Dansk Universitetspædagogisk Tidsskrift

Tema Universitetspædagogikum

Årgang 13 nr. 25 / 2018

Titel **Validity assumptions for a multiple-choice test of medical knowledge with open-books and web access. A known groups comparison study**

Forfattere Lotte Dyhrberg O'Neill, Eivind Ortind Simonsen, Ulla Breth Knudsen, Jesper Stentoft, Anders Bonde Jensen og Charlotte Green Carlsen

Sidetal 134-150

Udgivet af Dansk Universitetspædagogisk Netværk, DUN

URL http://dun-net.dk/

Validity assumptions for a multiple-choice test of medical knowledge with open-books and web access. A known groups comparison study

Lotte Dyhrberg O'Neill, Associate Professor, Centre for Teaching and Learning, University of Southern Denmark

Eivind Ortind Simonsen, Computer Scientist, Centre for Health Sciences Education, INCUBA Science Park Skejby.

Ulla Breth Knudsen, Professor, Department of Clinical Medicine – Department of Obstetrics and Gynaecology, University of Aarhus

Jesper Stentoft, Professor, Department of Clinical Medicine - Department of Haematology, University of Aarhus

Anders Bonde Jensen, Professor, Department of Clinical Medicine - Department of Oncology, University of Aarhus

Charlotte Green Carlsen, Clinical Professor, Department of Quality Assurance, Aarhus University Hospital

Anne Mette Mørcke, Director of Copenhagen Academy for Medical Education and Simulation, Rigshospitalet

Research article, peer-reviewed

Relatively little evidence about the validity threats in open-book multiple-choice tests exist. The aim of this study was to examine validity aspects relating to generalization, extrapolation and decision of a multiple-choice test of medical knowledge with aids (open-book and internet access). The theoretical framework was modern validity theory, and the study was designed as a 'known groups comparison' study. Test performances of three known groups of test takers hypothesized to have different knowledge levels of the test content were compared, and analysis of pass/fail decisions was used to examine implications of decisions based on test scores. Results indicated that it was possible to discriminate between expert and non-expert test taker groups even with the access to aids. In contrast, an indefensible passing score was found to be the largest potential threat to test validity.

Keywords

Open-book assessment, Education, Medical Student, Performance Assessment, Validity.

Background

Paradoxically examinees are often denied access to check factual information - even in tests of applied knowledge, suggesting that in practice many test administrators treat factual and applied knowledge as directly interchangeable. There appears to be little published evidence documenting the necessity of denying examinees free information seeking. As a consequence, we know relatively little about the validity threats arising from the access to look up information in tests of applied knowledge.

Assessment is generally recognized as an extremely important driver of students' learning in higher education, and assessing higher order thinking skills has been considered to encourage 'deep learning' (J. Biggs & Tang, 2007; J. B. Biggs & Collis, 1982; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). This has led to an interest in open-book assessments, in which students can use textbooks, notes, journals etc. as reference materials during tests. Less focus on isolated factual knowledge recall could have the benefit of lowering the time and energy students tend to spend on cramming less relevant facts and on being stressed out in overloaded curricula. It seems that open-book assessment might reduce student anxiety and stress in higher education (Gharib, Phillips, & Mathew, 2012; Theophilides & Dionysiou, 1996; Zoller & Ben-Chaim, 1989), and encourage deep learning (Baillie & Toohey, 1997; Eilertsen & Valdermo, 2000; Theophilides & Koutselini, 2000), although these conclusions have been contested (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008).

