

## An Inter- and Intra-University Comparison With Short Case-Based Testing

L.W.T. SCHUWIRTH<sup>1</sup>, B.H. VERHOEVEN<sup>2</sup>, A.J.J.A. SCHERPBIER<sup>1</sup>, E.M.A. MOM<sup>3</sup>, J. COHEN-SCHOTANUS<sup>4</sup>, H.J.M. VAN ROSSUM<sup>5</sup> and C.P.M. VAN DER VLEUTEN<sup>1</sup>

<sup>1</sup>*Department of Educational Research, Maastricht;* <sup>2</sup>*Skillslab, Maastricht;* <sup>3</sup>*Department of General Practice, Maastricht;* <sup>4</sup>*Department of Medical Education, Groningen;* <sup>5</sup>*Office of Education, Groningen*

**Abstract.** Comparisons between PBL and non-PBL medical schools on problem-solving ability often show no differences. This could be either due to the fact that no difference in problem-solving skills exists or that 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 program half way a 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.

**Key words:** case-based testing, comparison, problem solving

### Introduction

A claim of problem-based learning curricula (PBL) is that their medical students would become better problem solvers than students in a more traditional curriculum (Barrows, 1984). This claim was largely founded on the assumption that PBL students would acquire a general problem-solving ability which they then could 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 (Norman, 1988; Schmidt et al., 1996; Elstein et al., 1978). 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 and Mitchell, 1993; Vernon and Blake, 1993; Schmidt et al., 1987). 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 skills 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 et al., 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 out of this theory (Schmidt et al., 1990; Boreham, 1994; Regehr and Norman, 1996). 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 originate 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 as 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 lies probably in the inadequacy of the instruments used. The developments in these instruments have followed the insights of the previously described concepts of problem solving. 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 et al., 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 generalizabilities 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 (Schmidt et al., 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 (Elstein et al., 1978). 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 exists, 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 (Swanson et al., 1987; Norman et al., 1996; Maatsch and Huang, 1986; Schuwirth et al., 1996; 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. (1996) recently conducted a comparison between universities in the Netherlands, comparing (amongst others) one PBL curriculum with one traditional curriculum. In their study they presented

30 short cases each prompting for the (differential) diagnosis to 612 students. 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 concrete, 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 followed 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 self study, 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 I. 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 the sample was significantly different (in favor 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

*Table 1.* Descriptive Statistics of the Scores (Percentages Correct)

Year group	Groningen students			Maastricht students		
	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

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 emulate the achievement-motivational aspects of a real examination 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 computerized test (CCT) in which 60 cases were presented. A detailed description of the test and some sample items are described elsewhere (Schuwirth et al., 1996). 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 to validate the cases and the answer keys as experts. In the test used in this study the experts agreed on the correct answer for nearly all cases included. In a few instances a defensible '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 short answer open-ended questions were administered using a computerised long-menu format. This enables computerized scoring of the answers. The selection of the question format is based on the content of the question: when only a limited number of realistic alternatives could be gener-

ated multiple-choice questions were used; in other cases open-ended question types were used (Schuwirth et al., 1996).

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

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  $\times$  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 I reports the descriptive statistics of scores of both faculties. Table I shows 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:  $F(5,172) = 58.75, p < 0.0000$ ; Groningen:  $F(5,171) = 12.50, p < 0.0000$ ). 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

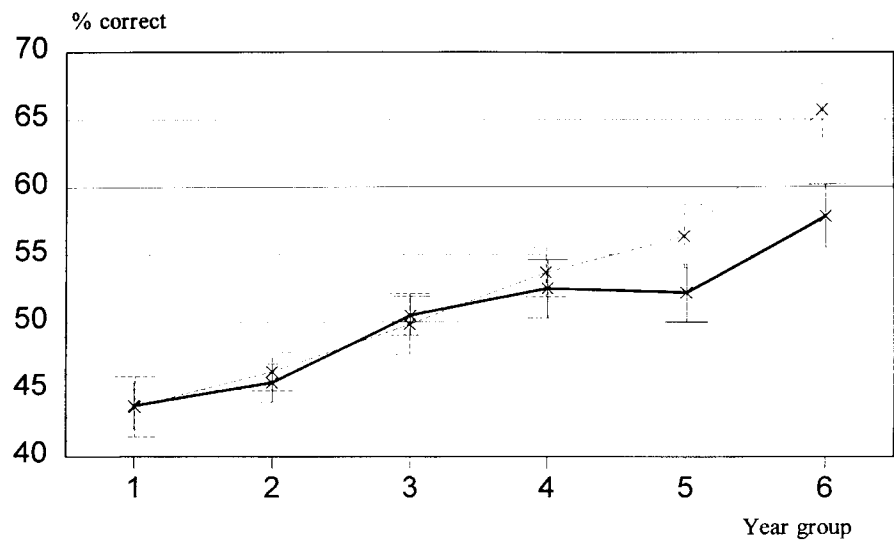


Figure 1. Mean Scores (Plus 95% CI) of Groningen and Maastricht Students (Broken Line).

