Open-book tests assessed

Heijne-Penninga, Marjolein

2

Open-book tests to complement assessment programmes: analysis of open and closed-book tests

M Heijne-Penninga , JBM Kuks , J Schönrock-Adema, TAB Snijders,
J Cohen-Schotanus
Adv Health Sci Educ 2008;**13**:263–73.

Abstract

Today's health sciences educational programmes have to deal with a growing and changing amount of knowledge. It is becoming increasingly important for students to be able to use and manage knowledge. We suggest incorporating open-book tests in assessment programmes to meet these changes. This view on the use of open-book tests is discussed and the influence on test quality is examined.

To cope with the growing amount of medical knowledge, we have divided the body of knowledge into *core knowledge*, which students must know without need for references, and *backup knowledge*, which students need to understand and use properly with the help of references if so desired. As a result, all tests consist of a subtest for reproduction and understanding of core knowledge (a closed-book test) and a subtest for the ability to understand and manage backup knowledge (an open-book test). Statistical data from 14 such double-subtest exams for first and second-year students were analyzed for two cohorts (N=435 and N=449) with multilevel analysis, in accordance with generalizability theory.

The reliability of the open and closed-book sections of the separate tests varied between 0.712 and 0.850. The open-book items reduce reliability somewhat. The estimated disattenuated correlation was 0.960 and 0.937 for cohorts 1 and 2 respectively.

It is concluded that the use of open-book items next to closed-book items slightly decreases test reliability, but the overall index is acceptable. In addition, open and closed-book sections are strongly positively related. Therefore, open-book tests could be helpful in complementing today's assessment programmes.

Introduction

'What we call "the body of knowledge", is doubling every ten years', Spetz stated as early as 1989.[1] Today, with newer technologies, knowledge is growing even faster. For students, it is impossible to remember this growing amount of facts, some of which will have changed or been disproved by the time they begin their professional careers. Therefore, it is important that students in health sciences education programmes are able to use and manage knowledge when dealing with new problems and changed situations. This change in learning objectives implies changes in assessment, because assessment drives students' learning behaviour.[2-4] The use of open-book tests seems to be better aligned with these new learning objectives because they reduce the need for cramming and memorization of facts.[5-7] Generally, open-book tests are implemented to encourage students to use deeper learning approaches and to assess higher cognitive levels.[8,9] In this paper the use of open-book tests is discussed as a way to handle the growing body of knowledge.

Open-book tests

In the past, two reasons underlaid the use of open-book tests, namely improving the representation of the professional setting and encouraging deeper learning. Firstly, open-book tests were seen to be more representative of the professional setting in offering access to references in order to find answers to questions and solutions for the problems assigned.[1,2,7,8,10] Professionals do not rely heavily on memory; the open-book test is therefore closer to what is expected when 'on the job'.

The second reason for implementing open-book tests is to encourage deeper learning.[8,9] Open-book tests were expected to encourage teachers to ask questions on cognitive levels beyond recall. According to general opinion, items of the reproduction type are not suitable for open-book tests because answers can simply be copied from the references. Items assessing comprehension and application are considered as more suitable.[11,12] As Hoffman formulated it: 'The "plug and chug" questions are replaced by problems which require deep thought, understanding,

and intellect'.[13] Items assessing higher cognitive levels could encourage students to use deeper learning strategies in preparation, especially when the need for recall is limited. Students are stimulated to prepare in a more constructive way, for example by consulting more references and improving note-taking and active listening during lectures.[14] However, in our opinion, testing at a higher cognitive level and stimulating students towards deeper learning approaches is also desirable for closed-book tests.

In addition, we decided to implement open-book tests for a third reason. We expect that students will be able to study more knowledge in the available preparation time, thus allowing for more subjects to be covered. This view is supported by several studies which found that students spent less time preparing for an open-book test than for a closed-book test.[9,15-17] However, a possible threat to open-book tests is that students underestimate the need for preparation.[7,17] This lack of preparation could influence the psychometric quality of the tests negatively. To prevent underestimation of preparation and to match the competency of using and managing knowledge better, exams containing a closed-book and an open-book section were used. It also remains possible to assess knowledge that students need to remember and which is best assessed in closed-book tests with these exams. These exams were implemented throughout our undergraduate medical curriculum from the first year on. After using these dual exams for two years, we analysed the psychometric quality of the assessment procedure using generalizability theory.