In medical education proponents of open-book assessment have argued that access to information makes clinical problem solving tests more authentic and aligned with what happens in everyday clinical practice (Broyles, Cyr, & Korsen, 2005; Feller, 1994; Frederiksen, 1984; Heijne-Penninga, Kuks, Schönrock-Adema, Snijders, & Cohen-Schotanus, 2008; Spetz, 1989), and that their use may encourage deeper clinical learning (Broyles et al., 2005; Heijne-Penninga, Kuks, Hofman, & Cohen-Schotanus, 2011; Heijne-Penninga et al., 2008), and enhance long-term retention (Heijne-Penninga, Kuks, Hofman, Muijtjens, & Cohen-Schotanus, 2013). Allowing books etc. in the exam may signal to students that memorizing all isolated facts should be less of a worry, and that the main aim of their learning should instead be the meaningful integration of knowledge. Against this backdrop of suggested benefits, validity researchers have also started to explore open-book assessment (Brightwell, Daniel, & Stewart, 2004; Krasne, Wimmers, Relan, & Drake, 2006). Nevertheless, *the literature on the validity of open-book assessments is sparse, and there is currently no evidence for exclusively using either closed-book or open-book exams* according to a recent review (Durning et al., 2016).

Validity framework

The modern theoretical framework for examining questions relating to test validity is the 'unified' validity framework (American Educational Research Association, 2014;

Kane, 2006; Messick, 1987). In this framework, all sources of validity evidence are considered as counting towards construct validity. The idea of the existence of different 'types' of validity, such as content validity, predictive validity, concurrent validity, and discriminant validity etc., has been abandoned. Instead, Kane (2006) outlined four major categories of inferences, which may be examined and challenged in validation research. These are inferences relating to: *scoring* (from observed performance to observed score), *generalization* (from observed score to 'universe score'), *extrapolation* (from the universe score to the level of skill), and *decision* (from conclusion about level of skill to decisions taken). Using Kane's (2006) approach, we propose the main inferences relating to scoring in our context to be: 1. the electronic recording of students' responses represents students' intended answers, 2. the answer key for items is appropriate, and 3. the answer key is applied accurately and consistently. The main inferences for generalization are: 4. the observations made in testing are representative of the universe of observations defining the testing procedure, and 5. the sample of observations is large enough to control for sampling error. We propose the main inferences relating to extrapolation in our context to be: 6. the test tasks require the competencies developed in the course, and we may safely extrapolate expertise levels from the test scores, and 7. there are no skill irrelevant sources of variability that would seriously bias the interpretation of scores as measures of students' subject knowledge. The main inference relating to decision in our context was perceived to be: 8. Students with no or low levels of subject knowledge are unlikely to pass the test and progress in the programme. All inferences and assumptions cannot be evaluated in one single validation study, but rather in a programme of validation research. The most relevant kinds of validity evidence to examine first are those that support the main inferences and assumptions in the interpretative argument, *particularly those main inferences which are most problematic* (Cook, Brydges, Ginsburg, & Hatala, 2015). It seemed to us, that what Cook et al. (2015) called 'the weakest assumptions in the evidentiary chain' in our context were those related to extrapolation and decision (validity assumptions 6.-8. above), because the open-books and web access were allowed in these exams. An obvious competing alternative to argument 6 was, that the test tasks did *not* require the competencies taught in the course (subject knowledge), but merely access to aids like books and the web combined with good information seeking skills. If this competing interpretation proved correct, arguments 7 and 8 would also be seriously challenged.

Aims and objectives

The overarching aim of this study was therefore to examine aspects of validity relating to generalization, extrapolation and decision for a multiple-choice examination of medical knowledge with open-book and web access. The objectives were to: 1) ex-

amine dependability of test scores, 2) compare test scores from test takers with known differences in expertise levels, and 3) examine the pass/fail decisions for test takers with known differences in expertise levels.

Methods

The context of the study

Approximately a quarter of the curriculum at Aarhus University Medical School is assessed using multiple-choice examinations of medical knowledge, which allow open-book and web access. The learning outcomes tested in these exams are in the knowledge and applied knowledge domains. For examples of multiple-choice items testing these two types of knowledge we refer readers to the item writing guidelines by Case & Swanson (2002). The exams consist of 80 One-Best-Answer multiple-choice items each with three answer choices and a test time of 1.5 minutes per item, i.e. 2 hours of test time in total. The guidelines and checklists used for item construction were based on the work of the National Board of Medical Examiners (NBME) in the USA (Case & Swanson, 2002). Each test set is checked by an external examiner for relevance and validity of content before the exam. In the exams students bring along and use any written material they find useful, and they may use personal electronic devises to look up information on the device or on the internet if needed. Communication with others during the exam is prohibited. The students record their responses on iPads supplied by the university. Six to eight invigilators, who constantly move about the room checking examinees' screens and behaviours, enforce the communication ban. Mobile phones are stored away in bags and may not be handled during the exam.

