.D.	alpha	mean	S.D	alpha
2	21.6	11.8	.68	31.8	10.9	.50	14.0	7.3	.41	16.0	6.3	.60
4	41.8	11.2	.56	52.7	10.7	.51	42.2	15.6	.77	39.6	14.5	.72
5	46.4	9.2	.28	57.6	9.1	.30	54.2	16.2	.77	50.9	11.5	.56
Total	36.6	15.2	.76	47.3	15.1	.74	36.2	21.2	.89	35.5	18.4	.8

training on mean scores is significant in all subsets ($p < 0.0001$).All interaction effects are not significant in both conditions. Some differences in mean scores are seen between the CLMs in condition 1 and the CLMs in condition 2, but they are all not significant. Reliability estimates

Table 3.2: Average response times needed to complete all items.

	Comparison condition 1				Comparison condition 2			
	CLM 1		MCQ 1		OEQ 2		CLM 2	
Year of training	mean	S.D	mean	S.D.	mean	S.D.	mean	S.D.
2	29.91	9.04	9.63	2.83	15.17	8.87	17.93	4.34
4	36.04	8.31	10.71	3.10	20.36	12.83	22.80	6.72
5	30.77	12.86	10.69	5.47	19.49	12.00	19.70	5.90
Total	31.41	10.76	10.39	3.96	18.93	11.68	20.14	5.96

within the year groups are moderate, but over the total of the examinees they are acceptable. In combination with the mean response times an estimate of the reliabilities per hour of testing time can be obtained using the Spearman-Brown prophecy formula. Reliabilities would then be 0.86 and 0.94 for the CLM1 and MCQ1, and 0.96 and 0.94 for the OEQ2 and the CLM2.The shortest response times are needed for the multiple choice. The difference in condition 2 appear to be smaller than in condition 1, but in both conditions the effect of question-format on response time is significant ($p < 0.0001$). The differences between the mean response times on the CLMs in condition 1 and 2 are remarkable. They are all statistically significant (independent T-test, $p < 0.0001$).
In table 3.3 the intertest correlations are shown.

Table 3.3: Correlations between the question formats.

Year of training	Comparison condition 1 CLM1 - MCQ 1	Comparison condition 2 CLM 2 - OEQ 2
2	.72	.74
4	.46	.88
5	.57	.95
Total	.78	.94

Correlations shown in this table are observed correlations. In condition 1 these are lower than in condition 2. The correlation in year group 4 in the first condition is considerably lower than all other correlations. This correlation estimate, however, is not statistically significant (p =

.232), indicating that the number of observations is too small in this sample to determine such low correlations. All other correlations can be considered significant ($p < 0.05$). The results of the cross tabulations of the possible answer combinations are shown in table 3.4.

Table 3.4: Percentages in which the four possible answer combinations occurred.

Cueing type	Comparison Condition	year 2	year 4	year 5	total
Positive cueing	1	17.1%	16.7%	15.8%	16.5%
	2	7.8%	6.7%	8.7%	7.7%
Negative cueing	1	6.9%	5.8%	4.7%	5.8%
	2	5.8%	9.3%	10.2%	8.4%
Total cueing	1	24.0%	22.5%	20.5%	22.3%
	2	13.6%	16.0%	18.9%	16.1%
Net cueing	1	10.2%	10.9%	11.1%	11.7%
	2	2.0%	-2.6%	-1.5%	-0.7%

The net cueing (being the positive cueing minus the negative cueing) indicates the size of score differences resulting from the cueing (which has in fact been assessed by the analysis of variance). The total amount of cueing, however, indicates the percentages in which both question formats did not agree about the competence of the examinee (based on the same case). In the comparison of CLMs with MCQs more positive cueing occurs than in the comparison CLMs and OEQs. Negative cueing, however, occurs slightly more often in the second comparison. A reasonable amount of net cueing remains in the first comparison, in the second comparison the net cueing deviates only marginally from 0%.
The effect of possible "computer anxiety" on the mean score is not significant in either condition ($p = 0.133$ in condition 1 and $p = 0.678$ in condition 2), nor is the effect on response time ($p = 0.991$ and 0.858 respectively). All interaction effects are not statistically significant.

DISCUSSION

The aim of the study described here was not to establish whether or not CLMs tap into a different kind of cognitive skills than OEQs or MCQs. The design of this study, the number of questions used and the number of examinees involved would probably not allow such conclusions.

Since the CLMs are meant to be an acceptable replacement of open-ended questions, it is more practical to assess whether scores obtained using CLMs are more comparable to OEQs than to MCQs.

An issue that strongly influences the comparability of CLMs to either of the other question formats is the kind of standard setting used. The use of an absolute standard requires the scores on the question formats to be identical in absolute terms, whereas a relative standard setting method would only require the question formats to rank order students similarly. Both parameters have been studied here.

It appears that an effect of question format on mean scores could only be detected when comparing the CLMs with MCQs. In view of the fact that no interaction effects of level of training could be detected it seems that the mean effect is fairly stable.

The mean scores on the CLMs in condition 2 are somewhat lower than those on the CLMs in condition 1. Since this difference is not statistically significant to any acceptable level, the most appropriate explanation is that it is an artefact. Still it would not have been surprising to have found a difference favouring the CLM2 since they were presented as the second set in a condition. Mobilisation of prior knowledge could have been expected to result in a higher mean score on the CLM2. But an effect on proficiency seems to occur only when time allocation to the cases is restricted (Machiels-Bongaerts, Schmidt, & Boshuizen, 1993), which has not been the case here. Nevertheless some inequality in the mean level of competence could be present, although 45 examinees were fully randomly assigned to both conditions. This has to be borne in mind in all of the comparisons between conditions described.

Answering CLMs appears to be more time-consuming than answering MCQs and even than OEQs. Reasons for this could be unfamiliarity with this question-type, or incompleteness of the list used. Although over 2500 options have been included in the list it is certainly possible that some of the alternatives considered by the candidates have not been included yet. This can of course be regarded as a (temporary) setback of this method. A clear difference exists between the mean response times for the CLMs in both conditions.

The most probable explanation lies in the fact that in condition 1 the CLMs were the first subset and in condition 2 the second. So in the second condition students already had to formulate their answer, thus allowing them to find the answer in the CLM2 sooner than in CLM1. Another effect which could have attributed to this difference may be the mobilisation of prior (related) knowledge (Machiels-Bongaerts et al., 1993).

This mobilisation appears to shorten the response times needed. This then could account for a decrease in response times on CLM2 even when the answer on the OEQ2 had been incorrect. The difference in response times did not influence the reliability estimate per hour of testing time. In these reliabilities only minor differences were found.

The observed correlations reported are suppressed by the unreliability of the subsets. Norman et al. (Norman et al., 1996) suggest therefore that a disattenuation of the observed correlations should be performed. This procedure, however, corrects also for content differences of both subsets. Since all subsets had an identical content, this would have resulted in an overestimation of the true correlation. It may, however, be expected that some of the difference in correlation may be attributable to the difference in reliabilities, since these are slightly lower in condition 1.

It appears that the net cueing effect of CLMs can be neglected, but even in condition 2 still in 13 to 19% disagreement of both question formats about the ability of the examinee (total cueing) exists. Negative cueing seems to appear more in condition 2 than in condition 1. A

possible explanation for this could be that the number of synonyms and alternatives in the list used for the CLM is still too low. An examinee would then be able to answer the OEQ correctly, but not to find the answer in the CLM. In view of the already large numbers of alternatives, this may be expected to have accounted only for a small portion. Extension of this list would eventually lead to an elimination of this problem. In all year groups MCQs cue more than CLMs. In MCQs the total amount of cueing averages about 20%, which is congruent with one of our earlier studies (Schuwirth et al., 1992).
The level of perceived computer anxiety seems to influence neither the mean scores nor the mean response times. The former is congruent with the findings of Norcini et al. (Norcini et al., 1986) The latter would indicate that either the self-perception of computer anxiety of the students is inadequate or that it is adequate but still does not interfere with performance. Our favourite explanation would be that the interface is so user-friendly that everybody can work with it. But whether this explanation is correct should certainly be studied further. In summary, it seems that CLMs resemble an OEQ more than an MCQ and could therefore be an acceptable replacement for the OEQ when using a computer. Since, however, the disadvantage of response times needed is also present in CLMs, it is still advisable to let the content of the question decide what the most appropriate format for the question must be (McGuire, 1987; McGuire, 1994).

REFERENCES

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Case, S. M., & Swanson, D. B. (1993). Extended-macthing items: a practical alternative to free response questions. *Teaching and Learning in Medicine, 5*(2), 107 - 115.

Harden, R. M., Brown, R. A., Biran, L. A., Dallas Ross, W. P., & Wakeford, R. E. (1976). Multiple choice questions: to guess or not to guess. *Medical Education, 10*, 27-32.

Hettiaratchi, E. (1978). A comparison of student performance in two parallel physiology tests in multiple choice and short answer forms. *Medical Education, 12*, 290-296.

Hurlburt, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meaning, nouns, verbs, adjectives. *Journal of Educational Research, 47*(8), 561- 76.

Lord, F. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 13*(4), 663 - 72.

Maatsch, J., & Huang, R. (1986). *An evaluation of the construct validity of four alternative theories of clinical competence.* Paper presented at the Proceedings of the 25th annual RIME conference, Chicago.

Machiels-Bongaerts, M., Schmidt, H., & Boshuizen, H. (1993). Effects of mobilizing prior knowledge on information processing: Studies of free recall and allocation of study time. *British Journal of Psychology, 84*, 481-498.

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. *Educational and Psychological Measurement, 25*(3).

McCall, W. (1920). A new kind of school examination. *Journal of Educational Research, 1*(1).

McCarthy, W. H. (1966). An assessment of the influence of cueing items in objective examinations. *Journal of Medical Education, 41*, 263 - 66.

McGuire, C. (1987). Written methods for assessing clinical competence. In I. R. Hart & R. M. Harden (Eds.), *Further Developments in Assessing Clinical Competence* (pp. 46-58). Montreal: Can-heal Publications.

McGuire, C. (1994). Letter to the editor. *Teaching and Learning in Medicine, 6*(2), 74.

Naerssen, R. v. (1961). A scale for the measurement of subjective probability. *Acta Psychologica, 20*, 159-166.

Newble, D. I., Baxter, A., & Elsmlie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*, 263-268.

Norcini, J. J., J. Keskausas, J., L. Langdon, L., & G. Webster, G. (1986). An evaluation of a computer simulation in the assessment of clinical competence. *Evaluation and the Health Professions, 9*(3), 286-304.

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teaching and Learning in Medicine, 8*(4), 208-216.

Norman, G. R., Smith, E. K. M., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21*, 297-304.

Page, G., Bordage, G., Harasym, P., Bowmer, I., & Swanson, D. B. (1990). A new approach to assessing clinical problem-solving skills by written examination: Conceptual basis and initial pilot test results. In W. Bender, R. J. Hiemstra, A. Scherpbier, & R. J. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence, Proceedings of the fourth Ottawa conference* (pp. 403 - 7). Groningen: Boekwerk Publications.

Ruch, G., & Stoddard, G. (1925). Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology, 16*, 89-103.

Sabah, G. (1993). Knowledge respresentation and natural language understanding. *AICOM, 6*(3/4), 155 - 86.

Schuwirth, L. W. T., Jean, P., Van der Vleuten, C. P. M., & Van Santen, M. (1993). Problem-Analysis Questions, een korte casusvorm voor het preklinische domein [Problem Analysis Questions, a short case-based testformat for the pre-clinical domain]. In E. Houtkoop, J. Pols, M. C. Pollemans, A. J. J. A. Scherpbier, & G. M. Verwijnen (Eds.), *Proceedings van het Derde Gezond Onderwijs Congres* (pp. 104 - 11). 's Gravenhage, Nederland: Haagse Hogeschool.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1992). Open-Ended Questions versus Multiple Choice Questions. In R. Harden, I. R. Hart, & H. Mulholland (Eds.), *Approaches tot the Assessment of Clinical Competence, proceedings of the fifth Ottawa conference* (Vol. 2, pp. 486 -91). Norwich, Great Britain: Page Brothers.

Stalenhoef-Halling, B. F., Van der Vleuten, C. P. M., Jaspers, T. A. M., & Fiolet, J. B. F. M. (1990). A new approach to assessing clinical problem-solving skills by written examination: Conceptual basis and initial pilot test results. In W. Bender, R. J. Hiemstra, A. Scherpbier, & R. J. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence, Proceedings of the fourth Ottawa Conference* (pp. 552 - 7). Groningen: Boekwerk Publications, Groningen, The Netherlands.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Veloski, J., Rabinowitz, H., & Robeson, M. (1986). *Cueing in multiple choice questions: a reliable, valid and economical solution.* Paper presented at the 27 th annual RIME conference, Chicago.

Votaw, D. F. (1936). The effect of do-not-guess directions upon the validity of true-false or multiple choice tests. *Journal of Educational Psychology, 27*, 698-703.

West, P. V. (1923). A critical study of the right minus wrong method. *Journal of Educational Research, 8*(1), 1-9.

CHAPTER 4

DIRECT AND INDIRECT VALIDITY OF A TEST FOR COMPUTERISED CASE-BASED TESTING

SUMMARY

Current insight into the nature of problem solving have led to the development of so-called authentic assessment instruments. These are based on short cases with limited numbers of questions asking for essential decisions only. Computerised Case-based Testing (CCT) is one of these approaches. Evidence for the validity of these instruments is still scarce. An optimal validation procedure would consist of elements of both indirect and direct validity. The aim of this study is to provide evidence for the validity of CCT using viewpoints on validity. Direct validity was assessed using students' and experts' opinion on the tasks set by CCT. In addition the adequacy of the blueprint and the question formats used were assessed using multivariate generalisability theory. Indirect validity was studied by comparing pre-test and post-test scores of students and test scores of experts. In addition, test scores were compared to scores on different assessment instruments. Judgements of experts and students suggest that CCT measures problem-solving skills. The blueprinting and the question formats did not have a negative effect on the generalisability of CCT. A significant difference in mean pre-test and post-test scores was found. Student post-test scores were significantly lower than expert scores. Correlations with other tests ranged from low to moderate. In general, it can be concluded that evidence for the validity of CCT is present, although the correlation study is not conclusive.

INTRODUCTION

The measurement of problem-solving skills has posted quite a record in the research in medical education. A long history of proposed and abandoned strategies can be found in the literature (Barrows & Tamblyn, 1977; Helfer & Slater, 1971; Knox, 1980; Swanson, Norcini, & Grosso, 1987). The mainstream of these developments consisted of the development of high fidelity simulations in which patient cases were used as stimulus material. Mostly complete and integral patient simulations were used of which some followed a realistic branched scenario. Two assumptions underlied these approaches. First, problem-solving ability was considered to be more or less a generic skill, i.e. the score on one case would predict the score on any other case. Second, experts could be distinguished from novices by their possession of more elaborate algorithms or skills to solve a problem (Barrows, 1984).

The outcome of further psychometric research on these integral case examinations, however, revealed evidence contradicting these two assumptions. First, it was found that inter-case correlations were actually extremely low, which implied that the score on one case is in fact a very poor predictor for the score on any other case. As a consequence, tests containing only a few cases would lead to a severe sampling bias and thus to unreliable test scores. To overcome this problem it would be necessary to use many cases leading to an infeasibly long testing time (Swanson, 1987; Swanson et al., 1987).

Second, it was discovered that experienced experts were outperformed by novices. This so-called intermediate effect cast doubts on the construct validity of these tests (Grant & Marsden, 1988; Schmidt, Boshuizen, & Hobus, 1988). In addition, repeated findings of high correlations between most of these long case tests and plain (multiple choice-based) knowledge tests gave rise to further doubts with respect to the construct validity of the method (Maatsch & Huang, 1986; Norcini, Swanson, Grosso, & Webster, 1985).

These disappointing findings introduced a shift in thinking about the nature of problem solving and development of professional expertise herein (Schmidt, Norman, & Boshuizen, 1990). In modern theories, problem-solving ability is considered to be dependent on the specific problem and its clinical context, and not at all generic as was assumed originally. Furthermore, it is currently theorised that solving a particular problem is not so much a matter of applying a fixed set of algorithms on a specific problem, but that problem-solving strategies are highly individualised and idiosyncratic, depending on individual experiences and clinical exposure.

This in turn led to a change in thinking about how tests should be designed to measure problem-solving ability. One of the suggestions consisted of the so-called 'key-feature' approach (Bordage, 1987; Page & Bordage, 1995; Page, Bordage, & Allen, 1995). In this approach a test consists of cases in which all irrelevant information is removed and only relevant clinical and contextual information is reported. Only a limited number of questions are asked per case, prompting only for decisions that are essential and specific to that particular case. This way the focus is more on the solution of essential problems than on the process (algorithms) leading to this solution. Through their brevity the administration of large numbers of cases per test is possible, thereby improving the content coverage and the reproducibility of the test scores.

Using this approach a computerised case-based testing (CCT) procedure has been proposed (Schuwirth, Van der Vleuten, De Kock, Peperkamp, & Donkers, 1996a). CCT is used in our

context to test the problem-solving ability of the students during their clerkships. It uses short cases that are based on real patients and they are written by the doctor who has actually done the consultation with the patient. Questions in the cases are aimed at various decision moments. In addition, different question formats (multiple choice, true-false and short answer open-ended questions) are used, based on the number of realistic alternatives that would exist in real practice. Therefore the content of the question determines the format. This is done to increase the authenticity of the test. Reasons for doing so lie in the literature on cueing that has repeatedly shown no significant format effect (Norman, Swanson, & Case, 1996; Norman et al., 1987; Schuwirth, Vleuten, & Donkers, 1996b; Ward, 1982). The cases for a test, which is generated for each individual, are selected according to a predefined blueprint. A more detailed description of CCT has been published elsewhere (Schuwirth et al., 1996a). Evidence for the validity of the 'key-feature' approach in general or CCT in particular is scarce (Bordage, Brailovsky, Carretier, & Page, 1995; Page & Bordage, 1995). Therefore, this chaptere addresses the validity of the CCT as an attempt to validate the key-feature approach.

The choice for an optimal validation procedure, however, is not an easy one. Controversy has long existed in the literature about the bets approach to validate new examination methods (Cronbach, 1983; Ebel, 1983). The most widely used approach in the literature views an examination implicitly as a psychological test. This implies that psychological test validation approaches should be applied Cronbach, 1983 #175]. Central in this approach is to postulate a construct with its characteristics. In the validation process test scores are used to assess whether and to what extent these scores support the postulated characteristics of the construct. Usually this leads to correlational validation approaches or to predictions of difference in performance between reference groups.

A totally different point of view has been advocated by Ebel (Ebel, 1983). He argues that, contrary to psychological tests, educational tests have an intrinsic direct meaning, directly determined by their contents. Validation should therefore depend more on direct rational analyses of the content of an examination (e.g. by careful blueprinting) or by direct analyses of the way the measurement is operationalised.

Other authors advocate a more intermediate standpoint by suggesting both direct (Ebel) and indirect (Cronbach) validations into a more comprehensive validation procedure (Flanagan, 1983; Gardner, 1983).

Regardless of whether direct or indirect validation should be more important,

including both viewpoints into a validation procedure for CCT appears to offer the broadest view on its validity. Therefore, this study assesses the validity of CCT in such way.

In the process of direct validation of CCT three aspects were considered essential: authenticity of the case material (stimulus), authenticity of the question formats (response formats) and blueprinting (content sampling). This led to the first three research questions:

1 Do students and experts judge the cases and the content of the questions to be sufficiently authentic or realistic? In the production of CCT material considerable care was taken to the selection of the essential problems and to the authenticity of the case description. Therefore, an overall positive judgment of both groups was expected.

2 Do different response formats contribute differentially to the capability of CCT to discriminate between ability? In CCT different question formats are used which are summed to a single overall test score. The question is whether these different formats contribute differently to the true score (or universe score) variance. If this is the case, it might be wise to use more items of a given format and to avoid others. However, given the

high correlations usually found between scores derived from different test formats, this was not to be expected.

3 Do different content areas contribute differentially to the capability of CCT to discriminate between ability? This is in fact a similar question to the previous one, but now focussed on (clinical) content. The cases for CCT are selected through a blueprint covering different content areas. It might be that some content areas are more heterogeneous of homogeneous than others, therefore differentially contributing to the true score variance. If this were the case, a strategy could be to weigh the content areas differently in the summation to a total score. It was, however, primarily expected that the different categories would equally contribute to the true score variance.