Method

Context

The undergraduate medical curriculum of the University of Groningen and the University Medical Center Groningen is composed of ten-week modules. The content of each module is divided into *core knowledge* and *backup knowledge* (Figure 1). Core knowledge is the knowledge that every health science professional should know immediately and without needing to consult outside sources. Backup

knowledge is defined as knowledge that students need to understand and use properly with the use of reference sources if so desired.

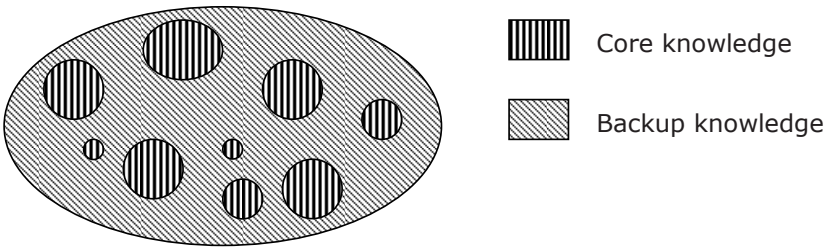

Figure 1. Total body of knowledge divided in core knowledge and backup knowledge.

Core knowledge is assessed in closed-book tests and backup knowledge in open-book tests. Teachers and experts decided which knowledge is core knowledge and which is backup knowledge.

Subjects

The first and second-year test results of two cohorts of medical students from the University of Groningen were analysed. The first cohort of students (n = 435) enrolled in 2003, the second cohort (n = 499) in 2004. The data from the two cohorts were analysed separately.

Procedure

Each cohort completed eight exams. The first exam in the first year was totally closed-book in order to allow students to get used to the new training course. These results were not included in this study. The following seven exams all consisted of a closed-book and an open-book section, together calculated as one final result. Each examination assessed the students' performance after completing an integrated module. Three modules from the first year and four modules from the second year were included. Table 1 shows the titles and the subjects dealt with in each module.

Within each module, a team of teachers was responsible for organizing the learning events and formulating problems and questions for the examinations. These teams varied per module.

Table 1. Titles and subjects of first and second year educational modules

Module	Title	Subjects
1.2	Foundations of Medicine	Medical science, basics of cell biology, endocrinology and genetics
1.3	Building on Health	Physiology and homeostasis
1.4	Care	Heath psychology, infection and immunity, training at a nursing home
2.1	Observe and React	Anatomy, physiology, disease symptoms of the nervous system and the sense-organs
2.2	Notice and Process	Clinical concepts of nervous-system diseases and sense organs
2.3	Regulation and Disorder	Chronic diseases, especially internal medicine
2.4	Chronic Loss of functions	Practical aspects of chronic diseases, medical science

Both the open-book sections and the closed-book sections were in the multiple choice format; this assessment form is standard because of the large number of students. Items had two, three or four alternatives. The number of alternatives per item varied between exams. The items were constructed by expert teachers and edited by specialists in test-item construction. Some teachers constructed only closed-book questions, other teachers only open-book questions, but most teachers constructed closed-book and open-book questions. Questions concerning different levels of understanding were formulated, although items assessing only the recall of facts were not allowed in the open-book sections.

All the examinations were in-class. The resources permitted for consultation during the open-book sections were only the literature supplied. Each exam started with the closed-book section. After a fixed period the answers were collected and the open-book section started. The time frame allowed for the open-book section was also fixed.

Statistics

The items were scored dichotomously (1 or 0 for the right or wrong answer). Questions of poor statistical quality were eliminated afterwards – a standard procedure in calculating students' results. The right answers were added to a total score for each section for every student.

Multilevel analysis was used to analyse the data.[18] Multilevel analysis is a flexible method to estimate models with several sources of variance and it allows taking into account the differences between tests in their number of items. This can be regarded as a more versatile way of implementing the approach of generalizability theory (G theory). G theory explicitly recognizes multiple sources of variance that contribute to the undifferentiated *E* (random error term) in classic theory.[19] G-theory consists of two kinds of studies – G (generalizability) study, in which the various sources of variance are estimated, and D (decision) study, in which the estimated variances are used to calculate different reliabilities.

The variances of three components were estimated – variance at the student level (level 3), variance at the test moment level (level 2) and variance at the test-item-set level (level 1). Level 1 indicates the set of items representing an exam section. There are two kinds of sections in this study, namely open-book and closed-book, each with different sets of items.