Design

The study was designed as a 'known-groups comparison' study. Although a known-groups comparison study in isolation is never sufficient to claim validity of scores (Cook et al., 2015), accurate discrimination between groups with different expertise levels is an absolute necessary prerequisite for validity. The ability to differentiate between low-ability and high-ability test takers has been described as a 'fundamental principle of all educational measurement and a basic validity principle' (Downing & Yudkowsky, 2009). The known-groups comparison design allowed us to examine to what extent the test tasks in an open book/web setting *seemed* to require the competencies developed in the course, whether hypothesized expertise levels could be extrapolated from the test scores in the open book/web setting, and whether students with *no* or *lower* level of subject knowledge were likely to pass the test and progress in the program under open book/web conditions. In other words: our primary, initial validity concern was, whether unprepared students could pass the test under the open book/web conditions. This concern influenced our choice of compar-

ison groups, and made us compare relatively equally advanced students assumed to have different subject expertise levels depending on course participation/non-participation and on educational background (medical/non-medical). If the performances of test takers with no medical background would turn out to be indistinguishable from one or both groups with a medical background, it could indicate a considerable validity threat arising from the open-book web access conditions alone.

Participants

We compared three groups of students, which we labelled as medical experts, medical non-experts, and other non-experts.

As our 'expert' group sample we chose fourth year medical students who had previously completed the course 'Inflammation' and the corresponding open-book/web examination. They were labelled 'experts' in this study, because we were particularly interested in *challenging* whether the test tasks *seemed* to require the competencies developed in the course at all (see validity assumption 7) given the open book/web conditions.

As our non-expert test takers, we invited 2 groups of university students (medical students and other university students) to complete the same exam as the expert students and under similar test conditions. These two groups completed the test in February 2015. It was a requirement that non-experts were bachelor students in their last (third) year or just about to embark on their fourth year in order to avoid comparing our experts with novice students, and to allow for generic (i.e. subject independent) information seeking skills in the three groups to be as equally developed as possible. The medical non-experts were assumed to have some level of relevant medical background knowledge although they had not yet embarked on the Inflammation course, and so they were assumed to perform less well than the course takers (expert group), but better than university students with no medical background knowledge at all (other non-experts). The latter group was assumed to rely mainly on their generic (subject independent) information seeking skills or guessing, i.e. they were assumed to rely mainly on competing test constructs. Together these three groups were assumed to cover the full spectrum of expertise. In order to allow for the worst possible outcome for the test validity to be able to take place, and thereby to seriously challenge validity assumptions 6-8 above, it was essential that students with no medical backgrounds (i.e. the 'other' non-experts) were also allowed a go at performing well on the test. All volunteering students were asked to supply their names, e-mail addresses, their programmes of study and the semester they were on. The subject to be tested was not known to non-expert participants before the test.

Procedure

The physical facilities booked for the test of non-experts were the same as used in the ordinary exam. The test time (2 hours) was the same for experts and non-experts. Before the test started the purpose of the test and study was explained to the non-expert participants, along with the test time and the number of items to be answered. They were encouraged to use the open-book and web resources as they saw fit. Before the test started the non-expert participants were instructed on how to get access to the course e-book on their devices and the search facilities within this e-book. The non-experts answered the test electronically on the same iPads as used in the original examination. After 2 hours of test time the test was stopped.

Materials

The e-course book which participants were given access to was a basic book in medicine and surgery (Schroeder, Schulze, Hilsted, & Aldershvile, 2012). The exam paper was an exact copy of that used in the examination of the course 'Inflammation' on 14 June 2013. We chose this exam paper because these test results were amongst the most internally consistent test results available, and because the majority of the items in that paper tested applied knowledge as opposed to factual knowledge (Case & Swanson, 2002). Of the 80 items in this test, 64 tested applied knowledge while 16 tested factual knowledge. Scoring was 'dichotomous', i.e. 1 point was given for each item answered correctly, and 0 points were given for incorrectly answered items. This meant that test takers could obtain total test scores of between 0-80 items correct.