Since an increase in scores with increasing expertise and low associations with tests measuring other constructs may be expected, the following 3 research questions for indirect validation were formulated.

4 Linked to a (practical) clerkship, can a significant difference be found between the scores on an entrance level test and an exit level test? What is the shift in ratio of incompetent to competent students as identified by the test? It was expected that scores on the exit level test would be significantly higher than those on the entrance level test, and that a large shift from incompetent to competent students would occur.

5 Do expert scores differ from student scores on the exit level test? It was expected that experts would outperform students.

6 Which pattern of correlations between CCT and different examinations can be found? In this comparison CCT is compared to different tests. Some of these aim to focus exclusively on the testing of (recall of) factual knowledge, others focus more on the application of knowledge or the testing of clinical knowledge. Although often the results of studies using such a methodology lack evidence for divergent validity, it was expected that in the case of CCT this divergent validity would be found. It was therefore expected that overall correlations between CCT and the other tests would be low, but that a trend would occur in which correlations between CCT and those other tests that examine (the application of) clinical knowledge would be higher.

METHOD

Instruments

- CCT: two tests (pre-test and post-test) are used in this study. Both consisted of 40 cases each. Both tests are constructed according to a predefined blueprint, based on ICPC categories. Eight cases were the same in the pre- and post-test and were used as anchor items. All case material incorporated in the test had passed two review sessions, the first purported to optimise wording of cases and questions, and the second to judge correctness of answer keys, pass-fail level and reality value.
  The CCT tests were part of the formal assessment of the 10-week clerkship of general practice (GP) and lasted each for about 1.5 hours.
- A questionnaire was used to assess students' opinion on CCT. Two of the questions were aimed at the impression of students on the competence being measured by CCT namely:

(1) Does CCT distinguish itself sufficiently from examinations that are aimed at measuring factual knowledge? (2) Does the test adequately resemble the work you have been doing during your clerkship? The answers to these open-ended questions were then converted to three categories: positive, neutral, and negative.

- Clinical performance ratings: the GP of the practice in which the students did their clerkship (GP supervisor) gave four judgements about student performance in practice: knowledge (GP_K), skills (GP_S), application of knowledge (GP_{AK}) and attitudes (GP_A). The GP supervisors qualified performance as "insufficient", "doubtful", "sufficient", "good" or "excellent", judgements which were converted into marks on a five point scale ('insufficient' = 1; 'excellent' = 5).
- Tutor ratings: in addition to spending time in practice, students spent one day per week at the university in tutorial groups, led by a GP tutor. In these tutorial groups students had to perform three assignments: case presentations, thematic presentations and a family report about a family of which one of the members has a chronic illness. The GP tutor gave a mark out of 10 for each of the three assignments (T_C, T_T and T_R).
- Progress test: during the clerkship students had to sit a progress test. This is a written test, testing overall factual medical knowledge which consists of 250 true-false-don't know items. These items are sampled form the whole domain of medicine (Verwijnen et al., 1989). Within the test three subtests can be identified: "Basic sciences", "Clinical sciences", "Social sciences". In the analysis for the present study the subtests are used separately. They will be indicated with PT_B (basic sciences), PT_C (clinical sciences) and PT_S (social sciences).

Experts

In the second review cycle a panel of seven experts (third year GP trainees) were used to assess the pass-fail level and reality value of the cases. To determine a pass-fail level an Angoff procedure was used, without a consensus meeting (Angoff, 1971).

To obtain the judgements for the reality value of the cases and questions, the seven experts were asked to give a score based on two impressions: whether they thought it reasonable to be visited by a patient like the one described, and whether they considered the case description realistic and adequate. For this a five-point Likert scale (good, sufficient, doubtful, insufficient and poor) was used. They were instructed to be critical and if in doubt, to give a negative judgement. Three increasingly strict criteria for a valid case were defined: criterion 1 considered a case valid if the majority of the judges gave a rating of 'sufficient' or 'good'. Criterion 2 considered a case valid if no more than two experts considered the case 'doubtful', 'insufficient' or 'bad'. Criterion 3 considered a case valid if no expert rated the case 'doubtful', 'insufficient' or 'bad'.

Participants

All students who completed the clerkship of general practice in the academic year 1995/1996 (n=128) filled in the questionnaire about CCT. These results were used for research question 1. For a group of 63 students the clinical ratings, the scores on the CCT pre-test, post-test and the progress test could be obtained. In addition, the post-test was taken by 8 experienced GPs (GP tutors). All scores were used to answer research questions 2 to 6.

Analyses

The answers of the students on the questionnaire were categorised into three categories: "agree", "neutral" and "disagree". The numbers of answers in each of the categories were converted to percentages. These results and the relevancy rating of the experts were used to answer research question 1.

To answer research questions 2 and 3 multivariate generalisability theory was used (Brennan, 1983). In multivariate generalisability theory multiple true (or universe) and error scores (variances) can be estimated for different components of a test, i.e. format and content areas here, and these can be combined into a single composite reliability estimate. Technically, this involves a crossed person-by-item (p x i) design nested within a fixed facet (i.e., format and content area). This will reveal the relative contributions of each component, which can be used to optimise the overall composite reliability by varying the sample size within the component or the weight of the component. The composite universe score variance can be broken down into relative proportions of subtest (or component) variance in relation to the composite universe score variance. By comparing the relative universe score contribution with the actual number of items, it is possible to infer an optimal number of items. When the actual number is close to the optimal number suggested, one can conclude that there is no difference in relative contribution.

To answer research question 4, two analyses were performed. First, mean scores on the pre-test were compared to those on the post-test scores using a paired samples t-test. Then the tests were equated using an anchor item method (MacCann, 1990). The expected pre-test scores were compared to the scores on the post-test using an independent samples t-test. In the second analysis the percentages of passing or failing students on the pre-test and post-test are calculated and compared. The pass-fail scores were 55% for the pre-test and 51% for the post-test. The stability of the pass-fail score set by the experts is assessed using generalisability theory. A root mean square error (RMSE): σ_{Δ} of 3.64% (cf. Brennan, 198, p. 19) was found, indicating a fairly reliable judgement of the passing score.

Research question 5 was studied by comparing scores of the GPs to those of the students. For both groups the post-test was used. An independent samples t-test was used as a test for significance.

To answer research question 6 Pearson product moment correlations were calculated between the CCT post-test and the other tests. Since these can be spuriously low due to unreliability of the measurements, true correlations were calculated. These are estimations of the correlation if both measures had a perfect reliability. When reliability estimates were not available, unilateral or no corrections were made.

RESULTS

To answer research question 1 the expert and student judgements of the cases are reported below.Table 4.1 reports the numbers of valid cases per criterion according to the experts.

Table 4.1: Numbers and percentages of valid cases per criterion

Criterion	numbers of valid cases	% of valid cases	numbers of invalid cases	% of invalid cases
1	35	87.5	5	12.5
2	32	80.0	8	20.0
3	19	47.5	21	52.5

The number of valid cases depends on the criterion set. The difference between the first two criteria is rather small, but the number of cases on which all experts agreed that the case is sufficiently realistic is only about 50%.In table 4.2 the percentages are given of the responses of the students to the questions of the questionnaire.

Table 4.2: Percentages of student responses to the questions of the inventory.

	CCT Distinct from factual knowledge	CCT resembles problem solving during clerkship
Agree	60.94	88.28
Neutral	17.19	7.81
Disagree	21.88	3.91

For both questions the percentage of students agreeing is considerably larger than those disagreeing. Since the students were allowed to comment on their answer it was possible to establish the nature of the criticism. The most abundant comment made by those students disagreeing was that in order to solve the cases correctly, knowledge was indispensable, and therefore still a prerequisite for passing the examination. A large proportion of students, however, had the impression that the cases and the problem-solving strategies used to solve them closely resembled the patient cases they had encountered and the problem-solving strategies they had to apply during their clerkship.
The results of the generalisability analyses were used to answer research questions 2 and 3. A slight difference in generalisability coefficients was found between the situations with actual numbers of items per item format and optimal numbers of items per format (.67 and .62). The difference for items within blueprint categories is negligible (.71 and .73)

The results pertaining to research questions 4 and 5 are reported in table 4.3, giving mean percentage scores and standard deviations. The expected score is that on the pre-test after

Table 4.3: Descriptives of test scores, numbers of students passing and GP tutors' results on pre-test and post-test.

	Raw score		Equated score		GP Tutors		Number of students	Number of students
Test	Mean	Sd	Mean	Sd	Mean	Sd	failing	passing
pre-test	47.20	7.03	41.26	7.04			51 (80.92%)	12 (19.05%)
post-test	68.61	8.28			71.88	2.64	2 (3.17%)	61 (96.83%)

equating, which would have been the result of the pre-test if both tests had been equally difficult. The difference in scores between pre- and post-test is significant (paired samples t-test, $p < .0001$). The equated pre-test score is even lower indicating that the pre-test was in fact generally easier than the post-test (independent samples t-test, $p < .0001$), and that differences in difficulty of both test actually decreased the effect size. The difference between post-test and expected pre-test must therefore be significant (paired samples t-test, $p < .0001$)
The difference between the scores of the students on the total post-test and those of the tutors is also significant (independent samples t-test, $p < .0001$).

A considerable shift in numbers of students passing and failing occurs between pre and post-test. No students passed the pre-test and subsequently failed the post-test. All GP tutors passed the test when the student cut-off score was applied.

The correlation matrix between all tests pertaining to research question 6 is shown in table 4.4.
Correlations between CCT and all other tests are low, both observed and true correlations. The only significant correlations exist between different subtests within one measurement: all GP judgements correlate significantly with each other, as do the tutor judgements and the different clusters on the progress test. Between measures the numbers of significant correlations is low. CCT correlates only significantly but moderately ($R_{xy} = .40$) with the GP judgement about the skills of the student. Further significant correlations are between the family report and the GP judgement of application of knowledge, and between the cluster clinical sciences of the progress test and the GP judgement about knowledge and skills.

Table 4.4: Observed correlations (bottom diagonal) and true correlations (upper diagonal) between the scores on different tests.

	CCT	GP_K	GP_S	GP_{AK}	GP_A	T_R	T_C	T_T	PT_B	PT_C	PT_S
CCT	α=.67	.3349	.4012	.2289	.0473	.2537	.2791	.2394	.1551	.3215	.0476
	¢	¢	¢	¢	¢	¢	¢	¢			
GP_K	.2244	α=?	†	†	†	†	†	†	.1983	.3600	.1146
	.077								¢	¢	¢
GP_S	.2688	**.7625**	α=?	†	†	†	†	†	.0830	.3630	.2399
	.033	**.000**							¢	¢	¢
GP_{AK}	.1534	**.7180**	**.6758**	α=?	†	†	†	†	.2962	.3232	.2678
	.230	**.000**	**.000**						¢	¢	¢
GP_A	.0317	**.3613**	**.2807**	**.4739**	α=?	†	†	†	.2515	.1531	.1550
	.805	**.004**	**.026**	**.000**					¢	¢	¢
T_R	.1700	.1546	.1716	.2883	.1423	α=?	†	†	.1158	.1334	.1211
	.183	.226	.179	.022	.266				¢	¢	¢
T_C	.1870	.1166	.1459	.2418	.0960	**.9789**	α=?	†	.1123	.1035	.1070
	.142	.363	.254	.056	.454	**.000**			¢	¢	¢
T_T	.1604	.0713	.1170	.2002	.0801	**.9764**	**.9805**	α=?	.0996	.1045	.1205
	.209	.579	.361	.116	.533	**.000**	**.000**		¢	¢	¢
PT_B	.1085	.1382	.0606	.2162	.1836	.0845	.0820	.0727	**α=.73**	.7739	.4015
	.397	.280	.637	.089	.150	.510	.523	.571			
PT_C	.2264	.2664	.2686	.2392	.1133	.0987	.0766	.0773	**.5688**	**α= .74**	.4651
	.074	.035	.033	.059	.377	.442	.551	.547	**.000**		
PT_S	.0335	.0848	.1775	.1982	.1147	.0896	.0792	.0892	**.2951**	**.3442**	**α=.74**
	.795	.509	.164	.119	.371	.485	.537	.487	**.019**	**.006**	

†:True correlations not calculated, reliability could not be computed. ¢:True correlations are based on one reliability ?: alpha could not be computed.

DISCUSSION

Some support for the validity of CCT as a measure for problem-solving ability emerges from this study, although the support from all substudies is not equally strong.
A part of the experts used to judge the direct validity of the cases expressed doubts. Complete consensus that the cases were fit to test problem-solving ability could only be reached for about 50% of the cases. Even when using the least stringent criterion (the majority of the experts judging in favour), still 12.5% (k=5) of the cases could not be considered valid. This is a somewhat disappointing finding, considering the amount of time and resources that were devoted to the production of the cases (about 2 hours for writing and review of experienced item writers). Of course, the exact assignment to the experts and in particular the instruction to judge negatively in case of doubt may have had an negative impact on the figures. Nevertheless, in a repetition of such a procedure more detailed information as to why a case is considered invalid should be incorporated. This could probably provide some concrete suggestions for improvement.
In noticeable contrast to the expert judgements, student judgements were favourable. The vast majority of the students agreed that the problems and the decisions in CCT did resemble closely the problem solving during the clerkship. They also agreed that CCT can be distinguished from other more factual knowledge-oriented examinations. Often students are suspicious when confronted with a new and strictly formal examination. The fact that their overall judgement was so positive is therefore quite promising. It must be kept in mind, however, that the value of using introspection as a means to assess thought processes is limited (Nisbett & Wilson, 1977). The numbers of students, however, agreeing with the hypothesis seem large enough to conclude that at least another and more 'problem-solving-like' aspect is addressed by CCT.
The use of different question formats and the blueprinting does not have a negative impact on the generalisability. Especially for the different item formats this is quite gratifying. From a direct validity standpoint it is most sensible to use the number of alternatives in a question (either infinite in an open-ended question or finite in a multiple choice question) that most closely resembles the number of realistic options in real practice. This results in varying numbers of items per format. Simply adding the scores of all the items without adjustment or weighting for the different numbers could be perceived as too simplistic an approach. The results, however, indicate that this is not the case. The homogeneity of the scores within each question format is about equal. Therefore, no empirical objections to this approach exist. A similar conclusion can be drawn with respect to the blueprinting. The distribution of items within categories is fully adequate; no reasons exist to modify the sampling of the cases from the blueprint.
The increase in mean scores between pre-test and post-test is substantial and cannot be accounted for by differences in mean difficulty of the tests used. This supports the assumption that CCT measures a competence which is acquired during the clerkship of general practice. In addition the shift from large percentages of students failing the pre-test to large percentages of students passing the post-test can be seen as considerable support for the validity of CCT. Apparently after the clerkships students have developed sufficient competence in general practice for a clerk.

The fact that students at the end of their clerkship are outperformed by their tutors contributes further to the validity of CCT. In studies describing the intermediate effect often the reverse occurred (Goran, Williamson, & Gonella, 1973; Grant & Marsden, 1988; Marshall, 1977; Newble, Hoare, & Baxter, 1982; Schmidt et al., 1988). Still the conclusion that no intermediate effect exists should be made with some caution, since GP trainees for example were not used in this study. Therefore, to enable a firmer conclusion with respect to the absence of an intermediate effect further research is necessary.

The results of the correlational comparison are not very conclusive; all correlations between measures were low and the expected stronger correlations were not found. Two reasons for this can be mentioned. First, within examination procedures (GP judgements, Tutor judgement, Progress Test) a pattern of high intercorrelations between the subtests occur. If any specific pattern existed, its detection would probably be impaired by the fact that different subtests within one test are not independent of each other. Specifically within the GP judgements and the GP tutor ratings high correlations exist, probably due to a halo effect. Second, in many studies comparing test procedures correlational patterns are found that are hard to interpret (Swanson & Norcini, 1989). There is some doubt then whether this should be seen as indicative for the lack of validity of the instruments or rather as a weakness of using correlations as a method of validity assessment.

With caution, one conclusion can be drawn from the correlation matrix. Overall correlations are low. This indicates that the information obtained with CCT is unique with respect to the information from the other measures.

In general it can be concluded that the combination of the six substudies provides evidence for the validity of CCT as a measure of problem-solving ability. The direct validity studies are more conclusive than the indirect. Often, however, direct validity is considered 'softer' and indirect 'harder'. This renders the overall result of our study not decisive enough: conclusive but 'soft' evidence plus less conclusive but 'hard' evidence. Therefore, to complete the assessment of the validity of CCT as a measure of problem-solving ability additional information is needed. A third approach to validity, in the form of a cognitive psychological study in which more direct insight is gained in the nature of the thought processes involved in solving CCT cases, would be a practical area of research to find more conclusive evidence.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scales. In R. L. Thorndike (Ed.), *Educational Measurement* . Washington: American Council on Education.

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Barrows, H. S., & Tamblyn, R. M. (1977). The Portable Patient Problem Pack: a problem-based learning unit. *Journal of Medical Education, 52*, 1002 - 4.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Bordage, G., Brailovsky, C., Carretier, H., & Page, G. (1995). Content validation of key features on a national examination of clinical decision-making skills. *Academic Medicine, 70*(4), 276-81.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa: ACT Publications.

Cronbach, L. J. (1983). What price simplicity? *Educational Measurement: Issues and Practice, 2*(2), 11-12.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7 - 10.

Flanagan, J. C. (1983). A rational rationale. *Educational Measurement: Issues and Practice, 2*(2), 12.

Gardner, E. F. (1983). Intrinsic rational validity: Necessary but not sufficient. *Educational Measurement: Issues and Practice, 2*(2), 13.

Goran, M. J., Williamson, J. W., & Gonella, J. S. (1973). The validity of Patient Management Problems. *Journal of Medical Education, 48*, 171 - 7.

Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education, 22*, 173 - 79.

Helfer, R. E., & Slater, C. H. (1971). Measuring the process of solving clinical diagnostic problems. *British Journal of Medical Education, 5*, 48-52.

Knox, J. (1980). How to use Modified Essay Questions. *Medical Teacher,, 2*(1), 20-24.

Maatsch, J., & Huang, R. (1986). *An evaluation of the construct validity of four alternative theories of clinical competence.* Paper presented at the Proceedings of the 25th annual RIME conference, Chicago.

MacCann, R. G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. *Journal of Educational Statistics, 15*, 146 - 70.

Marshall, J. (1977). Assessment of problem-solving ability. *Medical Education, 11*, 329 - 34.

Newble, D., Hoare, J., & Baxter, A. (1982). Patient Management Problems, issues of validity. *Medical Education, 16*, 137 - 42.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*(3), 231 - 59.

Norcini, J. J., Swanson, D. B., Grosso, L. J., & Webster, G. D. (1985). Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education, 19*(3), 238-47.

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teaching and Learning in Medicine, 8*(4), 208-216.

Norman, G. R., Smith, E. K. M., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21*, 297-304.

Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Academic Medicine, 70*(2), 104-10.

Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194-201.

Schmidt, H. G., Boshuizen, H. P. A., & Hobus, P. P. M. (1988). Transitory stages in the development of medical expertise: The "intermediate effect" in clinical case representation studies, *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 139 - 45). Montreal, Canada: Lawrence Erlbaum Associates.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611- 22.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., De Kock, C. A., Peperkamp, A. G. W., & Donkers, H. H. L. M. (1996a). Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher, 18*(4), 295 - 300.

Schuwirth, L. W. T., Vleuten, C. P. M. v. d., & Donkers, H. H. L. M. (1996b). A closer look at cueing effects in multiple-choice questions. *Medical Education, 30*, 44 - 9.

Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in Assessing Clinical Competence* (pp. 13 - 45). Montreal: Can-Heal publications.