Seven exams were analysed for each cohort and each exam took place at a different moment and contained an open-book and a closed-book section. Every exam was intended to measure students' performance at that specific moment, thus variance at level 3 (representing the overall differences between students) and variance at level 2 (representing the differences within a student at several moments) are desirable variances. Variance at level 1 refers to the non-systematic differences between the scores obtained by a given student on the open-book and closed-book sections of a single exam; this is the unreliable variance, also called error variance. Because of the differences in test length (number of items) and the potential differences between the reliabilities of open-book and closed-book exams, the level-1 variance was allowed to differ between the two types of exam and to depend inversely proportionally on test length.

All variances were estimated using the MLwiN 2.02 statistical programme of Rasbash et al.[20] These variances can be combined into a single composite reliability value for each test, which gives the reliability of a complete exam, including both the open-book and the closed-book sections. The approach allows optimal weighting of the two kinds of sections, open-book and closed-book, and assessment of the contribution of each to the composite reliability. The latter was used to find the impact on reliability by changing the number of items within open-book or closed-book sections. The approach also allows estimation of true disattenuated correlations between subtests. The effect of measurement error has been removed, leading to a measure of the relationship between the open-book and closed-book scores of students in the hypothetical situation over completely reliable tests. Because the open-book section of exam seven of cohort 1 seemed to be an outlier (see Table 2, results section), this exam was excluded from the analysis.

Results

The exams were taken by 351 – 471 students, with a mean of 402.95 (377.6 in cohort 1 and 428.3 in cohort 2).

Table 2 . Description of the open and closed-book sections of the different exams of cohort 1

		No of students	No of items	Average score, %	SD
1.2 2003	CB	416	43	74.3	10.6
	OB		49	74.2	9.3
1.3 2003	CB	406	134	76.2	9.3
	OB		50	73.7	11.4
1.4 2003	CB	410	97	77.5	7.1
	OB		37	74.2	8.1
2.1 2004	CB	354	90	73.6	9.2
	OB		48	79.6	8.1
2.2 2004	CB	352	33	71.8	9.4
	OB		55	76.4	8.0
2.3 2004	CB	354	110	75.5	9.2
	OB		65	72.0	8.3
2.4 2004	CB	351	65	72.1	10.3
	OB		29	50.1	10.7

Tables 2 and 3 provide details of the number of items, number of students, average percentage score and standard deviation (SD) for each cohort.

Table 3. Description of the open and closed-book sections of the different exams of cohort 2

		No of students	No of items	Average score, %	SD
1.2 2004	CB	471	47	72.3	10.4
	OB		54	68.1	10.4
1.3 2004	CB	452	101	75.6	10.0
	OB		41	64.7	10.4
1.4 2004	CB	378	83	67.8	8.8
	OB		39	70.3	10.6
2.1 2005	CB	417	84	76.2	8.2
	OB		40	73.8	9.2
2.2 2005	CB	421	36	74.5	8.6
	OB		54	83.1	6.3
2.3 2005	CB	410	110	68.6	9.9
	OB		57	69.8	8.9
2.4 2005	CB	449	68	72.1	9.2
	OB		25	66.6	10.1

Generally, closed-book sections consisted of more items than the open-book sections. Only in exams 1.2 and 2.2 for both cohorts did the items in the open-book sections outnumber the items in the closed-book sections. On average, the cohort 1 exams consisted of 135.2 items, with 84.5 items in the closed-book and 50.7 items in the open-book sections. In cohort 2 the mean number of items was 119.9, with 75.6 items in the closed-book section and 44.3 items in the open-book.

The average percentage of correct answers ranged from 71.8 to 77.5 for the closed-book sections in cohort 1 and from 50.1 to 79.6 for the open-book sections. In cohort 2 these percentages were 67.8 to 75.6 and 64.7 to 83.1 respectively. Within the exams in cohort 1, these percentages were higher for the closed-book sections in five out of seven exams and in cohort 2 this was true for four out of seven. The differences in average scores were all significant, except for 1.2 2003. Exam 2.4 of cohort 1 seemed difficult, with a mean percentage of correct answers of 50.1 compared to the more than 70% in the other exams. Therefore, the open-book

section of this exam was indicated as an outlier and this exam was excluded from further analysis.

The estimated disattenuated correlation between the open-book and closed-book scores was 0.960 and 0.937 for cohort 1 and cohort 2 respectively.