Analyses

We calculated the mean score, the range of scores, and the standard deviation (SD) of the test scores for the three groups. As our item discrimination index we calculated the correlation between students' performance on the item and their performance on the entire test, also known as the point-biserial correlation coefficient (PBS) (Case & Swanson, 2002; Haladyna, 2012). The level of PBS reflects 'the degree to which an item contributes to the measurement objective of the test' (Downing & Yudkowsky, 2009). It is an item characteristic which quantifies the item's ability to measure existing differences among individual test takers sensitively (Haladyna, 2012). The PBS coefficient values may range from -1 to 1, and at minimum PBS should be a positive number (Downing & Yudkowsky, 2009). PBS coefficients yield approximately the same information when dichotomous scoring is used, as the discrimination parameter from a two or three parameter model rooted in Item Response Theory (IRT) (Haladyna, 2012). In addition, PBS calculations do not require more than 500 examinees unlike the simplest IRT alternative (a two-parameter mod-

el), and it is therefore recommended for more modest sample sizes of examinees (De Champlain, 2010).

Analysis of generalizability

As validity evidence in the category generalization, we examined the dependability coefficients (Φ) of test takers' scores for each of the three groups, with a 'person crossed with item' design based on Generalizability Theory (Brennan, 2001). The phi (Φ) coefficient is a way of quantifying the *relative* influence of error on test scores. A phi coefficient of 0.70 for example, would - with our generalizability design - indicate that 70% of the observed score variance was due to real student performance differences, while 30% of the variance in scores was caused by error either related to the sample of items used or occurring at random.

Analysis of extrapolation

The test scores of the three groups of test takers were examined for equal variances with Levene's test, which confirmed unequal variances. After a quadratic transformation of data the problem of unequal variances was resolved, and the transformed data was examined for group differences with a one-way analysis of variance (ANOVA) test. Tukey's test was used to examine the significance of differences between all possible participant group pairs.

Analysis of decisions

The phi coefficients from the generalizability analyses were subsequently used to calculate the standard error of the measurement (SEM), which is another way of quantifying measurement error - this time in the same units as the test scores. A score difference between two test takers of > 1.96 x SEM may be considered statistically significant (Harvill, 1991). We inspected box plots of the scores of the three groups of test takers, and examined for differences in scores of the best performing non-expert and the poorest performing experts by checking for overlap in their 'reasonable limits' score bands (Harvill, 1991). Secondly, we calculated the number of non-expert participants who would have passed the test with the cut-score used in the original exam, as well as the percentage of the expert examinees who had scores >1.96 x SEM above the best performing non-expert test taker (Harvill, 1991). Thirdly, because of the test conditions (open-book and web), we also checked for the consequences of the presence of the 16 factual items in the test, by examining the scores and the pass/fail decisions of the three groups on the test with the factual items removed, to determine whether their presence made a difference.

All statistical analyses were performed using the statistical package STATA/IC 14, except the Φ coefficient which was estimated with the software GENOVA for PC.

Results

Of the 79 non-experts who volunteered to participate in the test, 71 turned up to participate on the test day. Of these 71 non-expert participants, 41 were medical students and 30 were other university students. The 30 'other' non-experts' academic backgrounds were: Arabic and Islam studies (n=2), Economics (n=10), Engineering (n=1), Physics (n=1), Japanese (n=1), Molecular Medicine (n=5), Molecular Biology (n=1), Psychology (n=6), Social Sciences/Philosophy (n=1), Political Science (n=2). Of the 71 non-experts 50 were third year students and 21 were fourth year students.

Item point-biserial correlations (item discrimination measures) ranged from 0.05-0.87 with a mean of 0.62 across the full range of behaviour in all groups (n=249). Two of the 79 items had point-biserials below the recommended minimum (<0.15), but none were close to 0 (Case & Swanson, 2002).