Swanson, D. B., & Norcini, J. J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine, 1*(3), 158-166.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Verwijnen, M., Imbos, T., Snellen, H., Stalenhoef, B., Pollemans, M., Luijk, S. J. v., Sprooten, M., Leeuwen, Y. D. v., & Vleuten, C. P. M. v. d. (1989). The evaluation system at the Maastricht medical school. In H. G. L. Schmidt, M. Jr.; Vries, M.W. de Greep, J.M. (Eds) (Ed.), *New directions for medical education, problem-based learning and community-oriented medical education*. New York, etc.: Springer Verla

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

CHAPTER 5

VALIDATION OF SHORT CASE-BASED TESTING USING A COGNITIVE PSYCHOLOGICAL METHODOLOGY

SUMMARY

The purpose of this study is to assess whether case-based questions elicit other thinking processes than factual knowledge-based questions. Twenty general practitioners (GPs) and twenty students solved case-based questions and matched factual knowledge-based questions while thinking aloud. Verbatim protocols were analyzed. Five indicators were defined: extent of the protocols, immediate responses, re-addressed information, order of re-addressing information, and type of considerations: 'true-false' type or deliberations that have a magnitude and a direction: 'vectors'.Cases elicited longer protocol word counts than factual knowledge questions. Students re-addressed more given information than GPs. GPs gave an immediate response on twice as many occasions as students. GPs re-ordered the case information, whereas students re-addressed the information in the order it was presented. This ordering difference was not found in the factual knowledge questions. Factual knowledge questions led to true-false considerations whereas cases elicited mainly vectors. Short case-based questions lead to thinking processes representing problem-solving ability better than factual knowledge questions.

INTRODUCTION

The assessment of problem-solving ability has progressed substantially through research in the past few decades. The progression has led to a change in thinking about the nature of problem-solving assessment. The implicit traditional assumption viewed problem-solving ability as a generic and independent trait (Barrows, 1984). In this view an expert is someone who possesses a general ability enabling him/her to solve any given problem within the domain. The process of problem solving was assumed to rely heavily on the applications of general algorithms and predefined analytical pathways. This would imply that different experts would solve a similar problem using similar strategies.

Assessment instruments that were designed on the basis of this viewpoint focused on the assessment of the problem-solving *process* displayed by the examinee whilst solving the problems presented in an examination. A common denominator of these instruments is the use of complete and integral and realistic (often branched) simulations of patient cases (Barrows & Tamblyn, 1977; Berner, Hamilton, & Best, 1974; Helfer & Slater, 1971; McGuire & Babbott, 1967; Rimoldi, 1961).

The methods applied to validate these examination formats use a similar approach as those used to validate psychological tests. Most of these procedures are based on comparing scores on different test formats or comparing test scores on different occasions or of different examinee populations. These procedures are called *indirect validation* procedures, because they use the test scores to obtain an impression about the thinking processes of the candidates: an inference has to be made from the observed test score to a "latent" trait or ability (Cronbach, 1983).

Cognitive psychology, however, provided a different view on the nature of problem-solving ability (Norman, 1988; Regehr & Norman, 1996; Schmidt & Boshuizen, 1993; Schmidt, Norman, & Boshuizen, 1990). Problem-solving ability proved not as generic as formerly assumed. Instead the amount of transfer from one problem to a similar other one was shown to be quite limited. This was even true with problems within the same domain (Elstein, Shulmann, & Sprafka, 1978).

Probably even more surprising was the finding that problem-solving ability was in fact highly idiosyncratic. The processes that different experts apply to solve the same problem differed considerably from expert to expert. In most instances experts even did not 'solve' the problem in the traditional meaning of the word, but simply recognised the problem and automatically applied the correct solution.

The consequences of these findings for the assessment of problem solving were that more emphasis was placed on short linear cases focusing rather on the *outcome* of problem solving than the process itself (Bordage, 1987; De Graaf, Post, & Drop, 1987; Norman, 1988; Schuwirth, Van der Vleuten, De Kock, Peperkamp, & Donkers, 1996).

Validation procedures, however, did not change accordingly, but were still mainly based on the correlational approach. The outcomes of these studies were usually disappointing. In comparisons between instruments developed for the assessment of problem solving and for other abilities (particularly knowledge) intermediate to high correlations were found (Norman, Van der Vleuten, & De Graaff, 1991). The high correlations indicate that the amount of unique information of either method is limited; the intermediate correlations did not allow any

conclusion. Depending on the purpose and the expectations, researchers concluded that the "glass was half full", while others concluded that it was "half empty" (Norman et al., 1991).
A refreshing viewpoint on validity has been advocated by Ebel (Ebel, 1983; Flanagan, 1983; Gardner, 1983). He suggested not perceiving an educational test as a psychological test, but regarding it as a measurement with an intrinsic educational and "visible" meaning. In his opinion validity must be built into a test, and is therefore based on a careful construction of the test by applying quality control procedures including blueprinting, item review by experts, careful determination of scoring systems, etc. These validation procedures are referred to as *direct validation* procedures. Essentially, one does not consider most important correlations with other methods, but rather the carefulness of setting tasks to examinees and of deriving test scores from these (Van der Vleuten & Swanson, 1990).
Direct validation, however, is equally unable to reveal whether the desired problem-solving processes are actually assessed. Even the most careful direct validation procedure can only determine which thought processes test developers think will be elicited by the test. It does not determine whether these steps actually take place. An attractive way would be to adopt the methodology used in cognitive psychological research to study cognitive processes. A common technology in this field is the use of think aloud protocols. This allows a direct assessment of the cognitive processes by having the subjects verbalise their thoughts while they are tackling a problem (Ericsson & Simon, 1993). For that matter, a think aloud methodology could actually be viewed as an extension to Ebel's direct validation procedure.
This paper addresses the validity of short cases by applying the cognitive psychological methodology of think aloud protocols to assessment. The research question is: do short case-based questions elicit different thinking processes than factual knowledge-based questions do? For this case-based questions and matched factual knowledge questions were presented to experts and novices to study whether differences could be detected in their reasoning processes on both question types.
Using the results in cognitive psychological research on (medical) expertise the present research question was studied using five indicators.
A prominent and recurrent finding in the cognitive psychological research is that experts differ from novices in the way their knowledge is organised (Chi, Glaser, & Rees, 1982; Glaser & Chi, 1988; Schmidt & Boshuizen, 1993). In the development of expertise the building of so-called semantic networks, in which knowledge aspects are embedded within meaningful links with other knowledge aspects, is considered an important step. These networks enable experts to process information more rapidly and in larger steps ('chunking'). For the present study it is assumed that short case-based tests require the presence and use of the networks more than factual knowledge questions. The following indicators were used.
The first indicator consisted of extent of protocols, expressed in word counts. Since cases contain more information that needs to be processed than factual knowledge questions, it was expected that word counts on case-based questions would be higher. In addition it was expected that experts would generate less extensive protocols in solving cases than novices, whereas such a difference would not be present in the factual knowledge questions.
A second indicator was the number of occasions in which an immediate answer was given. With increasing expertise, semantic networks aggregate even further into illness or instance scripts. Therefore, experts were expected to produce an immediate answer more often than novices on the case-based questions, whereas no such difference was expected in the factual knowledge questions.

A third indicator was based on the amount of given information that is reread after reading the question. It was expected that experts would be better able to store and retrieve the information in the case, as a result of their more elaborate semantic networks, and thus would reread less information than students. Again it was expected that such a difference would not occur in the factual knowledge questions.
The order in which subjects processed the information given was the fourth indicator. An important characteristic of expert illness scripts is that their activation in a reasoning process automatically generates a set of expectations regarding other patient findings. It was therefore expected that in the case-based part, experts would pick up information non-sequentially following their own networks or scripts. Students, who do not have these knowledge structures, were expected to pick up information sequentially, following the order in which the information is provided in the case.
The final indicator was the difference in considerations needed to solve cases and factual knowledge questions. Solving a case often implies that many different probabilities must be weighed against each other. Therefore two types of considerations were studied. The first would only have the value true or false ("HIV penetrates cells"); the second would have a direction and a magnitude ("With respect to the hypothesis that a male 45-year old patient has gallstones, his gender is a weaker argument against the hypothesis than his age is in favor"). The first were called true-false considerations, the second vectors. It was expected that in the case-based questions the percentage of vectors would be higher than the true-false considerations.

METHOD

Instrument

A set of four short cases was developed based on the so-called 'key-feature' concept (Bordage, 1987). All cases were based on real patients and were carefully reviewed on content and wording in two separate review cycles. Each of the cases was matched on content to four factual knowledge questions. Thus, each candidate was presented 4 cases and 16 factual knowledge questions. Question formats were adapted to the content of the questions, (Schuwirth et al., Submitted for publication) which resulted in 8 multiple choice and 8 open-ended factual knowledge questions, and 2 multiple choice and 2 open-ended case-based questions. An example of a case and its four matching factual knowledge questions is presented in table 5.1.

Table 5.1: An example of a short case-based question with its factual knowledge questions.

Case

You are a general practitioner. You see Sylvia, a 17-year old girl, again at your practice. She is now accompanied by her mother. About 10 months ago the problems started. She was tired all the time and her common cold and her sore throat would not go away. Because she had examinations she "forced you" to prescribe her antibiotics for which she then appeared to be allergic. You could not find anything. Even a quite extensive lab investigation did not reveal anything abnormal; the monosticon test was negative. In the subsequent visits it became clearer to you that her choice of school had not been the correct one, but that her parent insisted that she finishes her year. Now she has a sore throat again and her mother tells you that she keeps her parents awake all night with her coughing.

On examination you find that her tonsils are swollen with a pale greyish appearance. In her neck you find a distinct lymphadenopathy. Further ENT examination reveals no abnormal findings. She feels tired and has difficulties swallowing. Her temperature is slightly increased (38.2 °C). Her mother asks you whether Sylvia can go to school or that she has to stay home.

Which is the most correct answer to the question whether she can go to school?

a she must not go to school with these symptoms
b she can try to go to school depending on how the symptoms develop
c there is no reason for her to stay home

Related knowledge questions

- Name one angina that involves in the majority of the cases only one tonsil.
- All mononucleosis infectiosas that are caused by an infection with the Epstein-Barr virus produce heterophilous antibodies (Paul-Bunnell test) that are detectable in the blood. In a certain group of patients, however, these antibodies are NOT detectable in the majority of the cases.
  This group is:
  a pregnant women
  b children under 5 years old
  c children over 5 years old
  d adolescents
- Certain viruses are lymphocytotrophic, meaning that on infection they infiltrate lymphocytes. Name two lymphocytotrophic viruses that are pathogenic for humans. (write-in format)
- Which is the most frequent route of transmission of the Epstein-Barr virus?
  a direct skin contact
  b hematogenetic route
  c fecal-oral route
  d via saliva

Subjects

Subjects were 20 general practitioners with a minimum of 5 years of experience and 20 medical students at the end of their fifth year. In this year students have completed the clerkships in internal medicine, surgery and general practice in random order. All participants received a financial compensation for their efforts.

Procedure

Cases and factual knowledge questions were randomly presented. Candidates were tested individually. All test material was presented in written form. The candidates were instructed to think aloud whilst answering the questions. This was audiotaped and typed out. After the instruction the research leader did not speak to the participants other than to remind them to think aloud.

Analysis

The five indicators studied were:

- *Extent of protocols*

The extent of protocols was operationalised as word counts. In these count analyses, all words were included, with exception of those verbalised when reading the case and those used to 'fill' a thinking pause (e.g. "eh", "ah let me think"). Results are expressed as number of words.

- *Immediate final response*

Pattern recognition was operationalised based on the speed with which the first answer was given. Correctness of the response was not taken into account. Two exclusion criteria for this step were defined; the pause between reading the question and giving the answer may not be longer than 10 seconds, and no reasoning process may have taken place either before or after the answer was given, with the exception of a confirmation of the answer (e.g. "Eh, yes that is it"). Results are expressed as a percentage.

- *Returning to the stem or case after having read the question*

For this indicator, all stems and cases were divided into information units and remaining words (needed simply to make correct sentences, e.g. articles) All information units used were scored. If a candidate read the same information unit again, it was counted as a look up. Re-addressing information units before the question was read was not taken into account. The number of information units readdressed has been used for analysis.

- *Sequential or non-sequential re-addressing of information units*

It was further evaluated whether information units were picked up in a sequential or in a non-sequential manner when the candidate re-addressed the stem or case. Results are expressed as percentages of information that was retrieved sequentially or non-sequentially.

- *'True-false considerations' and 'vectors'*

All considerations were classified in either true-false types or vectors. Results are expressed as percentage of cases of either type.

Statistical analyses

Because of the non-normality of the data, non-parametric testing of significance was used for testing differences between groups and/or question types. For the discrete variables (word count and the number of readdressed information units) Mann-Whitney U tests were used. For the dichotomous variables a χ^2 test for contingency tables was used. In all tests $p< 0.05$ was considered significant.

RESULTS

Extent of protocols

General practitioners' protocols were less extensive than students'. In the case-based question the mean number of words used by the GPs was 121.33, whereas the students' mean was 135.29. This difference was not significant (U = 3074.00, p=.667). In the factual knowledge-based questions the mean for the GPs was 50.09, and for the students 38.68. This difference was significant (U=41300.00, p<.0001).

Immediate response

Figure 5.1 presents the percentages in which a candidate gave an immediate response. The cases are numbered 1 to 4, the corresponding factual knowledge-based questions numbered 1 to 16. The means of the case and factual knowledge questions are denoted by "Mean cases" and "Mean fact". Overall immediate responses seem to occur more frequently in factual knowledge questions, with no difference between GPs and students (χ^2 = 0.163, df=1, $p>1$). In the case-based questions GPs give an immediate response on about twice as many occasions (χ^2 = 7,56, df 1, $p<.01$).

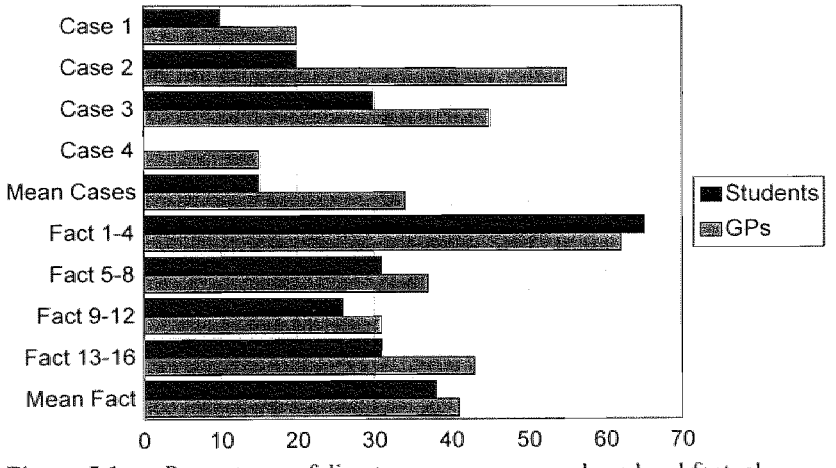


Figure 5.1: Percentages of direct responses on case-based and factual knowledge-based questions

Information units

Figure 5.2 provides an overview of the number of information units that are retrieved from the stem or case after the question was read. In the factual knowledge questions the numbers are quite low for both groups, which is logical because a stem often contained less information than a case. In the cases, students needed to pick up about double the amount of information compared to the GPs (U=1707.5, $p<.001$). As expected, this difference does not occur in the factual knowledge questions (U=6840.5 $p=.161$).

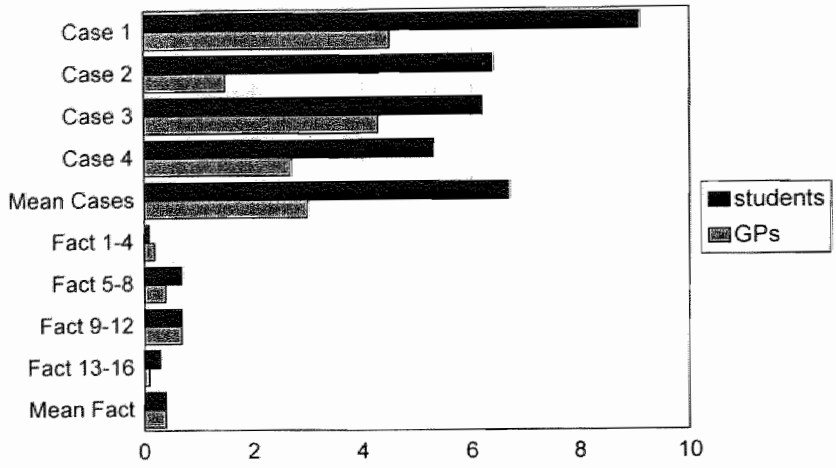
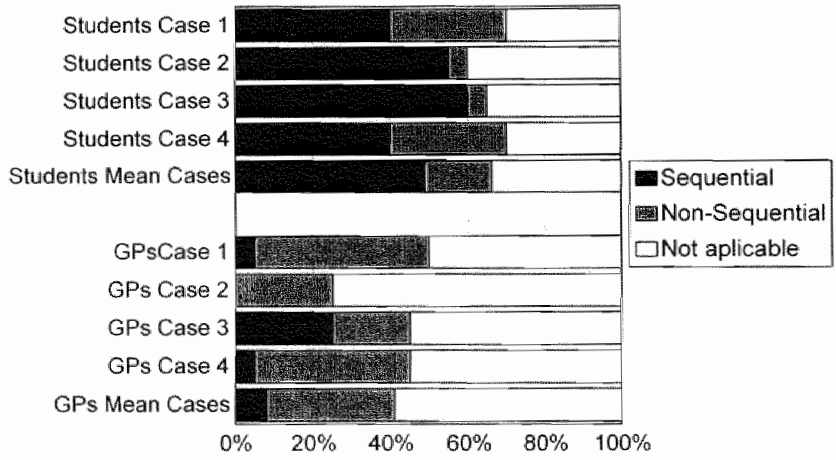


Figure 5.2: Number of information units retrieved on case-based and factual knowledge-based questions

Sequential versus non-sequential case processing

Figure 5.3 provides the percentages of sequential and non-sequential information processing in the student and GP group. Despite the definition of inclusion and exclusion criteria, the



distinction between sequential and non-sequential was too subtle to be categorised in 34 % to 59% of the occasions (for students and GPs respectively). These are denoted 'not applicable' in the figure. It is clear that the students mainly retrieve information sequentially, whereas GPs mainly retrieve information non-sequentially (χ^2 =22.63, df=1, $p<.001$).

True-false considerations or vectors

Figures 5.4 and 5.5 present the percentages of true-false considerations and vectors occurring in the cases and factual knowledge questions in the student and the GP group. In both groups the majority of the considerations while solving the cases are of a vector type. The opposite is true with the factual knowledge questions where true-false type considerations occur more often (GPs: χ^2=194,8, df=1, *p<.0001* and Students: χ^2 = 296,7, df=1, *p<.001)*.

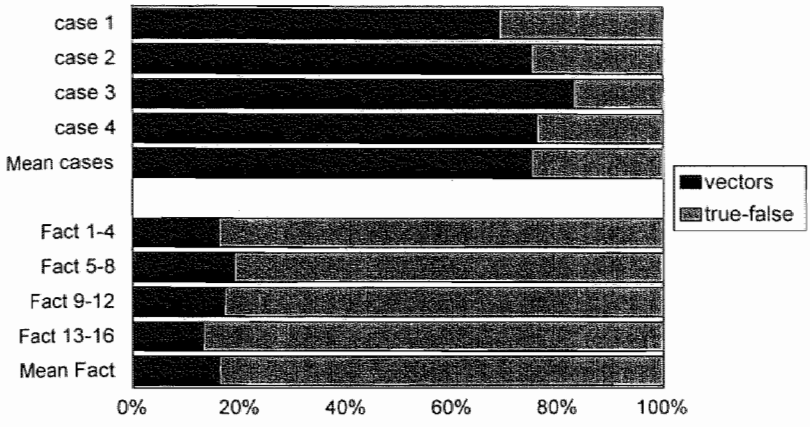


Figure 5.4: Percentages of 'vector' and 'true-false' thinking steps of students.

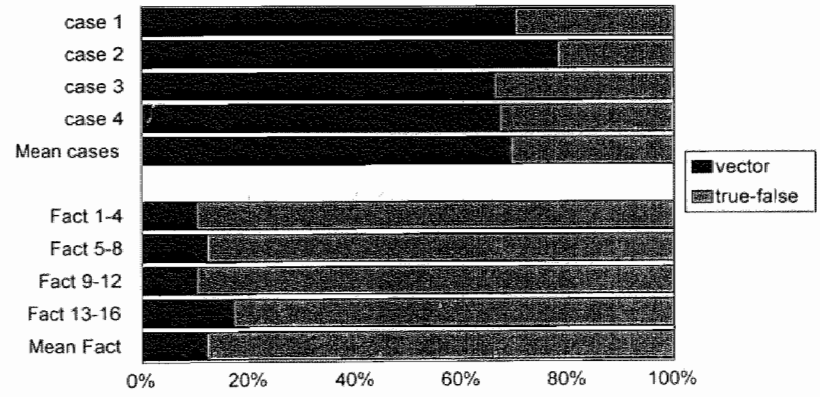


Figure 5.5: Percentages of 'vector' and 'true-false' thinking steps of general practitioners.