Table II. 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  $M_c$

Year group	Groningen students				Maastricht students			
	$M_o$	$M_i$	$M_c$	$p$	$M_o$	$M_i$	$M_c$	$p$
1	16.7	19.6	16.0	0.001	20.7	22.6	18.7	0.0001
2	16.2	17.0	16.7	0.915	20.9	22.2	19.3	0.0080
3	18.1	20.4	17.8	0.069	19.4	21.7	17.3	0.0001
4	18.4	21.2	18.6	0.048	19.2	22.1	17.1	0.0001
5	19.5	21.0	21.1	0.981	21.5	25.0	17.9	0.0001
6	17.9	17.2	19.1	0.004	17.5	22.1	14.8	0.0001

is mainly due to the difference between the first two year groups and the last four. The  $2 \times 2$  ANOVA yields two main effects (year group and faculty) and a significant interaction effect ( $F(1,5,354) = 5.50, p < 0.000$ ). 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 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. Striking is the fact that the lines virtually converge in the first three year groups and begin to diverge in the fourth year in favor of Maastricht students.

In Table II the results of the median response times are described. A difference in patterns was found in the mean scores on all items ( $M_o$ ). In Maastricht the response times tend to decrease with increasing year group, whereas in Groningen they tend to increase. The columns denoted by  $M_i$  and  $M_c$  present the median response times on those items that were answered incorrectly and those that were answered correctly. Overall, Maastricht students took longer than Groningen students in all but the last year group, although these differences are not significant (except for the first two year groups).

In both schools students took longer to answer the items to which they failed to produce the correct answer compared to the correctly answered items. This difference is significant ( $p < 0.01$ ) in all year groups of Maastricht, whereas none of them are significant for the Groningen results.

### 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. (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 the study of Schmidt et al. 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 et al. 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 in the highest three year groups supports this assumption. An argument against the hypothesis is the fact that the nature of the clerkships in Maastricht differ only little from those in more traditional schools. In the light of this it is strange to find a difference in scores in the clerkship years. In the study of Schmidt et al. a similar pattern was found. They assumed that if a PBL effect occurred, this would only occur in the clerkship years after an 'incubation type of period'. Our results cannot show 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.

Response times show a remarkable trend; in Groningen they tend to increase with increasing level of expertise, whereas in Maastricht they decrease with

increasing expertise. The latter seems to be congruent with the results found by Van der Vleuten et al. (1995), who found a distinct decrease in mean response times with increasing expertise. The former, however, does not. 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 problem solving of the Maastricht students occurs at the down-bending end of the curve whereas that of the Groningen students is still at the up-bending or horizontal part of the curve.

In this view it is interesting to see the differences between the response-times of the correctly solved and incorrectly solved cases in both schools. These differences are significant in every year group of Maastricht and nowhere in Groningen. This would also support the assumption that the problem-solving ability of the Maastricht students is at a more aggregated level and would therefore be more efficient. When confronted with a case they do not 'recognize', however, these students would again use lower level problem-solving strategies that require more extensive (and more time consuming) analytical processes. The nature of the data of the present study, however, is not suited to yield further evidence for these rather speculative conclusions and further research is needed.

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 categorized into three categories: population effects, sample effects and instrumentation effects.

Population effects would occur by a self-selection of students 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 could only marginally occur. Second, once passed the lottery system they cannot simply choose the university they prefer, but are assigned to a university by a national committee by 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 on 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 randomization 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 for the best achiever might have emulated an actual examination situation somewhat more.

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 look. 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 computerized testing. But this is neutralized 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 the fact that in Maastricht the clerkship of General Practice was completed in the fifth year by most of the students and in Groningen in the sixth year. If this would have had an effect on the mean scores it would have pointed into an opposite direction: the difference in the fifth year would have been larger than in the sixth year since the sixth year students of Maastricht were most likely to have forgotten what they have learned during the GP clerkship (Semb and 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 favoring 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 skills 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.
- Albano, M.G., Cavallo, F. & Hoogenboom, R. et al. (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. Van Gorcum, Assen.
- Bordage, G. (1987). An alternative approach to PMPs: 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. Can-Heal Publications Inc, Montreal.
- Borcham, 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*. Harvard University Press, Cambridge, MA.
- Maatsch, J. & Huang, R. (1986). *An Evaluation of the Construct Validity of Four Alternative Theories of Clinical Competence. 25th Ann. RIME Conference*, pp. 69–74. AAMC, Chicago.
- Norman, G.R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education* **22**: 270–286.
- Norman, G.R. (1989). Reliability and construct validity of some cognitive measures of clinical reasoning. *Teaching and Learning in Medicine* **1**(4): 194–199.
- 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.
- 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 [Interfaculty skills test Groningen-Maastricht: first results]. In T.J. Ten Cate, J.H. Dijkers, E. Houtkoop, M.C. Pollemans, J. Pols & J.A. Smal (Eds.), *Gezond Onderwijs 5*, pp. 351–357. Bohn, Stafleu, Van Loghum, Houten/Diegem.
- 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. Springer, New York.
- 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): 305–315.
- Schmidt, H., Norman, G. & Boshuizen, H. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine* **65**(10): 611–622.
- 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.
- 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.
- Schuwirth, L.W.T., van der Vleuten, C.P.M. & Donkers, H.H.L.M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education* **30**: 50–55.
- Schuwirth, L.W.T., van der Vleuten, C.P.M., Stoffers, H.E.J.M. & Peperkamp, A.G.W. (1996). Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. *Medical Education* **30**: 50–55.
- 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): 220–246.
- 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. University of Toronto Bookstore Custom Publishing, Toronto.

- Vernon, D.T. & Blake, R.L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine* **68**(7): 550-563.
- 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. Springer Publishing Company, New York.
- Ward, W.C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement* **6**(1): 1-11.