The average item completion rates were 100% (79/79) for the expert students, 92% (73/79) for the medical non-expert students and 71% (56/79) for the other non-expert students.

Table 1 reports the overall test characteristics and results based on 79 of the original 80 items in the test. One of the 80 items had to be removed from the analysis, due to an error occurring in the test of the non-experts (table 1).

Table 1. Main test indices (n_{items}=79).

Group	n	Mean scores	Range of scores	SD	SEM	Φ
Expert, medical	178	72.70	52-78	4.14	2.28	0.70
Non-expert, medical	41	37.20	9-50	8.46	4.26	0.75
Non-expert, other	30	23.63	12-39	7.54	3.94	0.73

SD=Standard Deviation, SEM=Standard Error of Measurement, Φ=the dependability coefficient for the absolute score values.

Generalizability

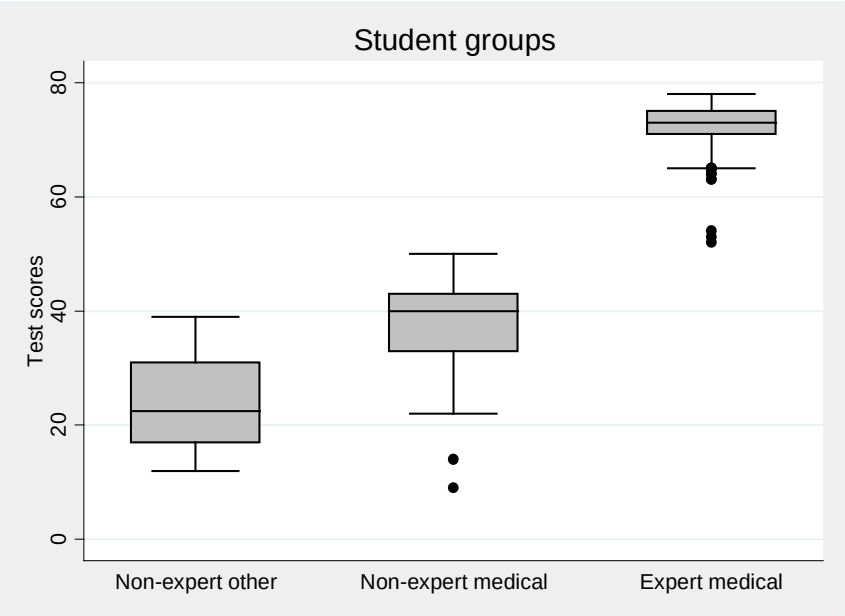
The dependability coefficients for the three groups of test takers are reported in table 1.

Extrapolation (comparison of test scores)

Figure 1 visualizes the test scores of the three groups expected to be at different levels of subject expertise by means of a 'box-and-whisker' plot. In this plot type, the grey box displays the interquartile range (the 25th to 75th percentiles). The 'floor' of

the boxes indicates the lower quartile, the 'roof' of the box indicates the upper quartile, and the central line is the median. The horizontal ends of the 'whiskers' attached to the boxes display the upper and lower values of scores falling within 1.5 times the interquartile range, while values outside this range are plotted individually (dots).

Figure 1. Test scores and outliers visualized.



The highest scoring medical non-expert scored 50 items correct. The three lowest scoring medical experts scored 52, 53 and 54 items correct respectively, while the fourth lowest scoring expert scored 63 items correct.

The ANOVA test showed that there were significant group differences in the test scores reported in table 1 and figure 1 ($F=1439.39$, $df_{between\ groups}=2$, $df_{within\ groups}=246$, $p<0.001$), and Tukey's post hoc test confirmed the statistical significance of all inter-group differences.

Decisions

The best performing non-expert scored 50 while the three poorest performing expert students scored 52, 53 and 54 respectively (see fig. 1). The reasonable limits score bands for the best performing non-expert and the three worst performing experts overlapped indicating no differences in performance. There was no overlap in the reasonable limits score band of the third and fourth lowest scoring expert; i.e. 175 of the 178 experts (98.3%) most likely scored significantly different than the best performing non-expert, and the three lowest scoring experts could be considered outliers relative to the other experts.