DISCUSSION

The results of the present study strongly indicate that short case-based examinations elicit different reasoning processes than factual knowledge questions do. In addition, thinking steps found in the short cases are theoretically meaningful and congruent with what research on clinical expertise development has delivered.

The results of the word counts, however, partly contradicted our expectation. In line with our expectations was the finding that cases elicited more extensive protocols suggesting more information processing. On the other hand, though, the expected pattern, in which a difference between groups of expertise would be found in the cases but not in the factual knowledge questions, did not occur. This could be due to the open instructions to the subjects (e.g. not correcting them when deviating from the main lines of reasoning).

An immediate answer was given by the GPs on the case-based questions on more than twice as many occasions as the students, whereas no difference was found in the factual knowledge questions. A direct answer must be the consequence of a recognition of the problem (and its solution) and must be based on a stored illness script or at least on extremely quick (but not verbalised) analyses. Since illness scripts are seen as the consequence of expertise development and experience, it can be concluded that short cases appeal more to higher-order expertise than factual knowledge questions do.

A further indication that cases appeal to illness scripts is the fact that in the cases GPs retrieved only half the information units that students did, whereas again no difference was found in the factual knowledge questions. This would imply that the information in the cases has more semantic meaning to GPs than to students, because it is possibly better embedded into existing scripts.

The difference between sequential and non-sequential retrieval of information underscores this even further. In order to recall and process information experts often re-order the information to match the order in which they are used to processing it. This has been demonstrated in previous research, such as in the work of Norman et al. (Norman, Brooks, & Smith, 1987). They presented lab data to experts and novices and asked them to memorise the material. Experts outperformed the novices, but in their turn were outperformed by the novices when confronted with nonsense values of the lab data. When confronted with authentic lab data but in a different order than that they were used to, they started rearranging the information in order to process it in the sequence they were used to.

Finally, the results show that placing the questions in a realistic context has a considerable effect on the type of cognitive operations that take place. 'True-false' considerations are the main ingredients of the solution of factual knowledge questions, whereas most considerations used to solve a case-based question are 'vectors'. In these steps knowledge is not merely a simple fact but it has a meaning to the decision. It can be either in favor or against a hypothesis, and it can be either strong or weak. This 'true-false/vector model' fits well with the theories on clinical expertise.

In the theory of expertise development of Schmidt and Boshuizen, expertise is viewed as a process in which isolated knowledge is gradually integrated into meaningful causal networks and aggregated into illness and instance scripts, comparable to the development of reading skills from isolated recognition of letters to direct recognition of (combination of) words (Schmidt & Boshuizen, 1993). In this view the true-false considerations must be the

representation of facts, the vectors would be the representations of causal networks and illness scripts would be the representation of memorised vector systems. In this view it can be assumed that an immediate answer to a case-based question is the result of more compiled knowledge (internalised vector systems).
The theory of Papa and Stone suggests that despite the idiosyncrasy and domain specificity of expertise a case contains an in-built diagnosability (Harasym, Papa, & Schumacker, 1997; Papa & Elieson, 1993). The vector systems could relate to this intrinsic diagnosability of a case. If the differences in direction and magnitude of the different vectors in a case are small, the case would be intrinsically more difficult to solve, i.e. the resultant vector would be more difficult to 'calculate'. It must, however, be kept in mind that the chance that a case is solved is a combination of this intrinsic homogeneity of vectors and the specific exposure of the problem solver to the clinical problem. This diagnosability problem would, however, not exist for an experienced person (with much exposure to these cases). The vector homogeneity is of no importance then, since no matter how complicated the vector system is, the case is simply recognised by this person.
The change from a rather fixed unidimensional thinking (true-false) towards multidimensional variable considerations (vectors) may explain the difficulties that students often face during the transition from a theoretical learning environment to a practical setting (Boshuizen, 1996). The theoretical part of many training programs is mainly based on the learning of (many) true-false facts, whereas practice often requires 'vector thinking'.
The numbers of candidates tested and the number of questions used in this study are small, restricting the generalisability. Although statistical testing reveals that the differences are all significant and largely in accordance with the expectations, replication studies need to be done. In these studies larger numbers of cases and candidates must be incorporated to replicate the effects found in this study. Still, the conclusion that case-based tests assess different thinking processes than factual knowledge-oriented questions do, seems to be justified.
The use of a cognitive psychological methodology appears to be a successful approach as a direct validation procedure. It seems better able to demonstrate difference between test formats and (theory-grounded) differences between groups of different expertise levels than methodologies based on comparisons or correlations between scores.

REFERENCES

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Barrows, H. S., & Tamblyn, R. M. (1977). The Portable Patient Problem Pack: a problem-based learning unit. *Journal of Medical Education, 52*, 1002 - 4.

Berner, E. S., Hamilton, L. A., & Best, W. R. (1974). A new approach to evaluating problem-solving in medical students. *Journal of Medical Education, 49*, 666 - 72.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Boshuizen, H. P. A. (1996). *The shock of practice; the effects on clinical reasoning.* Paper presented at the Paper presented at the 1996 annual meeting of the American Educational Research Association, New York, April 8-14.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7 - 76). Hillsdale NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1983). What price simplicity? *Educational Measurement: Issues and Practice, 2*(2), 11-12.

De Graaf, E., Post, G. J., & Drop, M. J. (1987). Validation of a new measure of clinical problem-solving. *Medical Education, 21*, 213-218.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7 - 10.

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis.* Cambridge Massachusetts: Massachusetts Institute of Technology.

Flanagan, J. C. (1983). A rational rationale. *Educational Measurement: Issues and Practice, 2*(2), 12.

Gardner, E. F. (1983). Intrinsic rational validity: Necessary but not sufficient. *Educational Measurement: Issues and Practice, 2*(2), 13.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv - xxviii). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Harasym, P. H., Papa, F. J., & Schumacker, R. E. (1997). The structure of medical knowledge reflected in clincians' estimates of the probabilities of signs/symptoms within diseases. In A. J. J. A. Scherpbier, C. P. M. Van der Vleuten, J. J. Rethans, & A. F. W. Van der Steeg (Eds.), *Advances in Medical Education, proceedings of the 7th Ottawa Conference* (pp. 602 - 7). Dordecht, The Netherlands: Kluwer Academic Publishers.

Helfer, R. E., & Slater, C. H. (1971). Measuring the process of solving clinical diagnostic problems. *British Journal of Medical Education, 5*, 48-52.

McGuire, C. H., & Babbott, D. (1967). Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement, 4*, 1 - 10.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 86.

Norman, G. R., Brooks, L. R., & Smith, E. K. M. (1987). *Expert-novice differences in recall of numerical laboratory data: Resolution of a paradox.* Paper presented at the 26 th Conference on Research in Medical Education, Washington DC.

Norman, G. R., Van der Vleuten, C. P., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education, 25*(2), 119-26.

Papa, F. J., & Elieson, B. (1993). Diagnostic Accuracy as a Function of Case Prototypicality. *Academic Medicine, 68*(10), S58-S60.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: Implications for professional education. *Academic Medicine, 71*(9), 988 - 1001.

Rimoldi, H. J. A. (1961). The test of diagnostic skills. *Journal of Medical Education, 36*, 73 - 9.

Schmidt, H. G., & Boshuizen, H. P. (1993). On acquiring expertise in medicine. Special Issue: European educational psychology. *Educational Psychology Review, 5*(3), 205-221.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611- 22.

Schuwirth, L. W. T., Blackmore, D. B., Mom, E., Van de Wildenberg, F., Stoffers, H., & Van der Vleuten, C. P. M. (Submitted for publication). How to write short cases for assessing problem-solving skills.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., De Kock, C. A., Peperkamp, A. G. W., & Donkers, H. H. L. M. (1996). Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher, 18*(4), 295 - 300.

Van der Vleuten, C. P. M., & D. Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58 - 76.

CHAPTER 6

AN INTER- AND INTRA-UNIVERSITY COMPARISON WITH SHORT CASE-BASED TESTING

SUMMARY

Comparisons between PBL and non-PBL medical schools on problem-solving ability often show no differences. This could be because no difference in problem-solving ability exists or because the instruments used are inadequate. In this study a key-feature approach case-based examination was used to compare two medical schools in the Netherlands, one of which has a PBL curriculum (Maastricht) and one which has a programme in transition from a non-PBL towards a PBL curriculum (Groningen). Differences were found both in proficiency scores and in the pattern of response times, both supporting the assumption that a PBL approach would lead to a higher level of problem-solving ability. The effect size, however, is not as large as originally assumed by the PBL proponents. Conclusions must be drawn with caution, but it seems likely that a test based on large numbers of short cases is the most sensitive in detecting differences in problem-solving ability between students of different curricula.

INTRODUCTION

A claim of problem-based learning (PBL) curricula is that their medical students would become better problem solvers than students in a more traditional curriculum (Barrows & Tamblyn, 1977). This claim was largely founded on the assumption that PBL students would acquire a general problem-solving ability which they could then apply to the medical problems they would encounter in practice. In the past decade, however, serious doubts have arisen as to whether this assumption is correct (Elstein, Shulmann, & Sprafka, 1978; Norman, 1988; Schmidt et al., 1996). A large body of research has studied possible differences in problem-solving ability between students of different medical schools. The vast majority of the outcomes indicated that results of both groups of students were comparable. In some instances, however, a small superiority of PBL students on clinical knowledge was demonstrated (Albanese & Mitchell, 1993; Schmidt, Dauphinee, & Patel, 1987; Vernon & Blake, 1993). The question arises, however, whether this actually represents a difference in problem-solving ability or merely a difference in knowledge domains. More conclusive demonstrations of the assumed superiority in problem solving have not occurred. Two explanations for this lack of evidence can be suggested. Either there is no difference in problem-solving ability between different student groups or the measurement instruments used to assess problem solving were inadequate.

The question whether or not a difference in problem-solving ability exists between PBL and traditional students originates from the developments in the theory on the nature of problem-solving ability. Problem solving has long been considered a generic trait or skill, implying that someone's level of ability allows the person to deal with many different problems. The first claim concerning the superior problem-solving ability of PBL students is based on this assumption (Barrows, 1984).

Modern cognitive psychological theories, however, consider the acquisition of problem-solving ability as a multistage process highly related to knowledge (Schmidt, Norman, & Boshuizen, 1990). In these theories the process is supposed to start with the storage of knowledge in so-called semantic or causal networks of different facts and their interrelationships. These networks can be aggregated into more abridged ones, by forming clusters of information leading to higher level causal models. Furthermore, illness scripts (mental representations of a disease) can be formed which enable the problem solver to compare a certain problem with the stored illness scripts in order to come to a clear problem definition and its subsequent solution. The last and most expert stage involves the use of instance scripts in which previous encounters with similar patients form a basis on which the problem and the solution are found. This ongoing aggregation leads to a higher level of efficiency in problem solving which enables straightforward pattern recognition with immediate diagnosis and selection of management. Some general characteristics of problem solving emerge from this theory (Boreham, 1994; Regehr & Norman, 1996; Schmidt et al., 1990). First, the ability to solve a problem is embedded within the problem itself; the amount of transfer to another, similar problem is limited. Problem-solving ability is therefore highly domain specific. Second, since networks, models and scripts are highly individual, problem solving is idiosyncratic. The manner and the extent to which these various strategies are applied to a specific problem differ from person to person. Third, more experienced physicians may be only moderately more accurate in their assessment of

the problem (and the possible solution) than less experienced physicians but they certainly are more efficient (faster) at solving it. Finally, problem solving does not always involve a conscious analytical process, but is often based on pattern recognition.

This highly domain-specific and idiosyncratic nature of problem-solving ability is not in accordance with the assumption that PBL students would be educated to be generic problem solvers. The only potential difference between PBL and non-PBL students in problem-solving ability would therefore stem from the fact that PBL students have seen and solved more problems during their study (Schmidt et al., 1996). This would result both in a (slightly) higher proficiency in problem solving and in shorter response times.

A second potential reason why a difference in problem-solving ability between PBL and non-PBL students has not been found may lie in the inadequacy of the instruments used. The developments in these instruments have followed the insights of the concepts of problem solving previously described. The original idea of designing instruments that mimic real practice as closely as possible, such as the use of long branched cases (e.g. Patient Management Problems), suffered from serious flaws. The emphasis that these tests placed on thoroughness of data gathering disadvantaged experienced clinicians, since these clinicians needed much less data to come to a conclusion than novices (Swanson, Norcini, & Grosso, 1987). This casts serious doubts on the construct validity of these instruments. Furthermore, the length of these cases precluded the administration of large numbers per session. Because of the domain specificity only low generalisabilities could be obtained (Elstein et al., 1978). This has led to other developments in the testing of problem solving (Swanson et al., 1987).

Some experiments, for example, have attempted to measure the pattern recognition accuracy (Norman, 1989). In these experiments either brief presentations of cases with slides or series of lab results were used as stimuli. Experts were shown to process these data better and more quickly than novices. The translation to concrete assessment tools for medical students, however, is still not clear. Another development lies in the focussing on other outcome variables than proficiency, such as response time (Norman, Swanson, & Case, 1996). Although some effects can be found indicating that more expert students indeed outperform less expert ones, these effects and their practical implications are still not clear. The matter of how to combine response times with proficiency scores on a test to come to a relevant competence score, for example, is unclear.

The common denominator of modern problem-solving assessment methods, however, is the use of short linear cases instead of long branched ones (Bordage, 1987). These cases are reduced to their essential elements as are the questions. This enables the administration of many different cases per period of testing time, which leads to a more adequate sampling than the long case examinations without much loss in authenticity of the case.

In addition, a large number of studies exist, exploring the influence of item format on the competences being measured. The most consistent conclusion is that the influence of the format is trivial, whereas the influence of the content of the stimulus or question is of paramount importance (Maatsch & Huang, 1986; Norman et al., 1996; Schuwirth, Vleuten, & Donkers, 1996c; Swanson et al., 1987; Ward, 1982).

To detect a difference in problem-solving ability between different curricula, if any, would therefore need an instrument that incorporates the outcomes of all these developments.

Using such an approach Schmidt et al. recently conducted a study between universities in the Netherlands comparing (amongst others) one PBL curriculum with one traditional

curriculum (Schmidt et al., 1996). In their study they presented 612 students with 30 short cases each prompting for the (differential) diagnosis. An overall effect of curriculum was found in the sense that on graduation the PBL students outperformed the non-PBL students in about 5% of the cases. Although the authors advocate prudence in drawing conclusions, one of their explanations is the assumption that the subject matter integration and active processing in PBL curricula helps students acquire proficiency in diagnostic reasoning. Although this is a very elegant study in terms of research questions and methodology, two comments can be made. All cases used in this study prompted for a diagnosis. In some medical cases, however, the actual diagnosis may not be the key element, but treatment or management may be more significant. In addition, the authors did not use other outcome variables than proficiency scores. These limitations might have diminished the sensitivity of the instrument to detect a difference.

All of this has led to the research question for the present study: is the absence of differences between PBL and non-PBL curricula due to the inadequacy of the instruments or to the fact that no difference exists?

More concretely, when using an optimally designed short case-based examination to compare a medical school with an integrated problem-oriented curriculum with a school in which part of the students followed a traditional lecture based curriculum and part of the students an integrated curriculum, it was expected to find differences between those students following an integrated curriculum and those following the more traditional curriculum. Differences were expected both in proficiency and in efficiency.

METHOD

Subjects

Students of two universities (Maastricht and Groningen) in the Netherlands were used. Both medical curricula take six years and are divided into four pre-clinical and two clinical (clerkship years). In Maastricht all students follow a problem-based curriculum in which tutorial groups, lectures and clinical skills training are built around (patient) cases. In Groningen the students of the first three year groups follow a thematic patient-based curriculum in with patient demonstrations, followed by selfstudy, tutorial groups, clinical skills training, and interactive lectures. Per week a different theme is addressed. In this school students of the last three year groups followed a lecture-based non-integrated curriculum. In both medical faculties students from all six year groups were involved. In total 355 students participated. Numbers of participants per year group and per faculty are reported in table 6.1. All students were volunteers. The participating students of Maastricht were compared to the rest of their class on regular test results to detect whether the sample was comparable to the population. Only in the first year group was the sample significantly different (in favour of the sample) from their class; in all other year groups mean scores of the sample and those of the rest of the year groups were virtually equal. Unfortunately similar data were unobtainable for the students of Groningen. Participants received a financial compensation for their efforts, but in addition an extra sum of money (50 Dutch guilders) was given to the student obtaining the highest score within each year group and faculty. This was done to simulate the achievement-motivational aspects of a real exami-

nation to some extent. All test administrations within each school were held within a time frame of 2 weeks. The pause between the administration of both schools was 2 weeks.

Instrument

The test used in this comparison was a case-based computerised test (CCT) in which 60 cases were presented. A detailed description of the test and some sample items are described elsewhere (Schuwirth, Van der Vleuten, De Kock, Peperkamp, & Donkers, 1996a). In all cases the 'key-feature approach' was used (Bordage, 1987). In this approach case descriptions are kept brief and questions are only aimed at essential decisions. To cover a wide and general domain in medicine, a test of general practice was used for this comparison. All cases were written by different general practitioners (GPs) and were based on real life patient contacts. Subsequently all cases were extensively reviewed by two other GPs and a medical educationalist. After this, in a second review process, 10 residents in general practice with three years of practical experience (originating from different medical faculties in the Netherlands) were asked as experts to validate the cases and the answer keys. In the test used in this study the experts agreed on the correct answer for nearly all cases included. In a few instances a defendable 'minority opinion' existed. In those cases partial credit was given.

Different question formats were used varying from short answer open-ended questions to multiple choice types of questions. The selection of the question format is based on the content of the question: when only a limited number of realistic alternatives could be generated multiple choice questions were used; in other cases open-ended question types were used (Schuwirth, Van der Vleuten, Stoffers, & Peperkamp, 1996b).

Participants were instructed to read the case description carefully and then click on a button to display the question. Because differences in case reading time were not of interest and since the question was only visible after the case had been read, the time needed to read the cases was recorded separately from the time needed to read and answer the questions. The latter was considered to be a more valid indicator for the response time than the former or a combination of both. Therefore, response times in this chapter pertain only to the latter.

Testing time was limited to 2.5 hours, which was ample for all participants.

Statistical analysis

Scoring of correct answers (1 point each) and registration of response times (in seconds) was performed automatically. Scores are expressed as percentage correct answers. Response times are calculated by the total time needed to answer all the questions divided by the number of cases. Means and standard deviations were calculated for the scores per year group and per medical school.

The effect of year group on scores was tested using a one-way ANOVA with a Tukey's HSD as post-hoc analysis. Subsequently, a two-way ANOVA (faculty x year group tested against mean score) was performed. Independent samples T-tests after a Bonferroni step-down procedure were used to determine the significance of the individual differences between equal year groups of both universities. As a result of the transformation $p \leq 0.01$ was considered significant.

Since the distributions of the response times contain some extremes, medians and ranges were calculated. Subsequently these were broken down to median response times on correctly and incorrectly solved cases. To establish the significance of the differences

between the faculties and between the correctly and incorrectly solved cases within each faculty, Kruskal-Wallis multiple comparisons were used. A Bonferroni step-down procedure was used to estimate the level of significance required. As a result of this procedure $p \leq 0.015$ was considered significant.

RESULTS

Table 6.1 reports the descriptive statistics of scores of both faculties. From this table it is clear that in the first four year groups virtually no differences are found between both schools. In the fifth and the sixth year group, however, differences emerge. The one-way ANOVAs show a significant main effect of year group on score (Maastricht: $F(5,172) = 58.75$, $p < .0000$; Groningen: $F(5,171) = 12.50$, $p < .0000$).

Table 6.1: Descriptive statistics of the scores (percentages correct)

	Groningen students			Maastricht students		
Year group	N	Mean	SD	N	Mean	SD
1	32	43.8	5.0	30	44.1	5.9
2	30	45.5	4.4	30	46.6	4.2
3	30	50.5	4.4	30	50.0	5.6
4	30	52.5	6.3	29	53.7	5.0
5	30	52.2	5.6	27	56.3	5.9
6	25	57.9	5.9	32	65.3	6.3

Post-hoc analysis shows that the effect in Maastricht can be explained mainly by differences between the last two year groups and the first four, whereas the effect in Groningen is mainly due to the difference between the first two year groups and the last four. The 2 x 2 ANOVA yields two main effects (year group and faculty) and a significant interaction effect (F (1,5,354) = 5.50, $p < .000$). These effects can be explained by the differences found between the faculties in the year groups 5 and 6. Both effects are significant ($p \leq 0.01$ and $p \leq 0.001$ respectively).