As Table 4 shows, reliability varies from 0.712 to 0.845 for cohort 1 and from 0.747 to 0.850 for cohort 2. The reliabilities are somewhat higher for the second cohort than for the first. The lowest reliabilities, except in exam 2.4 in cohort 2, are exams containing more open-book items than closed-book items – test 1.2 and 2.2.

Table 4, composite reliability of the different test moments

	Cohort 1	Cohort 2
1.2	.723	.769
1.3	.845	.829
1.4	.799	.806
2.1	.769	.808
2.2	.712	.747
2.3	.836	.850
2.4	-	.761

The estimated reliabilities of varying numbers of items in the open and closed-book sections are shown in Table 5. The mean number of items and their distribution for both cohorts together was 128, with 80 in the closed-book section and 48 in the open-book section. These numbers were used to calculate the various reliabilities in Table 5.

As Table 5 shows, the estimated reliability, although not depending strongly on the mix of the two types of items, is highest for both cohorts when the test consists of only closed-book items. The reliabilities were all higher for cohort 2, these were all above 0.80.

Table 5. Estimated reliability with different numbers of items and different distribution of items in open-book and closed-book sections

Reliability when an exam consist of…	Cohort 1	Cohort 2
…80 closed-book items and 48 open-book items	.789	.812
…equal number of items (both 64)	.785	.809
…only closed-book items (128)	.800	.821
…only open-book items (128)	.775	.803

Discussion

In this study the psychometric quality of an assessment procedure with open and closed-book sections was examined. The results show that the use of open-book items alongside closed-book items is possible without much decrease in psychometric quality. The reliabilities were around 0.80, which is a preferable value.[21] Students' scores in percentages were lower in open-book tests, but student ranking was almost the same for the open and closed-book sections.

All reliabilities were of acceptable values. The estimated reliability was highest for tests with only closed-book items. This reliability was reduced slightly when more open-book items were included. A first possible explanation for this small decrease could be the novelty of the open-book test format. The formulation of open-book items is new or at least unusual for most teachers. Shine et al. concluded that the formulation of open-book items demands more thought and skill from test constructors.[22] The training and development – and commitment – of academic staff is necessary when adopting open-book tests as the standard mode of assessment. This could also explain why reliabilities are higher in cohort 2, where teachers had more experience with constructing open-book tests. A second explanation why reliability was slightly negatively influenced by more open-book items could be that students in health sciences are accustomed to preparing for closed-book rather than open-book tests. Most students were being confronted with open-book tests for the first time. It is possible that students did not prepare properly for the open-book tests. A study in a different educational setting

revealed that preparing for an open-book test requires a deeper approach.[23] However, those results are difficult to generalize to our educational setting. It would be interesting to examine the learning approaches of health sciences students in preparing for open-book and closed-book tests.

The distribution used in this study, 1/3 open-book items and 2/3 closed-book items, seems optimal when introducing open-book tests. Once students and teachers become more accustomed to open-book tests and the reliability of open-book tests improves further, this distribution can shift to a more balanced number of items. However, too great an increase in the number of open-book items is unrealistic since a difficulty with open-book tests is that students need longer to answer the questions. They often report a lack of time for open-book tests and tend to use far more time than necessary.[8,15] The amount of time required by the average student to answer an open-book item is not known and has yet to be examined.

Open-book tests are often suspected of being easier.[15] However, the results in this study contradict this assumption. The high correlation indicates that students performing poorly on the closed-book sections do not suddenly perform excellently on open-book sections. Furthermore, there was a small but significant trend indicating that students performed less well on open-book sections. Several reasons may account for this – (i) the open-book questions were truly more complicated, (ii) students prepared less thoroughly for the open-book test, as shown by others,[9,15-17] or (iii) students tend not to use their reference manuals appropriately and spent too much time in searching for information.[15,24,25] Broyles et al. already recommend advising students on how to prepare for an open-book test and how to use references best during the test.[7] Research showed that students who frequently refer to their books tend to achieve lower grades,[15,17,26] but no research has been done on how preparation for open-book tests influences test results. Therefore, further research is required to study what caused the difference in our study and how results were influenced by learning approaches and differences in learning approaches.

This study was based on two large cohorts of students and seven tests for each cohort. The results are consistent and show that sufficient reliabilities can be attained using an open-book section. A limitation of this study was that it concentrated on only one curriculum. Research is needed to determine whether these results are consistent in other educational settings.