To obtain a clearer impression of the trend of the scores over the year groups figure 6.1 provides the means plus 95%-confidence intervals. Although the comparison has been a transversal instead of longitudinal, a line graphic is used in order to make the trend more clearly visible. What is striking is the fact that the lines virtually converge in the first three year groups and begin to diverge in the fourth year in favour of Maastricht students.

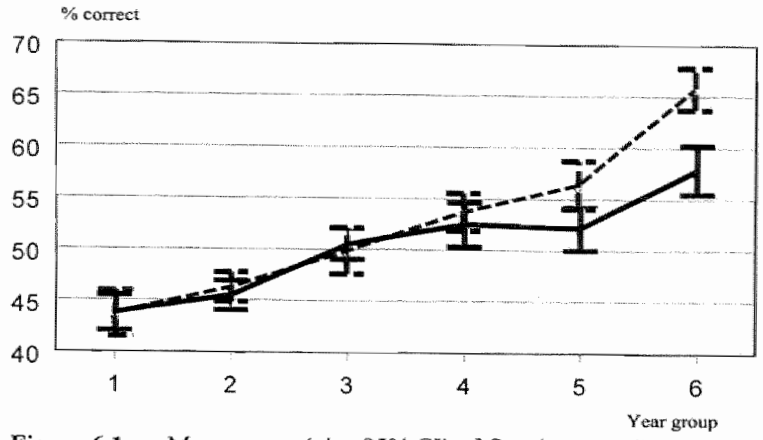


Figure 6.1: Mean scores (plus 95% CI) of Groningen (solid line) and Maastricht (broken line) students.

In table 6.2 the results of the median response times are described. Differences between both faculties are small.

Table 6.2: Response times medians of all items (M_o), incorrectly (M_i), correctly (M_c) solved cases and level of significance of the difference between M_i and M_c.

	Groningen students				Maastricht students			
Year group	M_o	M_i	M_c	p	M_o	M_i	M_c	p
1	16.7	19.6	16.0	*.001*	20.7	22.6	18.7	*.0001*
2	16.2	17.0	16.7	*.915*	20.9	22.2	19.3	*.008*
3	18.1	20.4	17.8	*.069*	19.4	21.7	17.3	*.0001*
4	18.4	21.2	18.6	*.048*	19.2	22.1	17.1	*.0001*
5	19.5	21.0	21.1	*.981*	21.5	25.0	17.9	*.0001*
6	17.9	17.2	19.1	*.004*	17.5	22.1	14.8	*.0001*

On average the Groningen students completed the test faster than their Maastricht colleagues in all but the final year group. In addition two aspects are remarkable. First, the differences in response times between the correctly and incorrectly solved cases are highly significant in the case of the Maastricht students, whereas these difference are only clearly significant in the first and last year group of the Groningen students. Second, the respons times of the correctly solved cases have a tendency to increase in the Groningen group, whereas those of the Maastricht students tend to decrease. Both effects were significant ($p < .0001$ for both schools)

DISCUSSION

In this study an effect of curriculum on proficiency scores was found. The differences in mean scores upon graduation level appear to be somewhat larger than those found by Schmidt et al. (Schmidt et al., 1996). The difference in their study was about 5%, whereas in this study a difference of about 8% was found. This difference is more than one standard deviation of the scoring scale. This partly supports the hypothesis that a test requiring more than diagnosis alone would be more sensitive in detecting differences in problem-solving ability, but the difference in effect sizes between the present study and the Schmidt et al. study is rather small. The present study therefore confirms the results found in Schmidt et al's study. This is striking because in the comparisons between medical schools in the Netherlands using tests of factual knowledge, no difference has ever been found before at graduation level (Verwijnen, Van der Vleuten, & Imbos, 1987).

These findings point to the conclusion that the lack of difference found in those comparisons is mainly due to the inadequacy of the instruments used and that using a short case-based testing approach is more sensitive in detecting differences in problem-solving ability between curricula.

This hypothesis is based on the assumption that there is a real difference between the students of both curricula. The fact that a difference was found supports this assumption. It is perhaps somewhat strange that, in view of fact that the nature of the clerkships in Maastricht differs only marginally from those in Groningen, the differences were found mainly in the last year two to three year groups. It could, however, be argued that any effect of curricula of different schools needs time to emerge. Schmidt et al. found a similar pattern and suggested a sort of incubation effect, that would not occur before that final years of the study. In their assumption such an effect would need a sort of 'incubation type of period' before becoming overt. Our results, however, cannot provide additional evidence for the existence of such an incubation period other than a replication of the tendency of the mean scores as found in the Schmidt et al. study.

The fact that response times in Groningen were lower than those in Maastricht is somewhat surprising. An explanation of this could be that the Maastricht students took the test more seriously, but no other evidence for this assumption has been found.

Still, the tendencies of the response times in the correctly solved cases in both faculties is striking: Maastricht students tend to get faster with increasing expertise, whereas in the Groningen group the opposite occurs. This seems to be congruent with the results found by Van der Vleuten et al., who found a distinct decrease in mean response times with increasing expertise (Van der Vleuten, Schuwirth, & Ronteltap, 1995). An explanation could be that the correlation between response times and expertise forms a parabolical curve. Starting with the totally ignorant student discarding the case immediately an increase in response time could occur, whereas the more knowledgeable student could need more time to come to the correct answer. In a further stage as knowledge becomes more integrated, more pattern recognition would occur, leading again to a decrease in response times. This would then suggest that the Maastricht student's problem solving occurs at the falling end of the curve whereas that of the Groningen students is still at the rising or horizontal part of the curve. This explanation would, however, only be plausible if the

overall differences in response times between Groningen and Maastricht would indeed not be the consequence of differences in level of taking the test seriously.
Drawing a firm conclusion that a PBL-effect is responsible for all the differences found is hazardous. Comparisons like these are not simple because many of the variables cannot be controlled for. Schmidt reports several error sources of this kind of comparisons (Schmidt, 1990). His concerns can be categorised into three categories: population effects, sample effects and instrumentation effects.
Population effects would occur student's self selection due to their preference for a certain university and by differences in selection during the study based on different examination systems. These effects do not seem to influence our study significantly. First, because students in the Netherlands can enter medical school only after a rather homogenous secondary school system, after passing a national high school examination and after passing a lottery procedure, so a prior selection would only be marginal. Second, once past the lottery system they cannot simply choose the university they prefer, but are assigned to a university by a national committee from which personal preferences is but one variable. Inter-university comparisons have furthermore shown that the average level of knowledge of all medical students in the Netherlands is fully comparable (Verwijnen et al., 1987).
Another suggested objection are major and minor curriculum revisions. These would alter the effects of the curriculum in some way, so the 'treatment' is not kept constant. Revisions in Groningen, however, must not be considered to be a bias here. Quite the contrary, in this case they enable an even better comparison between both schools.
Objections to the sampling procedure do apply to the present study. Participants were paid volunteers and no randomisation procedure could be used. Nevertheless, at least the students of Maastricht were comparable to the rest of their year group, and the awarding of extra money to the best achiever might have simulated an actual examination situation somewhat more closely.
An instrumentation effect - the local familiarity with the instrument used - appears to be present on first sight, but its influence can be shown to be insignificant on closer inspection. Maastricht students in the last two year groups are somewhat more familiar with the instrument in the sense that they have some prior experience with computerised testing. But this is neutralised by an extensive structured instruction to all participants. Experiences in Maastricht with this instruction have shown that this is more than sufficient to master the program perfectly.
A difference between both schools is that most of the Maastricht students completed the General Practice clerkship in the fifth year and the Groningen students in the sixth year. If this had had an effect on the mean scores, it would have pointed in an opposite direction: the difference in the fifth year would have been larger than in the sixth year since the sixth year Maastricht students were most likely to have forgotten what they have learned during the GP clerkship (Semb & Ellis, 1996).
A more probable alternative explanation could be a "skillslab" effect. Students in Maastricht follow a longitudinal skillslab program in which all the necessary physical examination and communication skills are taught and practised. A prior comparison using OSCEs and written tests on skills between both faculties showed significant differences favouring Maastricht students in all six year groups (Scherpbier et al., 1996). The assumption would then be that Maastricht students are prepared differently for their clerkships by this program, and therefore would not have to bother with learning these skills during the clerkship. Instead they

could focus more on solving patient problems. This would explain why especially in the last two year groups differences were found using a test focussing on patient problems and why Schmidt et al. found a similar pattern in their study. In summary, the results suggest that the lack of differences found between PBL and non-PBL schools on problem-solving assessment is mainly due to an inadequacy of the instruments used and that PBL versus non-PBL comparisons on problem-solving ability would benefit from using tests of a large sample of short key-feature approach cases.

REFERENCES

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues. *Academic Medicine, 68*(1), 52-81.

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Barrows, H. S., & Tamblyn, R. M. (1977). The Portable Patient Problem Pack: a problem-based learning unit. *Journal of Medical Education, 52*, 1002 - 4.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Boreham, N. C. (1994). The dangerous practice of thinking. *Medical Education, 28*, 172-179.

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Maatsch, J., & Huang, R. (1986). *An evaluation of the construct validity of four alternative theories of clinical competence.* Paper presented at the Proceedings of the 25th annual RIME conference, Chicago.

Norman, G., Swanson, D., & Case, S. (1996). Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. *Teaching and Learning in Medicine, 8*(4), 208-216.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 86.

Norman, G. R. (1989). Reliability and Construct validity of some cognitive measures of clinical reasoning. *Teaching and Learning in Medicine, 1*(4), 194 - 99.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: Implications for professional education. *Academic Medicine, 71*(9), 988 - 1001.

Scherpbier, A. J. J. A., Pols, J., Nieuwenhuijzen Kruseman, A. C., Schaper, N. C., Verwijnen, G. M., & Van der Vleuten, C. P. M. (1996). Interfacultaire vaardigheidstoets Groningen-Maastricht: eerste resultaten. In T. J. Ten Cate, J. H. Dikkers, E. Houtkoop, M. C. Pollemans, J. Pols, & J. A. Smal (Eds.), *Proceedings van het vijfde Gezond Onderwijs Congres* (pp. 351-357). Houten/Diegem: Bohn, Stafleu, Van Loghum.

Schmidt, H. G. (1990). Innovative and conventional curricula compared: what can be said about their effects? In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in Medical Education: An Evaluation of Its Present Status* (pp. 1-7). New York: Springer Publishing Company.

Schmidt, H. G., Dauphinee, W. D., & Patel, V. L. (1987). Comparing the effects of problem-based and conventional curricula in an international sample. *Journal of Medical Education, 62*(4), 305-15.

Schmidt, H. G., Machiels-Bongaerts, M., Hermans, H., ten Cate, T. J., Venekamp, R., & Boshuizen, H. P. A. (1996). The Development of Diagnostic Competence: Comparison of a Problem-based, and Integrated, and a Conventional Medical Curriculum. *Academic Medicine, 71*(6), 658-664.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611- 22.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., De Kock, C. A., Peperkamp, A. G. W., & Donkers, H. H. L. M. (1996a). Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher, 18*(4), 295 - 300.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., Stoffers, H. E. J. H., & Peperkamp, A. G. W. (1996b). Computerized Long-Menu Questions as an Alternative to Open-Ended Questions in computerized assessment. *Medical Education, 30*, 50 - 5.

Schuwirth, L. W. T., Vleuten, C. P. M. v. d., & Donkers, H. H. L. M. (1996c). A closer look at cueing effects in multiple-choice questions. *Medical Education, 30*, 44 - 9.

Semb, G. B., & Ellis, J. A. (1996). Knowledge Taught in School: What is Remembered? *Review of Educational Research, 64*(2), 253 - 286.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van der Vleuten, C. P. M., Schuwirth, L. W. T., & Ronteltap, C. F. M. (1995). A cognitive psychological interpretation of a few remarkable psychometric findings. In A. I. Rothman & R. Cohen (Eds.), *Proceedings of the Sixth Ottawa Conference on Medical Education* (pp. 506 - 508). Toronto: University of Toronto Bookstore Custom Publishing.

Vernon, D. T., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68*(7), 550-63.

Verwijnen, M., Van der Vleuten, C. P. M., & Imbos, T. (1987). A comparison of an innovative medical school with traditional schools: an analysis in the cognitive domain. In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in Medical Education: An Evaluation of Its Present Status* (pp. 40-49). New York: Springer Publishing Company.

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

CHAPTER 7

HOW TO WRITE SHORT CASES FOR ASSESSING PROBLEM-SOLVING SKILLS

SUMMARY

In assessment of problem solving the use of short case-based testing is a promising development. In this approach an examination consists of large numbers of short cases each of which contain a small number of questions. These questions are aimed at essential decisions. Writing such cases, however, is not easy. In this chapter a description of this type of examination is provided. Also strategies and pitfalls are described in writing these cases. These strategies pertain to the selection of the essential decisions, the careful writing of the cases and the questions and the selection of the question formats.

INTRODUCTION

Although in the domain of testing of problem-solving skills long patient case simulations have been favoured for quite some time, there has been a tendency towards use of short cases since the mid-1980s (Bordage, 1987; Graaf, 1988; Schuwirth, Van der Vleuten, De Kock, Peperkamp, & Donkers, 1996; Van der Vleuten et al., 1994). The reason for this is that the assumptions on the nature of problem-solving ability underlying the long simulations have proven to be incorrect. The first of these assumptions considered problem-solving ability to be generic, implying that considerable transfer of the problem-solving process would occur from one case to another (Barrows, 1984). The second considered the problem-solving process to be uniform across experts, i.e. that different experts would solve the same problem the same way. The contrary assumptions, however, are more appropriate; problem solving is highly domain-specific (Elstein, Shulmann, & Sprafka, 1978; Swanson, 1987; Swanson, Norcini, & Grosso, 1987). The score an examinee obtains on one problem is often a poor predictor of the score on any other given case, even within the same domain or on the same topic. Regarding the second assumption it appeared that experts in most cases had difficulty to reach consensus on the optimal strategy to solve a certain problem, although they did agree on what the correct solution should be (Swanson et al., 1987). The process of problem solving is apparently highly idiosyncratic, even when all experts agree on its outcome.

These findings have had implications for the design of test instruments for the assessment of problem-solving ability. Instead of using a small number of long simulations focussing on the process of problem solving, tests should use a rather large number of short cases. Each case includes only a small number of problems and focusses on the outcome of the problem-solving process.

In the construction of these short case-based tests, however, it is of paramount importance to take considerable care of the description of the case, the selection of the problems to be asked and a clear and unambiguous description of the questions or decision points to be assessed.

Both at the Medical Council of Canada and the University of Maastricht in the Netherlands extensive experience exists with this format. The Medical Council has been using it for their national licencing examination since 1992. At the University of Maastricht such an approach has been used for end-of-clerkship examinations since 1993.

In describing a method to write short cases, this chapter outlines the most important developmental strategies and pitfalls in writing this type of test material. First, however, a general framework of a case with a question is presented to serve as a general model. Then more specific strategies and pitfalls are described.

A GENERAL FRAMEWORK OF A CASE

A case-based question consists of two (rather obvious) parts: the case and the question(s). Within the case a situation is described or information is presented which the examinee will have to use to solve the problem. In some cases the problem might already be presented within the case text. An example is provided in figure 7.1.

You are a general practitioner. You make a house call on Mrs Van Doorn (65 years old). She consults you because of an acute pain in her abdomen. She tells you further that..... On physical examination you find..... *You wonder whether she should be admitted to the casualty immediately or whether it is best to keep her home for the time being and treat her there unless her condition worsens.*
What is the most appropriate action?
A You have her admitted to casualty
B You keep her at home for the time being

Figure 7.1: The problem is already present in the case.

In other cases, as in figure 7.2, it might be easier to present the problem in the question:

You are a general practitioner. You make a house call on Mrs Van Doorn (65 years old). She consults you because of an acute pain in her abdomen, She tells you further that..... On physical examination you find..... *What is the most probable diagnosis?*

Figure 7.2: The problem is presented in the question.

Thus the information in the case is of vital importance. It should contain sufficient information to enable the examinee to make the decisions which are asked in the questions, but at the same time should not contain too much irrelevant information to divert the student. Different types of information are included within the case. A patient case for example contains both clinical and contextual information. The clinical information reports aspects such as signs, symptoms and findings, but also negative or normal signs, symptoms and findings (i.e. non-abnormal signs). The contextual information reports on aspects that are not the result of the patient's illness but which do have an influence on the decision an examinee has to take. This information consists of aspects such as physical surroundings, gender, age, family circumstances, non-verbal information.
Authenticity of a case is vital and highly dependent on an adequate report of the information. Thus, verbal information is presented 'raw' without any 'chunking' ("respiratory rate of 30/minute" instead of "tachypnea"), relevant visual and auditive information should be presented directly (e.g. by using multimedia), instead of a textual description.
The above problems are examples oriented to assessing essential decisions. Diagnosis and treatment, however, may not always be the essential problem of a case. Other aspects like processing information (in history taking or physical examination), (efficient) diagnostic management, informing the patient, etcetera may be more essential to the case.

Authenticity does not only pertain to the case description but also to the questions asked. Options or questions that are out of context of the case are to be avoided whenever possible. The case writer must be aware of the realistic options that exist in the real life situation. Some examples may clarify this point. When faced with a severe accident with two casualties with different traumas and only one ambulance, the decision which patient to send in first has only

two options, but is far from trivial. Deciding which labtest should be ordered to confirm a possible hyperthyroidism has only a limited number of options in real life. The recognition of a pattern of symptoms as a certain disease or syndrome often has an 'infinite' number of options in reality. These considerations should be taken into account when determining the format for each question.

STRATEGIES AND PITFALLS

This section discusses the strategies and pitfalls in three parts: the case description, the selection of problems and the questions.

CASE DESCRIPTION

1 Use the representation of real patients.

Three arguments support this strategy. First, real practice provides a rich source for possible cases and prevents the mind boggling about choosing a suitable subject for a case. Second, by randomly selecting patients to form the basis for cases, the test or item bank will cover daily practice more congruently than constructed cases. It will prevent the author from hobbyhorse riding or asking only "exotic cases". Third, and most important, constructed cases (out of a textbook) without representation to genuine patients are often very artificial. Be aware that it is always advisable to preserve the anonymity of your patients when constructing examination material from real case files.

2 Ensure that the description of the information is as clear as possible.

In order for the student to process the information the case should be unambiguous and written clearly. It should deal with the exact features that are present in real life. Phrases like "possible masses on palpation" constitute in fact non-information, because a mass could be either palpated or not. This doubt may exist in real life, but then the physician will at least have the memory of his/her sensations during the examination. An example of such a flawed case is presented in figure 7.3.

Mrs Whiteless consults you in your practice because of complaints of a vague pain in the lower abdomen. The pain is "somewhere in my abdomen, doctor". It has been present for quite a while now. The pain has a cramping aspect.
During further history she tells you that her period is regular, but that her last menstruation was one week later than normal.
On physical examination you find a dubious pressure tenderness in the lower abdomen and you feel a possible mass there. During gynaecological examination a drop of blood is seen on the portio.

- What is the most likely diagnosis?

Figure 7.3: A flawed case description.

The many flaws in this case may be obvious: the age of the patient is not stated, the exact location of the pain is not given, "a dubious pressure tenderness" is vague, etc. Putting in some extra lines to make the case clearer may cost a little extra reading time for the examinee, but will probably decrease total testing time, because the student does not need to doubt about the meaning of the case. An example of a better case presentation is presented in figure 7.4.

You are a general practitioner. Mrs Whiteless (28 years old) consults you in your practice because of complaints of pain in the right part of the lower abdomen. It has been present for about 8 days now. In the beginning she barely felt it, but it has increased in the last four days. The pain comes in cramps, occurring about 5 - 10 times per day each attack lasting for about 15-30 minutes. During these cramps she feels the urge to lie down and keep as calm as possible.
She tells you further that her period is regular, but that her last menstruation was 1 week later than normal. On inspection of the abdomen you see no abnormalities except for an appendectomy scar. The peristalsis is normal. On palpation there is a pressure tenderness in the lower right abdomen and you feel a mass barely touching your fingers so you cannot determine its size. Percussion is slightly painful, but reveals no abnormal sounds. During gynaecological examination you see a drop of blood on the portio, bimanual palpation is painful at the right side (left side is normal). Further gynaecological examination yields no abnormal findings..

- What is the most likely diagnosis?

Figure 7.4: A better case presentation.

3 Provide sufficient realistic clinical information.

Not only should all information that is needed be present in the case, its presentation should be explicit: "normal routine examination" is often inadequate as a description; mostly the actions should be mentioned separately. As mentioned earlier, the disadvantage of the extra reading time is outweighed by the time that would be used by the candidate doubting what is meant by the description.

4 Provide sufficient realistic contextual information.

To solve a patient case, contextual information is very relevant. Information about previous consultations, the frequency of consultations, etc., can be relevant in order for the examinee to come to a judgement of probabilities on a certain diagnosis. This contextual information may also serve to present non-verbal behaviour, such as "the way the patient presents his complaint gives you the impression that he exaggerates". It is advisable to report some contextual information as a standard procedure (e.g. gender, age, frequency of previous visit, profession, etc.).

5 Provide sufficient negative information.

All information that is necessary to answer the question must be in the case. So not only the type of action taken (e.g. liver palpation) and the results (1-2 cm. below lower costa, sharp edge), but also the realistic actions not taken or actions that did not show abnormalities should be presented (e.g. "There is no rebound tenderness."). It is dangerous to assume that every candidate will automatically know whether an action not reported is either not taken or did not yield any abnormal findings.

6 *Provide information that is not pre-interpreted ('raw')*

For efficiency reasons often the information is presented in an pre-interpreted style. This means that some of the interpretation was already done by the case author. In real practice, however, interpretation of findings into larger meaningful 'chunks' is part of the problem-solving process. So instead of reporting a pulse deficit or liver enlargement, it is better to report the pulse rate and the heart rate or the size of the liver to let the examinee draw his/her own conclusions, i.e., to interpret the data for themselves.

THE PROBLEM

7 *Link the problems directly to the case*

The case should not be presented as an illustration followed by a question that asks for general knowledge. It is essential that case and questions form an inseparable unit. A minimum criterion can therefore be that a candidate should not be able to answer the question without having read the case. A question with that particular flaw is for example shown in figure 7.5.

Mrs Smith suffers from attacks of heavy cramping pain in the upper abdomen. When such an attack occurs, she has the urgent need to stand up and walk around. The attack occur 5-6 times per day and have been present for 4 days now. An attack lasts for about 4 minutes. She has noticed that most of the attacks occurred after she had eaten meals with fatty food. She has taken her temperature, which was 37.3 °C. On physical examination you notice a slight pressure tenderness in the upper right quadrant of the abdomen. Inspection, auscultation and percussion of the abdomen reveal no abnormal findings.

- Which of the alternatives lies most closely to the specificity of an ultrasound scan for detecting gallstones?

a 20%
b 40%
c 60%
d 80%

Figure 7.5: The question is not related to the case.

The next example (figure 7.6) appears to solve the problem,....

Mrs Smith suffers from attack of heavy cramping pain in the upper abdomen. When such an attack occurs, she has the urgent need to stand up and walk around. The attacks occur 5-6 times per day and have been present for 4 days now. An attack lasts for about 4 minutes. She has noticed that most of the attacks occurred after she had eaten meals with fatty food. She has taken her temperature, which was 37.3 °C. On physical examination you notice a slight pressure tenderness in the upper right quadrant of the abdomen. Inspection, auscultation and percussion of the abdomen reveal no abnormal findings.

- In this case the positive predictive value of an ultrasound scan for detecting gallstones lies closest to:

a 20%
b 40%
c 60%
d 80%

Figure 7.6: The question does not ask for an essential decision.

....but although the question cannot be answered without the case, it is still not aimed at decisions that are essential in practice. In the determination whether or not to perform an ultrasound scan of the gallbladder, not only the positive predictive value, but also the negative predictive value, the costs, etc., have to be included in the decision. Therefore, the most realistic way of asking this question is simply to ask whether or not an ultrasound scan of the gallbladder is indicated in this particular case (considering the signs, symptoms, probability of the diagnosis, positive predictive value, negative predictive value, costs, invasiveness, etc)

8 *Avoid problems or possibilities that are not present in real practice.*

Assessment is of course always an abstraction of real life. This, however, does not necessarily mean that real life is more difficult than an examination. Often possibilities exist in real life that are not present in a test. So, for example, do not focus in a case on aspects that can be easily looked up in real life (like normal lab values, or dosages of drugs). If critical issues of a case are asked, quick referencing does not reveal the answer. In contrast, if the answer can be easily and quickly found in a reference book, the question is probably not a key decision. So allowing the students to use reference books (within the time constraints of the examination) can help you to focus your cases on essentials. Another approach would be to include reference material such as normal laboratory values within the question or case scenario.
Distracting information may be present in real life. It is advisable to incorporate this in the case. It must be done, though, in such a manner that the student can extract some clues as to whether this information is relevant or not: for example, by describing the tone in which a patient says something or describing other non-verbal communication. While this is somewhat artificial in that the examinee cannot actually hear the tone of the voice of the patient, the case author needs to work within the limitations of the median of the case and question format which is often written text. Be cautious when introducing distracting information. This information should not be so confusing as to create an ambiguous case or question.

9 *Focus on essential problems only*

Most of the problems and decisions involved in solving a case are quite straightforward and follow automatically from others. When for example the decision to admit a patient as an emergency case has been taken, the subsequent problem whether or not to call for an ambulance is quite trivial. Some decisions, however, are essential for the case, though the distinction between essential and non-essential problems can be difficult.
Some prerequisites for a problem to be essential are:

- The problem must be based on combining the different information parts of the case.

Mr Johnson (67 years old) sees you in your practice because he has had a sharp pain in his right shoulder for three weeks. At first he thought that the pain would go away spontaneously, but since it has not, he wants you to take a look. The pain is not related to any movements of the shoulder. It irradiates to his back, mainly in the scapular region. Further history taking reveals that he has never been seriously ill. He has been a heavy smoker for almost 45 years. On physical examination you notice a myosis of the right pupil and a ptosis of the eyelid.
What is this combination of the last two symptoms called?

Figure 7.7: The question does not require the combination of several aspects in the case.

The case in figure 7.7 provides an example where this prerequisite is not met, because the essential part is not to recognise the Horner's syndrome, but for example to recognise that this is a serious disorder or even that this is probably a Pancoast tumour and should be dealt with accordingly. If only a small part of the information of the case is needed to solve the problem, it is probably not an essential one.

- An incorrect decision must lead to an incorrect management of the case.

You make a house call on Mrs Van Doorn (65 years old). She suddenly felt a fierce piercing pain in her abdomen about three hours ago. Any movement she makes increases the pain, so she keeps as still as possible. Abdominal examination is barely possible. On auscultation you hear no bowel sounds. Palpation is barely possible, not only because of the tenderness, but also because her abdominal muscles are contracted and cannot be relaxed. Percussion is not possible. She has a temperature of 39.5 °C. She has had abdominal pains before with periods of fever, but these were different in nature (less sharp) and subsided after a couple of days.

What is the most probable diagnosis?

A diverticulitis
B appendicitis
C perforated peptic ulcer

Figure 7.8: An incorrect answer would not lead to an incorrect patient management.

The exact diagnosis is not essential for effective management in the case in figure 7.8. Whether this is diagnosis A, B or C does not influence the management in this case. In each of the situations the patient would have to be admitted to emergency as quickly as possible.

- Do colleagues agree with your selection?

It is advisable to present the case to colleagues and ask them to comment on your selection of problems. Different experts often share the same opinion on what essential problems are. In the case of dissension it is recommended to rethink your selection process (Bordage, Brailovsky, Carretier, & Page, 1995).

10 Limit the number of problems to be asked

If a key feature of the case has been addressed by a question, it should not be pursued any further. It is more efficient to go on to the next (essential/key) problem or even to go on to the next case. The more different cases that can be presented and the more different questions that can be asked (addressing key elements in the problem-solving process), the higher the generalisability will be.

THE QUESTION

11 Phrase the questions as clearly as possible.

The rules that apply to normal test questions naturally apply to questions in case-based testing. They should be checked for common flaws. Flaws can work two ways: either they provide cues that might lead an incompetent examinee to produce the correct answer (i.e. lead to false positive results) or they consist of bad phrasing leading the competent candidate

astray (i.e. lead to false negative results). Examples of the former are: obvious differences in length of the options in a multiple choice, cueing words like 'can', 'always', 'never', etc., in multiple choice or true-false questions, unrestricted open-ended questions (which lead the candidate to write as much as he/she can). Examples of the latter are the use of ambiguous terms ('sometimes', 'often', etc.) or insufficient indication of the level of detail in an open-ended question. A more detailed description of these flaws goes beyond the scope of this chapter, but good examples are available in literature (e.g. WWW:\\NBME.ORG.COM).

12 Focus the question on the specific aspects of the problem

The problem in a case must be defined and presented as clearly as possible. It must be made clear which of the aspects must be considered by the candidates when making a decision. An example of a question where this is NOT the case is presented in figure 7.9.

John Provis (4 years old) has had a fever since yesterday (in the evening 38.9°C). His parents are worried, especially because he also has a severe cough. The general practitioner examines the child and concludes that no serious illness is present, but that it is a common cold. She advises the parents to give John an Aminocetophin® in the evening to lower the fever and to keep him inside for 3 days.

- Is this correct advice?

a Yes
b No

Figure 7.9: The question is not sufficiently focused.

The question cannot be answered because it is not clear what the aim of the question is. One could very well argue that the advice is incorrect since the GP should have informed the parents about the innocence of the disorder, but on the other hand the advice given is not incorrect. The problem can be avoided for example by including the considerations in the case, as is illustrated in figure 7.10.

John Provis (4 years old) has had a fever since yesterday (in the evening 38.9°C). His parents are worried, especially because he also has a severe cough. The general practitioner examines the child and concludes that it is nothing but a common cold. She asks herself whether the advice to give an Aminocetophin® in the evening and the advice to keep John inside for 2-3 days would be correct.

- The advice concerning the Aminocetophin® is correct. true/false
- The advice to keep John inside for 2-3 days is correct true/false

---

Figure 7.10: Using more and simpler questions can lead to a better focus on the essentials.

Breaking up the different elements in the question will often lessen ambiguity of combined elements.

13 Ensure that the answer is defensibly correct, distractors defensibly false.
This suggestion may look obvious but this is not always the case. Often questions are phrased in such a way that some of the boundaries between correct and incorrect answer are not made explicit. The example in figure 7.11 may clarify this.

Mrs White (45 years old) sees you in your GP practice. She tells you that she has this pain in her abdomen. The pain has been going on since one week and it seems to be getting worse.

- What questions do you ask?

Figure 7.11: No defensibly incorrect answer exists.

It may be clear than any answer is defendable. Every decision in practice is a trade-off in which the advantages, probabilities and disadvantages have to be weighed against each other. In this case, giving some more information about the complaint, and then focussing the question to parts of the normal history taking (family history, digestive tract, urogenital tract, etc.) and asking for the selection that is most likely to yield relevant information might resolve the problem.

14 Let the content of the question determine the format
It seems tempting to always use a certain question format and adapt all questions to that particular format, since this is most straightforward in psychometric sense. However, in terms of authenticity this would not be optimal. Decisions in real life often involve the selection between a limited number of alternatives (like the decision either to say "I do" at a wedding or not) and sometimes a nearly limitless number of options. To present questions with the same number of realistic options as in real practice may therefore be optimal in preparing the student to his/her future task.
A 'one format fits all' approach can be adopted for logistical efficiency or psychometric reasons, but cannot be based on the intrinsic superiority of a particular question format. Even open-ended questions can be lacking in authenticity. A defendant, for example, who is instructed by his lawyer in an open-ended way, will be ill-prepared when faced with "yes" or "no" questions in court. Similarly, a physician that is only used to one type of question for all problems may be sub-optimally prepared for future decisions. To determine what question type to use, the case author should try to determine the number of realistic alternatives that exist in real practice. If this number is very large, an open-ended question appears to be most appropriate or an extended menu format where the student is allowed to select several alternatives. If the number of realistic alternatives is limited, a multiple choice question is preferable. In the next case, in figure 7.12, the number of realistic options may be rather large, since the essence appears to be the recognition of the pattern.

Mr. Brown consults you about his knee. It is swollen and red and it hurts. He has had this for 2 days now and it has gradually become worse. He has not had this before. When you ask him for other complaints, he tells you that he has had a red spot on his left upper leg, about 5-10 cm in diameter. This spot then disappeared and reappeared on the other leg. Two weeks before the spots appeared he spent his holidays in the woods hiking. He had caught a common cold there, which lasted for only 3 days.

- What is the most appropriate diagnosis?

Figure 7.12: An example of a case that would need an open-ended question.

In the following (figure 7.13), however, the essence of the case is not necessarily to generate a differential diagnosis, but to discriminate between the probabilities of the individual diagnosis.

Mr. Thomas visits you in your practice because he has had chest pain yesterday. He was mowing the lawn, which he had not done for a long time. It was hard work cutting the grass. Suddenly he felt a stabbing crunching pain in his chest on the left side. The pain made it impossible to continue working, so he sat down. After 5-10 minutes the pain decreased, but he had the feeling that the pain was not totally gone until an hour after its onset. A very faint pain remained during that hour. He has never had this before and he is very worried. On cardiac and pulmonary examination (inspection, palpation, percussion and auscultation) you find no abnormalities. Pulse is 80/min (regular) and the blood pressure is 150/80 mmHg.

- Which of the following is the most probable diagnosis?

a myocardial infarction
b instable angina
c stable angina
d pulmonary embolism
e pneumothorax
f hyperventilation
g intercostal neuralgia

Figure 7.13: An example of a case where a multiple choice question would be better.

15 Have your material reviewed by others

Writing test material is not easy and the quality of the material can easily be negatively influenced by 'blind spots' of the author. In the field of scientific research and practical medicine these blind spots are already widely acknowledged and have led to quality control procedures such as cross checking and review. A similar approach should be adopted in the production of high quality test material: careful review increases the quality of the material. In figure 7.14 a summary of the strategies and tips with respect to case-based items is given.

1	Use the representation of real patients.
2	Ensure that the description of the information is as clear as possible.
3	Provide sufficient realistic clinical and contextual information.
4	Provide sufficient negative information.
5	Provide information that is not pre-interpreted ('raw')
6	Link the problems directly to the case
7	Avoid problems or possibilities that do not exist in real practice.
8	Focus on essential problems only
9	Limit the number of problems to be asked
10	Phrase the questions as clearly as possible.
11	Focus the question on the specific problem
12	Ensure that the answer is defensibly correct, distractors defensibly false.
13	Let the content of the question determine the format
14	Have your material reviewed by others

Figure 7.14: Tips and strategies for writing case-based test material.

EPILOGUE

Constructing short cases for examinations is not easy. In our experience an average time of 2-3 hours per case can be considered normal. In addition, it has proven to be ill-advised to make cases without consulting others. Nobody is immune to mistakes, blind spots, etc., (including the authors of this chapter). It is widely accepted that manuscripts for articles are shown to colleagues before sending them to a journal. It is widely acceptable to confer with colleagues when in doubt about a certain clinical strategy. Since case writing is a difficult task, it should be acceptable to show your test material to others and ask them for comments and criticism.

The nature of the cases and the selection of the essential problems varies of course with the educational context of the test. Tests in undergraduate courses will lead to the selection of different key-features than tests in post-graduate education or even in continuing medical education. This will be determined by the expected prior knowledge and experience of the candidates and the specific course goals (e.g. basic sciences versus clinical sciences).

The strategies and pitfalls of this chapter do not apply to clinical or patient cases only. Short descriptions of physiological or anatomical problems can be used also. Furthermore, clinical cases can be used to ask for basic science problems (Des Marchais, Jean, & Nu Viet Vu, 1993; Jean, Schuwirth, Van Santen, & Van der Vleuten, (under editorial review); Schuwirth, Jean, Van der Vleuten, & Van Santen, 1993).

A final piece of advice would be the suggestion to the reader to look for possibilities for co-operation with other departments or faculties. Since the production of high quality test material can be tedious and expensive, co-operation can often lead to a win-win situation. In any case, the use of the short-case approach to measuring problem-solving ability appears to be both viable and desirable.

REFERENCES

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Bordage, G. (1987). An alternative approach to PMP's: the "key-features" concept. In I. R. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence, Proceedings of the second Ottawa conference* (pp. 59-75). Montreal: Can-Heal Publications Inc.

Bordage, G., Brailovsky, C., Carretier, H., & Page, G. (1995). Content validation of key features on a national examination of clinical decision-making skills. *Academic Medicine, 70*(4), 276-81.

Des Marchais, J. E., Jean, P., & Nu Viet Vu, C. (1993). An attempt at measuring student ability to analyze problems in the Sherbrooke problem-based curriculum: A preliminary study. In P. A. J. Bouhuijs, H. G. Schmidt, & H. J. M. Van Berkel (Eds.), *Problem-Based Learning as an Educational Strategy* (Vol. 1, pp. 239-48). Maastricht: Network of Community-Oriented Educational Institutions for Health Services.

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Graaf, E. d. (1988). Simulation of initial medical problem-solving: A test for the assessment of medical problem-solving. *Medical Teacher, 10*(1), 49-55.

Jean, P., Schuwirth, L. W. T., Van Santen, M., & Van der Vleuten, C. P. M. (under editorial review). Do problem analysis questions (PAQs) and true/false questions (TFQs) measure different skills. .

Schuwirth, L. W. T., Jean, P., Van der Vleuten, C. P. M., & Van Santen, M. (1993). Problem-Analysis Questions, een korte casusvorm voor het preklinische domein [Problem Analysis Questions, a short case-based testformat for the pre-clinical domain]. In E. Houtkoop, J. Pols, M. C. Pollemans, A. J. J. A. Scherpbier, & G. M. Verwijnen (Eds.), *Proceedings van het Derde Gezond Onderwijs Congres* (pp. 104 - 11). 's Gravenhage, Nederland: Haagse Hogeschool.

Schuwirth, L. W. T., Van der Vleuten, C. P. M., De Kock, C. A., Peperkamp, A. G. W., & Donkers, H. H. L. M. (1996). Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher, 18*(4), 295 - 300.

Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in Assessing Clinical Competence* (pp. 13 - 45). Montreal: Can-Heal publications.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van der Vleuten, C. P. M., Newble, D. I., Case, S. M., Holsgrove, G., McCann, B., McGrae, C., & Saunders, N. (1994). Methods of Assessment in Certification. In D. I. Newble, B. Jolly, & R. Wakeford (Eds.), *The certification and recertification of doctors, issues in the assessment of clinical competence* (pp. 105 - 25). Cambridge: Cambridge University Press.

DISCUSSION

The purpose of this dissertation has been to assess the validity of Computerised Case-based Testing (CCT). Validity has been approached from the broadest possible perspective to obtain an optimal impression. Two main elements of study have been to investigate what cognitive tasks CCT poses to the candidate and what the enabling and boundary conditions are for these tasks. These elements were translated into 9 research questions or topics. In the majority of the cases these are addressed using the results of several of the chapters in this dissertation.

RESEARCH QUESTION 1: WHAT DO EXPERTS AND STUDENTS JUDGE THAT THE VALIDITY OF CCT IS?

Experts and students were asked to judge the validity of CCT. On average both groups gave a positive judgement on its validity. Experts judged the case presentations to be authentic and the decisions addressed in the questions to be realistic. A full consensus on all cases or questions could not be reached though. Of some cases, the majority of the experts expressed concerns about the validity. It did not become clear from this study where these concerns originated. For this a more detailed study should be undertaken. The purpose of such a study would be to identify common elements, medical and educational, in cases on which experts disagree about their validity.
Consensus was much higher in the student groups. The vast majority indicated that the tasks set by CCT are similar to those encountered when seeing patients in practice. They also indicated that a CCT test sets a different task from a more knowledge-oriented test. These results are reasonably positive, but it must be kept in mind that experts' judgements can only tell what experts think which cognitive tasks are set by CCT, and that the value of introspection of the students is limited (Ericsson & Simon, 1993; Nisbett & Wilson, 1977).