An implication for practice is training and instruction for teachers to improve their skills in constructing open-book test items. Moreover, students seem to need training and instruction to be better able to prepare for open-book tests and to use references during testing. These implications could improve reliabilities further and could allow the distribution of open and closed-book items to be more flexible.

To summarize, we implemented exams with an open-book and a closed-book section to deal with the growing body of knowledge. This combination (2/3 closed-book and 1/3 open-book) had an acceptable reliability and the scores of the open and closed-book sections were consistent as regards student ranking. Contrary to commonly held suspicions, open-book tests were not easier than closed-book tests. Although several aspects of open-book testing still have to be investigated, this study showed that it is possible to use open-book and closed-book tests together and thus complement today's assessment programmes.

Acknowledgements

We are grateful to Mrs J. Bouwkamp-Timmer for her support with the literature research and her constructive comments on this manuscript.

References

1. Spetz NS. No Right Answer. *Hist Soc Sci Teach* 1989;**24**:73–5.
2. Frederiksen N. The real test bias: influences of testing on teaching and learning. *Am Psychol* 1984;**39**:193–202.
3. Cohen-Schotanus J. Student assessment and examination rules. *Med Teach* 1999;**21**:318–21.
4. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programms *Med Educ* 2005;**39**:309–17.

5. Feldhusen JF. An evaluation of college students' reactions to open-book tests. *Educ Psychol Meas* 1961;**21**:637–46.
6. Cain JC. Continuing medical education. *JAMA* 1979;**242**:1145–6.
7. Broyles IL, Cyr PR, Korsen N. Open book tests: assessment of academic learning in clerkships. *Med Teach* 2005;**5**:456–62.
8. Baillie C, Toohey S. The power test: its impact on student learning in a materials science course for engineering students. *Assess Eval High Educ* 1997;**22**:33–49.
9. Eilertsen TV, Valdermo O. Open-book assessment: a contribution to improved learning? *Stud Educ Eval* 2000**;26**:91–103.
10. Feller M. Open-book testing and education for the future. *Stud Educ Eval* 1994;**20**:235–8.
11. O'Grady G. Open-book tests. *CDT-Link Triannual newsletter of centre for Dvelopment of Teaching and Learning*. 2000. Available at: www.sdtl.nus.edu.sg/link/jul2000/practice2.htm.
12. Mohanan KP *Open-book tests: a response to some recurrent concerns*. 2004. Available at: http://courses.nus.edu.sg/course/ellkpmoh/educ/cdtl-obe.pdf.
13. Hoffman A. *All tests should be open-book*. 1997. Available at: http://www.summation.net/openbook.html.
14. Theophilides C, Dionysiou O. The major functions of the open-book test at the university level: a factor analytic study. *Stud Educ Eval* 1996;**22**:157–70.
15. Boniface D. Candidates' use of notes and textbooks during an open-book test. *Educ Res* 1985;**27**:201–9.
16. Zeidner M. College students' reactions towards key facets of classroom testing. *Assess Eval High Educ* 1990;**15**:151–69.
17. Koutselini-Ioannidou M. Testing and life-long learning: open-book and closed-book test in a university course. *Stud Educ Eval* 1997;**23**:131–9.
18. Snijders TAB, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London, Thousand Oaks, New Delhi: SAGE Publications Ltd. 1999.
19. Brennan RL. *Generalizability theory*. New York: Springer-Verlag 2001.
20. Rasbash J, Healy M, Browne W, Cameron B, Charlton B. *MlwiN 2.02. multilevel application for windows*. 2004.
21. Nunnally JC, Bernstein IH. *Psychometric theory*, 3rd edn. New York: McGraw-Hill 2000.
22. Shine S, Kiravu C, Astley J. In defence of open-book engineering degree examinations. *Int J Mech Eng Educ* 2004**;32**:197–211.
23. Theophilides C, Koutselini M. Study behavior in the closed-book and the open-book test: a comparative analysis. *Educ Res Eval* 2000;**6**:379–93.
24. Kalish RA. An experimental evaluation of the open book test. *J Educ Psychol* 1958;**49**: 200–4.
25. Francis J. A case for open-book tests. *Educ Rev* 1982;**24**:13–26.
26. Lubaway W, Brandt B. A variable structure: less resource intensive modification of problem-based learning for pharmacology instruction to health science students. *Naunyn-Schmiedeberg's Arch Pharmacol* 2002;**366**:48–57.