RESEARCH QUESTION 2: DOES CCT ELICIT OTHER COGNITIVE OPERATIONS THAN FACTUAL KNOWLEDGE QUESTIONS?

For this question, a cognitive psychological approach was used. A think aloud protocol methodology was applied to gain better insight into the difference in cognitive operations when solving CCT cases compared to factual knowledge questions. The results of this study are striking. Differences between CCT and factual knowledge questions were large and highly congruent with the theoretical expectations. It was concluded that decisions presented within an authentic and meaningful context require the candidate to read and process the case information and to relate this to prior knowledge to be able to weigh up the probabilities of the different possibilities. In factual knowledge questions, less information processing is required, merely a simple retrieval of declarative knowledge is needed.

Although the study of these qualitative aspects of cases and questions yielded an insight into the nature of the thinking processes, it was also important to show that these have an impact on the test scores of CCT.

RESEARCH QUESTION 3: DOES PERFORMANCE ON CCT INCREASE WITH INCREASING EXPERTISE?

Scores and percentages of passing students were compared between an entrance level test and an exit level test in chapter four. The increases were considerable: mean scores increased by about 17% whereas the percentages of passing students increased by about 78%. The magnitude of the effect size strongly suggests that the tasks of CCT are similar to those in practice. This conclusion was underscored even further by the fact that experts scored significantly higher than students. This would lead to the conclusion that no intermediate effect occurs (Schmidt, Boshuizen, & Hobus, 1988). It must be said, however, that in this study residents were not included, and it is possible that they represent the intermediates. In two, for example, the mean scores of the residents were found to be slightly higher than those of the general practitioners. On the other hand, this difference was not statistically significant. Therefore, if an intermediate effect occurred in CCT, it is probably not as prominent as in other tests (Schmidt et al., 1988). Clearly, while further study of this effect is needed, in general scores on CCT do increase with increasing expertise.

RESEARCH QUESTION 4: DOES CCT YIELD INFORMATION ABOUT MEDICAL COMPETENCE OF STUDENTS THAT IS NOT PROVIDED BY OTHER MEASURES OF COMPETENCE?

Results of a comparison between CCT and other test formats are more difficult to interpret. In general low (true) correlations were found suggesting that the competence measured by CCT differs from that in other tests. Yet, the correlational pattern was not fully in line with the expectations. In particular the judgements of the general practitioners with respect to the application of knowledge in practice did not show an increased correlation. Two explanations can be found for this. First, the judgements of the general practitioners on the different aspects of the clerks (knowledge, skills, application of knowledge and attitudes) intercorrelated highly. This suggests that a halo effect is present, which would suppress any possible pattern of correlations within these judgements. A second possible explanation is the inadequacy of correlational studies for validity assessment. In the literature these methodologies often led to correlations ranging from moderate (and therefore uninterpretable) to high (arguing against construct validity) (Norman, Van der Vleuten, & De Graaff, 1991).

RESEARCH QUESTION 5: IS CCT ABLE TO DETECT DIFFERENCES BETWEEN STUDENTS OF DIFFERENT MEDICAL FACULTIES USING DIFFERENT APPROACHES TO PROBLEM SOLVING, BETTER THAN OTHER TESTS?

External evidence for the validity of CCT was provided in chapter six in which the results of the comparison between the medical faculties of Groningen and Maastricht were described. Many such comparisons have not shown any differences between students of problem-based learning (PBL) faculties and non-PBL faculties (Albanese & Mitchell, 1993; Albano et al., 1996; Vernon & Blake, 1993; Verwijnen, Van der Vleuten, & Imbos, 1987). Although the original assumption with respect to PBL that it would lead to better generic problem solvers (Barrows, 1984) may have been proven false (Elstein, Shulmann, & Sprafka, 1978), it would be still reasonable to expect some differences. First, learning within a relevant and meaningful context (as is done in PBL) enables a better storage and retrieval of relevant knowledge (Chi, Glaser, & Rees, 1982; Glaser & Chi, 1988). Second, students have solved many cases during their study which might lead to recognition of problems (and solutions) in more instances (Norman, 1988), or to a higher level of integration of knowledge which would not only lead to faster but also to better processing of information by a more efficient use of the 'cognitive room' (Chi et al., 1982; Schmidt & Boshuizen, 1993; Schmidt et al., 1996). A difference in mean scores and response times should therefore be found if a test were used that asks questions within such a relevant practical context. Indeed PBL students show a superiority in making the correct decisions in CCT, increasing with the year class. At graduation level the effect size is considerable. Also in the PBL group a difference was found in response times between the correctly solved cases and the incorrectly solved cases in every year group. In the non-PBL group no such difference was found. This is in line with the expectations, PBL students are not only more proficient in solving CCT cases, they appear to use a more integrated knowledge approach when they recognise the case, whereas in other instances they fall back on "lower level" problem-solving approaches. Because such differences were never demonstrated with factual knowledge-based tests, the results of our study can be seen as additional evidence for the validity of CCT.

RESEARCH QUESTION 6: HOW IS CCT EMBEDDED IN THE RESEARCH ON ASSESSMENT OF MEDICAL PROBLEM SOLVING?

The boundary and enabling conditions with respect to validity constitute a further important element of this dissertation.
A first aspect in this is the considerable history of research in the areas of medical problem solving and cognitive research on the nature of expertise development. The surprising and often counter-intuitive findings, as described in the introduction, and in chapters one and two of this dissertation have markedly influenced the shaping of CCT: short authentic cases with questions explicitly aimed at essential decisions. CCT must be seen as the result of and advancement on this research.

RESEARCH QUESTION 7: WHAT IS THE INFLUENCE OF QUESTION FORMATS ON THE VALIDITY OF CCT?

The question whether the use of multiple choice questions has a negative impact on the validity of CCT cannot be answered briefly. In chapter two it was found that in multiple choice questions cueing occurs, sometimes favouring the multiple choice scores (positive cueing), sometimes favouring the open-ended questions (negative cueing). The total percentage in which cueing occurs was found to be considerable (about 20%), but true correlations between the scores on the open-ended questions and the multiple choice questions were 1.00 in almost all year groups. The amount of cueing varied with the difficulty of the question: difficult cases generated more positive cueing, easy cases generated more negative cueing. This finding, in combination with the fact that the content of the question also influences the amount of cueing (Swanson, Norcini, & Grosso, 1987), would suggest that the optimal and most valid approach would be to select the question format according to the content of the question. This is one of the suggestions used in chapter seven. To do so, though, implies that in CCT it should be possible to ask and score (short answer) open-ended questions. Since computer scoring of open-ended questions is still problematic a computerised long menu question was developed. In chapter three it was demonstrated that such an approach is congruent with hand-scored open ended questions. The use of computerised long menu questions would therefore not negatively influence the validity of CCT. The use of different question formats within one test is therefore recommendable on grounds of authenticity. Potential psychometric concerns in this respect were investigated in chapter four. The results, however, indicate that no additional error-variance was introduced which can be based on the use of different question formats. This was, however, based on a relatively small sample of examinees and therefore warrants further study in the future.

RESEARCH QUESTION 8: WHAT IS THE INFLUENCE OF THE BLUEPRINT ON THE VALIDITY OF CCT?

A similar conclusion can be drawn with respect to the blueprint. The fact that different areas (represented by the different categories of the blueprint) are combined to one final score has no negative psychometric consequences, whereas from a direct validity viewpoint it is highly recommendable.

RESEARCH QUESTION 9: WHAT ARE THE GUIDELINES FOR DESIGNING VALID CASES AND QUESTIONS?

The experience gathered during the construction of many CCT cases and the research on the validity of the method have led to the definition of some explicit strategies, tips and pitfalls with respect to the writing of CCT cases. The description of these in chapter seven must be seen as the result of numerous review panel discussions, expert and student suggestions, quality control cycles and revisions of material.
In general the validity of CCT is concluded to be high, and no major boundary conditions were found. Enabling conditions are mainly the careful case construction processes with

thorough quality control cycles to produce authentic test material which is accompanied by a continuous study and implementation of the research results in the area of competence assessment. These aspects must be seen as crucial validity determinators of a test.

OVERALL CONCLUSIONS REGARDING CCT

The question remains, though, whether a valid CCT test is a sufficient solution to the problem with the assessment system described in the introduction (Swanson, 1988). To determine whether a test is useful within an institute information concerning parameters other than validity alone must be provided. In his model for the usefulness of an examination Van der Vleuten suggests representing the usefulness as a combination of five different parameters: validity, reliability, educational impact, costs and acceptability (Van der Vleuten, 1996). He further suggests perceiving the relation between the parameters to be multiplicative, because if one of them is nil, the usefulness of the examination would be nil. In formula this could be represented as:

$$U = V_{w_V} \times R_{w_R} \times E_{w_E} \times C_{w_C} \times A_{w_A}$$

The subscripts 'W' indicate that to all parameters a specific weighting must be attributed based on the specific demands of the institute in which the examination is to be used. This implies that assessment is always a compromise between the parameters.
The main focus of this dissertation has been on the validity of CCT. It can be concluded that the evidence for this validity is abundant and that CCT is a valid method for the assessment of problem-solving ability.
In addition, however, impressions about reliability, educational impact, costs and acceptability have been acquired which enable the setting of directions for future developments of CCT.
In two chapters reliability estimates of CCT could be given. In chapter two the reliability of the first set of 30 cases (the second set was in fact a repetition of the first) was .64, and in chapter four a reliability of .67 was found for a set of 40 cases (equalling about 1.5 hours of testing time). These values are lower than the .80 that is often referred to as the benchmark in literature. However, the level on which decisions about the students must be made by CCT is only to discriminate between students passing and failing. The percentage of students on which a pass-fail decision could not be done with sufficient certainty (with the current CCT test and the current cut-off scores) is about 10%.
To increase the reliability of the pass-fail decisions sheer prolongation of the test would not be the best solution. More efficient and more economical would be the use of sequential or adaptive testing. In the former case a fixed number of cases would be presented to all students, and only those students whose scores are too close to the cut-off score would receive an additional set of cases, the results of which are added to those of the first set. The others are excused. Adaptive testing requires a so-called calibrated item bank, in which the level of difficulty of all items is known. In this procedure the student solves the first case and based on whether the solution is correct or incorrect, a more difficult or an easier case is selected from the bank and presented to the student. Through this approach an exact estimation of the level of competence of a student can be reached with a limited number of cases. So, in sequential testing the length of the test may vary according to the scores of the student, whereas in

adaptive testing the content of the test may vary according to the scores of the student. To enable adaptive testing first considerable psychometric research has to be conducted to calibrate the item bank, but sequential testing can be implemented in a reasonably short time.
With respect to the educational impact of CCT only non-systematic debriefings have been conducted. They suggest that students perceive an active involvement and participation during their clerkship as an adequate preparation for the examination. At the Maastricht medical school educational impact of a clerkship examination is considered an important aspect. This implies that in the further development of CCT the building of sufficiently large item banks has priority for two reasons:

- If the bank is large enough, a student strategy aimed at memorising the cases and questions would not be efficient.

and

- If the bank is large enough it would allow for the possibility that every student has a limited number of formative test rounds. Because test administration is highly automated, this would not be logistically difficult. In that way students have ample opportunity to adjust their study behaviour during the clerkship.

Expansion of CCT to other disciplines would also be favourable. At this moment every CCT test is tightly interconnected with the specific clerkship it is produced for. The disadvantage of this is that student could still 'polish up' for the test occasion and could allow themselves to forget what they have learned after they have passed the examination. If, however, sufficient cases are produced in sufficient different clinical disciplines, a longitudinal approach could be adopted, in which students are tested periodically in a sort of case-based progress test (Van der Vleuten, Verwijnen, & Wijnen, 1996). This would probably encourage a more continuous learning with more in-built rehearsal (Van Til, 1998).
More detailed studies on the real educational impact of CCT as compared to other examinations are currently being conducted.
The costs involved in CCT are high. Production of high quality test material is tedious and expensive. On average 2-3 person-hours per case have been spent by authors, review committee member and experts in committees. Careful training of authors and review committee members might diminish this, but the investment per case is still considerable. The most obvious approach to reduce costs would lie in co-operation with other medical schools. Because all medical schools in the Netherlands have a curriculum that consists of four pre-clinical years followed by two clinical or clerkship years, CCT would be applicable at all medical schools. Co-operation would therefore lead to a reduction of the costs per faculty and an increase of the application of CCT. The first steps into the direction of inter-institutional co-operation have been made.
Acceptability can play an important role in assessment methods. Students and teachers must perceive it as a valuable contribution to the curriculum. In general student opinions on CCT are positive. In a debriefing using a questionnaire (n=128) the majority of the students carried a positive opinion on the testing format (80.2%) and the use of computers (78.4%). In general the acceptability among teachers is high, though the production of case material is sometimes considered tedious and difficult.
In general CCT appears to be an adequate solution for the lacunae in the clerkship assessment system as described in the introduction. The least positive aspect is the relatively high costs involved in producing test material. The incipient co-operation with other medical schools will reduce these costs significantly. Reliability, acceptability and educational impact seem to

be sufficient. Validity of CCT as a measure of medical problem-solving ability has been clearly demonstrated.

REFERENCES

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues. *Academic Medicine, 68*(1), 52-81.

Albano, M. G., Cavallo, F., Hoogenboom, R., Magni, F., Majoor, G., Manenti, F., Schuwirth, L., Stiegler, I., & Vleuten, C. P. M. v. d. (1996). An international comparison of knowledge levels of medical students: the Maastricht Progress Test. *Medical Education, 30*, 239-245.

Barrows, H. S. (1984). A specific, problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In H. G. Schmidt & M. L. De Volder (Eds.), *Tutorials in problem-based learning* (pp. 16 - 32). Assen: Van Gorcum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7 - 76). Hillsdale NJ: Lawrence Erlbaum Associates.

Elstein, A. S., Shulmann, L. S., & Sprafka, S. A. (1978). *Medical problem-solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis.* Cambridge Massachusetts: Massachusetts Institute of Technology.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv - xxviii). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*(3), 231 - 59.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 86.

Norman, G. R., Van der Vleuten, C. P., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education, 25*(2), 119-26.

Schmidt, H. G., & Boshuizen, H. P. (1993). On acquiring expertise in medicine. Special Issue: European educational psychology. *Educational Psychology Review, 5*(3), 205-221.

Schmidt, H. G., Boshuizen, H. P. A., & Hobus, P. P. M. (1988). Transitory stages in the development of medical expertise: The "intermediate effect" in clinical case representation studies, *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 139 - 45). Montreal, Canada: Lawrence Erlbaum Associates.

Schmidt, H. G., Machiels-Bongaerts, M., Hermans, H., ten Cate, T. J., Venekamp, R., & Boshuizen, H. P. A. (1996). The development of diagnostic competence: Comparison of a Problem-Based, and integrated, and a conventional medical curriculum. *Academic Medicine, 71*(6), 658-664.

Swanson, D. B. (1988). *Review of the assessment system used by the University of Limburg medical school* (PES nr 88 - 22). Maastricht: University of Maastricht.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220 - 46.

Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education, 1*(1), 41-67.

Van der Vleuten, C. P. M., Verwijnen, G. M., & Wijnen, W. H. F. W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher, 18*(2), 103 - 10.

Van Til, C. (1998). *Voortgang in Voortgangstoetsing. [Progress in Progess Testing.]*. , University of Maastricht, Maastricht.

Vernon, D. T., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68*(7), 550-63.

Verwijnen, M., Van der Vleuten, C. P. M., & Imbos, T. (1987). A comparison of an innovative medical school with traditional schools: an analysis in the cognitive domain. In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in Medical Education: An Evaluation of Its Present Status* (pp. 40-49). New York: Springer Publishing Company.

SUMMARY

When in 1974 the Maastricht University started its Problem-Based Learning curriculum, in its first faculty, the school of medicine, it was essential that a high quality assessment system was established which, on the one hand, would fit the educational approach and, on the other hand, would be useful in ensuring the quality of the graduates. A number of criteria were defined that each assessment instrument should meet: for example, congruence between assessment and education, separation of teacher and assessor role, comprehensive assessment that does not focus on certain aspects of medical competence only. In the past two decades several high-quality assessment instruments have been developed assessing mainly knowledge and skills. In the field of assessing problem-solving ability, however, a lacuna still existed. The examinations in this field, particularly those used in the clerkships varied considerably, ranging from structured orals with so-called long cases to written factual knowledge tests.
In response to this lacuna the development of Computerised Case-based Testing (CCT) was started. The idea to seek a CCT approach was founded on the many developments in psychometrics and cognitive psychology. In both disciplines it was found that problem-solving ability is not a generic skill that once mastered can be applied to any given situation, but that it is highly domain specific and idiosyncratic. This would imply that using small numbers of long and branched cases, in which the process of problem solving is measured, is not an optimal means to assess problem solving. An approach using large numbers of short cases with only a limited number of questions would be more favourable. The questions should not be aimed at general knowledge, but be focused on essential decisions. This approach is called the 'key-feature' approach. Logistically, it would be best to use computers for the administration of these examinations.
Although this approach may seem fruitful, a careful validation has not yet been performed. The main goal of this dissertation is therefore to provide evidence for the validity of CCT as a measure of problem-solving ability.

Chapter one describes the CCT method in more detail. First a description is given of the cases that are used in CCT. All case material is based on real patients, and the cases are written by the physician. This is done to optimise the authenticity of the cases. In addition, multimedia (such as pictures, sounds or moving images) are used when relevant. Cases are kept short to enable the presentation of many different topics in a test, and thus broaden the sample of topics covered.
A careful quality control system was installed to ensure the quality of the bank of cases. This entails extensive training of the authors before they start writing cases, a careful review of all case material on wording, content and relevance, and a validation of the answer keys and the authenticity by an expert panel.

Different question formats are used in CCT. The format of the question is determined by the number of realistic alternatives in real practice. Therefore, both multiple choice type and open-ended questions are used.
The main reasons to use computers in test administration are their logistical advantages, the possibility of including multimedia, and the possibility of revealing case information stepwise. Furthermore, by using a computerised database of cases, a (stratified) random test can be drawn for each student separately. A future advantage will be the use of sequential and/or adaptive testing.

A possible drawback in test design is the so-called cueing effect, which is studied in *chapter two*. This effect means that students can answer a multiple choice answer correctly by recognising the correct answer. In open-ended questions spontaneous generation is needed to produce the correct answer. It is because of this effect that multiple choice questions are assumed to be unsuitable for the testing of problem-solving ability. If this were true, it would pose a serious threat to the validity of CCT. The literature on cueing, though, is not conclusive. Results indicate that a cueing effect may exist, but that its influence on the competence being measured is negligible. By using the exactly same test twice, first presented with open-ended questions and then with multiple choice questions, a more detailed insight into the exact nature of the cueing effect could be obtained. It was found that cueing is a bidirectional effect, not only favouring the multiple choice questions, but often favouring the open-ended questions. The magnitude and direction of the cueing effect varies with the content and difficulty of the item. This led to the conclusion that the sole use of multiple choice questions would provide, to some extent, a biased estimate of the ability of the candidate.

In *chapter three* therefore, an item format specially developed for CCT was studied. It was hoped that this would have the advantage of being scorable by computers - which is impossible with open-ended questions - and would still require the student to generate the answer spontaneously. This item format was called Computerised Long Menu question (CLM). In the study CLMs were compared to hand-scored open-ended questions (OEQ) in one treatment group and to multiple choice questions (MCQ) in another. It was found that CLMs resemble open-ended questions more than they resemble multiple choice questions. Mean scores of CLMs and OEQs differed less than mean scores of CLMs and MCQs. In addition, (true) correlations between CLMs and open-ended questions were higher that those between CLMs and multiple choice questions. It was, however, also found that CLMs took more time to answer than multiple choice questions did. This may be seen as a disadvantage since it hinders broad content sampling. Therefore, CLMs can be a useful replacement for open-ended questions, but should be used only when a multiple choice question would not be valid.

In *chapter four* six research questions were studied to assess both the direct and indirect validity of CCT. For the direct validity it was studied whether students and teachers judge the content of cases and questions to be valid. Overall agreement was found that CCT poses tasks that closely resemble the problem-solving process in practice.
In addition, a multivariate generalisability procedure was used to assess the possible error variance due to the use of different question formats and different blueprint categories on the

total score. Negligible error source was found that could be attributed to either of these variables.
To assess the indirect validity it was studied whether a significant difference existed between the scores on an entrance level test and an exit level test, and whether expert scores differed from student scores on the exit level test. Finally, scores of CCT were compared to those on other measures of competence to study the pattern of correlations. It was found that a considerable increase in mean scores existed between the pre-test and the post-test, and that the experts scored even higher. Correlations between CCT and the other measures of competence were low to moderate and therefore difficult to interpret.
In summary, it was concluded that evidence for the validity of CCT emerged from the data, though the correlation matrix was not conclusive.

The next study was conducted to obtain a more direct insight into the nature of the thought processes elicited by CCT. This study is described in *chapter five*. General practitioners and students were presented with CCT cases and factual knowledge questions which were matched on content. They were asked to think aloud while solving these. This was audiotaped and typed out for further analyses. Specific indicators for analysis were defined before analysing the protocols. These were based on cognitive psychological research findings on problem solving. Indicators were word counts, the number of information units of the stem (or case) that were re-addressed after the question was read, the amount of sequentially and non-sequentially re-addressed information, and two assumed thinking steps: true-false and vectors. True-false would indicate thinking steps that considered whether something is factually true or false, whereas vector operations have a magnitude (probability) and a direction. Cases elicited thinking steps that were much more in line with the cognitive operations involved in problem solving than factual knowledge questions did. It was concluded that cases do measure problem solving in contrast to factual knowledge questions, and that the research methodology applied provides a useful addition to the 'armament' currently used in the validation of examinations.

In *chapter six* a comparison was made between students of the Maastricht University and students of the University of Groningen. All previous comparisons, using knowledge-based tests, had shown no significant differences between students of different medical schools. If with CCT differences could be found this would contribute to the external validity of CCT. Students of all year groups of both faculties were presented the same CCT test.
In the first three year groups the mean scores did not differ between schools (both groups of students of these year groups follow a similar type of curriculum). Starting from the fourth year group, however, the scores started to differ. In the final year a difference of about 8% was found in favour of the Maastricht students. In addition, the response times of the Maastricht student tended to decrease with increasing year group, whereas those of the Groningen students increased. It could be concluded form this that in the Maastricht students a higher level of problem solving takes place with increasing year groups, involving a processing of information in larger so-called 'chunks'.
It was concluded on both the proficiency scores and the response times that a difference between the student groups existed which could be attributed to differences between both curricula. CCT was, therefore, held to be sensitive to differences between curricula that cannot be detected with factual knowledge-based tests.

In *chapter seven* the experience gathered over the years of item writing for CCT is collected. Practical guidelines are given on how to write cases for authentic short case-based testing. First a general framework of a case is presented, and then strategies and pitfalls are described. The use of real patients is recommended since they provide a rich and authentic source. Considerable care must be taken to ensure a realistic and complete description in the case, both at the clinical and the contextual level. Information must be presented as 'raw' as possible; any pre-interpretation must be avoided.
The question must be aimed at essential decisions pertaining explicitly to the case. The number of questions per case must be limited to enable the administration of many different topics or problems per test. Specifically recommended is a careful review of all material by colleagues and the co-operation with other institutes to build a common database of cases.

In *chapter eight* the results are summarised and the research questions on validity of CCT are answered. Subsequently CCT is put into a broader perspective on the value of testing. Since other aspects, such as reliability, influence on learning behaviour, costs and acceptability are relevant too, these are discussed. Reliability of CCT is in range of .60 to .70 which is adequate for the setting in which CCT is used. The influence on student learning behaviour is still based on unsystematic debriefings of students. It seems to have the desired impact, although further study in this area is certainly needed. CCT is costly, with production time per case being about 2-3 person-hours. Therefore, co-operation with other medical schools in the Netherlands is being actively pursued. The acceptability is high among teachers as well as among students.
In general it is concluded that CCT provides an acceptable solution for the lacuna in the assessment system at Maastricht University.

SAMENVATTING

Toen de Universiteit Maastricht in 1974 startte met haar Probleem Gestuurd Onderwijssysteem (PGO) was deze onderwijsvorm een noviteit. Daarom was het van eminent belang dat binnen dit onderwijs een toetssysteem ontwikkeld zou worden dat aan de ene kant goed overeenstemde met de onderwijskundige uitgangspunten van PGO en anderzijds goed in staat zou zijn de kwaliteit van de afstudeerders te garanderen. Om hiertoe te kunnen komen werden een aantal uitgangspunten gedefinieerd waaraan alle toetsinstrumenten zouden moeten voldoen. Onder andere waren dit: congruentie tussen onderwijs en toetsing, scheiding van docent- en examinatorrol en een toetsing die gericht was op alle aspecten van medische competentie en niet slechts op enkele aspecten. In de afgelopen twee decennia zijn verschillende toetsinstrumenten ontwikkeld die voldoen aan veel van de gedefinieerde uitgangspunten, maar voornamelijk gericht waren op kennis en vaardigheden. Op het gebeid van probleem-oplossend vermogen echter bestond nog steeds een hiaat. De examens die hierop gericht waren, met name diegene die gebruikt werden voor de toetsing in de co-assistentschappen, waren erg divers. Ze varieerden van gestructureerde mondelinge examens, waarbij lange casus gebruikt werden, tot papieren op feitenkennis gerichte toetsen.
Als antwoord op deze hiaat is Computermatige Casusgerichte Toetsing (CCT) ontstaan. De idee hiervoor is gebaseerd op de ontwikkelingen binnen het gebied van de psychometrie en de cognitieve psychologie. In beide disciplines werd aangetoond dat probleem-oplossend vermogen niet de generieke vaardigheid was, die, wanneer ze eenmaal beheerst werd, in alle voorkomende situaties toegepast zou kunnen worden. Veeleer moest probleem-oplossen gezien worden als een domein-specifieke en idiosyncratische eigenschap. Dit zou inhouden dat een aanpak waarbij kleine aantallen van lange en vertakte casus, waarbij met name het proces van probleem-oplossen gemeten wordt, niet zinvol is. Daarentegen zou een aanpak met grote aantallen korte casus met een beperkt aantal vragen per casus een betere aanpak moeten zijn. De vragen zouden dan niet gericht moeten zijn op algemene kennisaspecten, maar veeleer op beslissingen die rechtstreeks betrekking hebben op de casus. Deze aanpak wordt de 'key-feature approach' genoemd. Om logistieke redenen zou het het beste zijn een dergelijk examen computermatig af te nemen.
Op het eerste gezicht moge een dergelijke aanpak de moeite waard lijken, een zorgvuldige studie naar de validiteit heeft nog niet plaatsgevonden. Het voornaamste doel van deze dissertatie is daarom de validiteit van CCT als een meetinstrument voor probleem-oplossend vermogen vast te stellen.

In *hoofdstuk 1* wordt CCT nader beschreven. Allereerst wordt een beschrijving gegeven van de soort casus die in CCT gebruikt worden. Alle casusmateriaal wordt gebaseerd op echte patiënten, en wordt geschreven door de arts die de patiënten behandeld heeft. Dit wordt gedaan om de authenticiteit van de casus te optimaliseren. Daarnaast wordt hiervoor multimedia (bijv. foto's, geluiden en filmpjes) gebruikt, doch alleen indien dit een

emeerwaarde heeft. De casus worden kort gehouden om het mogelijk te maken vele verschillende casus per toets te gebruiken, zodat de steekproef van onderwerpen zo breed mogelijk is.
Een zorgvuldige kwaliteitscontrole procedure is opgezet om de kwaliteit van de toetsen te verzekeren. Deze bestaat onder andere uit een uitgebreide training van de auteurs voordat ze casuïstiek beginnen te produceren, een zorgvuldige review procedure, waarin alle casus beoordeeld (en verbeterd worden) op relevantie, inhoud en formulering. Daarnaast worden alle casus nog gezien door een expert-panel om de antwoordsleutel en authenticiteit te beoordelen.
Verschillende vraagtypen worden gebruikt in CCT. De vraagvorm wordt bepaald door het aantal realistische opties dat in de praktijk bestaat. Zowel open als gesloten vraagvormen worden daarom gebruikt.
De voornaamste reden om computers te gebruiken in afname en scoring van de toets zijn logistieke voordelen, de mogelijkheid om multimedia te gebruiken, en de mogelijkheid om de casusinformatie stap voor stap te onthullen. Bovendien kan, door gebruik te maken van een computer database, voor iedere student een verschillende toets getrokken worden. Een verder (toekomstig) voordeel is het gebruik van sequentiële of adaptieve toetsing.

In *hoofdstuk 2* wordt het zogenaamde cueing-effect bestudeerd. Dit effect houdt in dat studenten een multiple-choice vraag correct kunnen beantwoorden doordat ze het juiste antwoord herkennen in het rijtje alternatieven. Bij open vragen is echter het spontaan genereren van het juiste antwoord een voorwaarde om de vraag correct te kunnen beantwoorden. Met name op basis van dit effect worden multiple-choice vragen ongeschikt geacht voor de toetsing van probleem-oplossend vermogen. Als dit waar zou zijn, dan zou dat serieuze negatieve consequenties hebben voor de validiteit van CCT. De literatuur over cueing, echter, is niet geheel eenduidig. Weliswaar suggereren de resultaten dat een cueing effect bestaat, maar ook dat de invloed van de vraagvorm op de te meten competentie beperkt is.
Door dezelfde toets twee maal aan proefpersonen voor te leggen, eerst met open vragen en vervolgens met gesloten vragen, kon een duidelijker beeld verkregen worden over de aard van het cueing effect. Er werd gevonden dat cueing een bidirectioneel effect is, dat niet alleen de gesloten vragen bevoordeelt, maar soms ook de open vragen. De grootte en richting van het effect varieert met de inhoud en moeilijkheidsgraad van het item. Indien CCT daarom alleen uit multiple choice vragen zou bestaan, zou de inschatting van de competentie van de studenten niet geheel accuraat zijn, hoewel de inschattingsfout beperkt zou zijn.

Daarom is in *hoofdstuk 3* een speciaal voor CCT ontwikkelde vraagvorm bestudeerd. Van deze vraagvorm werd gehoopt dat ze het voordeel zou hebben zowel per computer scoorbaar te zijn (wat niet mogelijk is met open vragen), en spontane generatie van het antwoord te vereisen. Deze vraagvorm is Computerised Long Menu (CLM) genoemd. In de studie van hoofdstuk 3 is deze vraagvorm vergeleken met handgescoorde open vragen in één onderzoeksgroep en met multiple choice vragen in een andere onderzoeksgroep.
Er werd gevonden dat CLM-vragen meer op open vragen lijken dan op multiple choice vragen. De verschillen in gemiddelde scores waren kleiner en de correlaties tussen CLM's en open vragen waren hoger dan die tussen CLM's en multiple choice vragen. Er werd echter ook gevonden dat de beantwoording van CLM's meer tijd kost dan multiple choice vragen.

Dat moet gezien worden als een nadeel daar hierdoor minder van deze vragen per uur toetstijd bevraagd kunnen worden, hetgeen de breedte van de steekproef kleiner maakt. Concluderend kan gesteld worden dat CLM's een goede vervanging kunnen zijn voor open vragen, maar dat ze alleen gebruikt moeten worden als het aangewezen is.

Hoofdstuk 4 bestaat uit een zestal deelstudies. In dit hoofdstuk wordt geprobeerd een indruk te krijgen van de directe en indirect validiteit van CCT.
Voor de deelstudies naar de indirecte validiteit zijn data gebruikt van een CCT toets die co-assistenten voorgelegd is aan het begin van hun co-assistentschap en van een andere CCT toets aan het einde van hun co-assistentschap. Ook zijn scores van experts verzameld. Daarnaast zijn de scores van de studenten op de eindtoets vergeleken met scores die zij behaald hadden op andere competentiedomeinen gerichte examens. Er werd een forse toename in scores tussen de begin- en eindtoets (CCT) gevonden; experts scoren nog hoger dan studenten op hun eindtoets. De correlaties tussen CCT en de andere examens zijn in het algemeen matig tot laag en daarom niet eenduidig te interpreteren.
Om een indruk te verkrijgen van de directe validiteit zijn studenten en experts gevraagd hun mening te geven over de authenticiteit en validiteit van CCT als een toets voor probleem-oplossen. Er was een hoge mate van overeenstemming onder de ondervraagden dat het oplossen van CCT casus grote overeenkomt vertoont met het oplossen van patiëntencasus in de praktijk. Ook is een multivariate generaliseerbaarheidsanalyse gedaan om te toetsen of er een mogelijke errorbron toegeschreven kan worden aan het gebruik van verschillende vraagvormen en de verschillende blauwdrukcategorieën in één toets. Geen storende errorbron kon gevonden worden die toegeschreven zou kunnen worden aan een van beide. Algemeen is geconcludeerd dat er weliswaar bewijzen voor de validiteit aanwezig zijn, hoewel de correlaties weinigzeggend zijn. Wat betreft de directe validiteit zijn de resultaten geheel positief.

In een volgende studie is geprobeerd een nog directer inzicht te verkrijgen in de denkprocessen die door CCT casus in gang worden gezet. Deze studie wordt beschreven in *hoofdstuk 5*. Huisartsen en vijfdejaars studenten geneeskunde hebben CCT casus en inhoudelijk gematchte feitenkennisvragen voorgelegd gekregen. Hen is gevraagd hardop te denken terwijl ze de casus en vragen oplosten. Dit is op cassette opgenomen en uitgetypt voor verdere analyse. Specifieke indicatoren voor analyse zijn gedefinieerd alvorens het materiaal geanalyseerd is. Deze indicatoren zijn gebaseerd op de bevindingen uit de cognitieve psychologie met betrekking tot probleem-oplossen. Als indicatoren zijn gebruikt: lengte van de protocollen, herlezen aantallen informatie-eenheden van de stam of casus nadat de vraag beantwoord is, de hoeveelheid sequentieel en niet-sequentieel herlezen informatie en twee geponeerde denkstappen: 'true-false' en 'vector'. 'True-false' stappen zijn denkstappen waarbij alleen maar afgevraagd wordt of een gegeven (geheel) waar of onwaar is, en 'vectoren' zijn denkstappen die een richting en een grootte (probabiliteit) hebben. De denkstappen die door de casus uitgelokt werden kwamen veel meer overeen met denkstappen die een rol spelen bij probleem-oplossen dan de denkstappen die door de feitenkennisvragen opgewekt werden. Er is derhalve geconcludeerd dat CCT casus meer probleem-oplossen meten dan feitenkennisvragen, en dat de hier gehanteerde onderzoeksmethodologie een waardevolle aanvulling kan zijn op het tot op heden gebruikte valideringsinstrumentarium van toetsinstrumenten.

In *hoofdstuk 6* wordt een vergelijking beschreven tussen studenten van de universiteit van Groningen en die van Maastricht. In Groningen heeft enige jaren geleden een curriculumherziening naar een probleem gestuurde onderwijsvorm plaatsgevonden. De Groningse studenten in de eerste drie jaargroepen volgden reeds het nieuwe, terwijl de studenten van de laatste drie jaargroepen nog het oude (traditionele) curriculum volgden.
In het verleden hebben studies waarbij studenten van verschillende type curricula werden vergeleken echter nooit enige significante verschillen op afstudeerniveau laten zien. Als daarom met CCT wel verschillen te zien zouden zijn zou dit een aanwijzing zijn voor een externe validiteit van CCT, daar op grond van het curriculumverschil met name verschillen in probleem-oplossen te verwachten zijn.
In de eerste drie jaargroepen verschillen de scores niet tussen beide universiteiten. Vanaf het vierde jaar beginnen de scores echter te verschillen. In het laatste jaar wordt zelfs een verschil van 8% gezien, ten faveure van de Maastrichtse studenten. Ook werd gevonden dat de responstijden van de Maastrichtse studenten de neiging hebben te dalen bij toenemende jaargroep, terwijl die van de Groningse studenten juist stijgen. Een mogelijk conclusie is dat de Maastrichtse studenten bij het stijgen van de jaargroep in staat zijn de informatie in steeds grotere eenheden (zogenaamde 'chunks') te verwerken. Dit kan gezien worden als een redelijk vergevorderde stap in het proces van expertise ontwikkeling. Bij de Groningse studenten zouden daarentegen juist uitgebreidere analytische processen plaatsvinden bij het toenemen van de jaargroep, hetgeen op een vroeger stadium van expertise ontwikkeling duidt.
Er is zowel op basis van de scorepatronen als van de responstijden geconcludeerd dat er een verschil tussen beide groepen bestaat dat kan worden toegeschreven aan aspecten van het curriculum. Er is daarom geconcludeerd dat CCT in staat is verschillen tussen curricula te detecteren, die niet gevonden kunnen worden met feitenkennisvragen.

Hoofdstuk 7 bevat de ervaringen met het schrijven van casus die in de afgelopen jaren zijn verzameld. Er worden praktische richtlijnen gegeven voor het schrijven van korte casus voor authentieke toetsing. Allereerst wordt een soort algemeen raamwerk van een casus gepresenteerd, en vervolgens worden tips en mogelijke valkuilen beschreven. Het wordt aanbevolen gebruik te maken van echte patiënten, omdat ze een rijke en authentieke bron vormen. Veel aandacht moet besteed worden aan een realistische en complete beschrijving van de casus, zowel voor wat betreft medisch technische als contextuele informatie. De informatie moet 'ruw' aangeboden worden, pre-interpretatie door de auteur moet vermeden worden.
De vragen moeten gericht zijn op essentiële beslissingen die expliciet betrekking hebben op de casus. Er moeten niet te veel vragen per casus gesteld worden, zodat per toets vele verschillende onderwerpen aan bod kunnen komen. Er wordt speciaal op gewezen dat een zorgvuldige review van alle materiaal moet plaatsvinden (door collega's), en dat samenwerking met andere instituten om tot een gemeenschappelijke database te komen sterk kostenbesparend kan zijn.

In *hoofdstuk 8* worden de resultaten tot slot samengevat, en worden de researchvragen met betrekking tot de validiteit beantwoord. Vervolgens wordt CCT in een breder perspectief over de waarde van toetsing geplaatst. Daar ook andere aspecten van een toets dan validiteit belangrijk zijn, zoals betrouwbaarheid, invloed op leergedrag, kosten en acceptatie, worden deze kort besproken. De betrouwbaarheid van CCT is in de orde van grootte van .60 - .70

hetgeen voor de setting waarin CCT nu gebruikt wordt voldoende is. De invloed op student leergedrag kan nog steeds slechts gebaseerd worden op een beperkt aantal niet-systematische interviews. Hieruit blijkt weliswaar dat CCT de gewenste invloed heeft, maar op dit gebied is zeker nog nadere studie nodig. CCT is duur, de productietijd voor één casus ligt in de orde van grootte van 2-3 persoon-uren. Daarom bestaat een actief beleid om te proberen een samenwerking met andere medische faculteiten in Nederland te starten. De mate van acceptatie van de CCT-methode is hoog, zowel onder docenten als onder studenten.
In het algemeen kan gesteld worden dat CCT een acceptabele oplossing is voor de in de inleiding geconstateerde hiaat in het toetssysteem van de Universiteit van Maastricht.

CURRICULUM VITAE

Lambert Schuwirth is geboren op 1 mei 1961 te Heerlen. In 1979 voltooide hij zijn atheneum β opleiding aan het Bernardinuscollge te Heerlen. In datzelfde jaar is hij gestart met zijn studie geneeskunde aan de toenmalige Rijksuniversiteit Limburg (heden: Universiteit Maastricht).
In 1988 is hij afgestudeerd, waarna hij een jaar als arts-assistent psychogeriatrie aan het Psychiatrisch Centrum Venray verbonden is geweest. Vervolgen heeft hij een jaar als opleidingsmedewerker aan de Vroedvrouwenschool te Heerlen gewerkt.
Sinds 1991 is hij verbonden aan de Universiteit Maastricht als universitair docent bij de capaciteitsgroep Onderwijsontwikkeling & Onderwijsresearch. Als coördinator en vice-coördinator is hij betrokken bij vaardigheidstoetsing, voortgangstoetsing en bloktoetsing. Tevens bekleedt hij het projectleiderschap van de projecten Computergestuurde Casusgerichte Toetsing met als doel de toetsing van probleem-oplossend vermogen te onderzoeken, en het project Computerised Problem-based Testing met als doel flexibele en interactieve toetsvormen te ontwikkelen. Daarnaast is hij als consulent verbonden aan diverse specialisten concilia met als doel de opzet en uitwerking van nationale arts-assitententoetsen.