When applying the existing passing cut score for the test (>50% correct) to the results, we found that 51% (19+2=21 of 41) of the medical non-experts would have passed the test, while none of the other non-experts would have passed (see table 2).

Table 2. Test consequences by group (n_{items}=79).

n items correct	ECTS grade	Consequence	Experts (n=178)	Medical non-experts (n=41)	Other non-experts (n=30)
0-25	F	Fail	0	3	18
26-39	F_x		0	17	12
40-47	E		0	19	0
48-55	D	Pass	3	2	0
56-63	C		2	0	0
64-71	B		41	0	0
72-79	A		132	0	0

ECTS=European Credit Transfer System. The grades were assigned as follows: students with ≤32.5% correct items received the grade of F, students with >32.5% and ≤50% correct items were graded Fx, students with >50% and ≤60% correct items were graded E, students with >60% and ≤70% correct items were graded D, students with >70% and ≤80% correct items scored grade C, students with >80% and ≤90% correct items scored grade B, and students with >90% correct received grade A. The dashed line demarcates the cut score for pass/fail decisions used in the examination of the experts in June 2013.

Discounting the 16 factual items and analysing the consequences of a test containing only the 63 applied knowledge items with the same cut score (>50% correct) - we found it made no difference to the pass/fail decisions in the expert group (table 3). In contrast, two additional non-experts (one 'medical' and one 'other') would have passed the test if the 16 factual items had been discounted (see tables 2 and 3).

Table 3. Test consequences by group when the 16 factual knowledge items were removed (n_{items}=63).

n items correct	ECTS grade	Consequence	Experts (n=178)	Medical non-experts (n=41)	Other non-experts (n=30)
0-20	F	Fail	0	3	18
21-31	Fx		0	16	11
32-37	E		0	19	1
38-44	D	Pass	3	3	0
45-50	C		1	0	0
51-56	B		42	0	0
57-63	A		132	0	0

ECTS=European Credit Transfer System. The grades were assigned as follows: students with ≤32.5% correct items received the grade of F, students with >32.5% and ≤50% correct items were graded Fx, students with >50% and ≤60% correct items were graded E, students with >60% and ≤70% correct items were graded D,

students with >70% and ≤80% correct items scored grade C, students with >80% and ≤90% correct items scored grade B, and students with >90% correct received grade A. The dashed line demarcates the cut score for pass/fail decisions used in the examination of the experts in June 2013.

Discussion

Open-book/web conditions in a multiple-choice test of medical knowledge with 79 items and 1.5 minutes of test time per item did not undermine the ability to distinguish between known groups. In contrast, the arbitrarily chosen cut score could pose a threat to test validity.

Generalizability

We found dependability coefficients for the three groups in the ranges 0.70-0.75 for a 79-item test (table 1). What constitutes *sufficient reliability depends on the stakes and purposes of a test situation* (Downing, 2004). Downing (2004) suggested that very high stakes testing, such as licensure or certification examinations in medicine would require very high levels of reliability (≥0.90). End-of-course or end-of-semester type exams (like the exam situation we investigated) could probably defend levels of reliability in the ranges of 0.80-0.89. While lower stakes assessments, such as formative or summative classroom-type assessments, created and administered by local faculty (like the exam paper we investigated) might be expected to be in the range of 0.70–0.79 (Downing, 2004). The level of generalizability estimated in this study (table 1) was perhaps somewhat lower than would typically be required for the stakes of the test situation in which they were used in practice. Too few items in the test, low item discrimination (PBS) and the access to look up information may all affect test reliability negatively. Test reliability reflects the extent to which a test can differentiate or tell apart test takers' performances (Streiner & Norman, 2003). *For the purposes of this study*, however, the levels of reliability were sufficiently high to allow significant and meaningful discrimination of known groups of test takers (figure 1).

Extrapolation (comparison of test scores)

The significant group differences in test scores is evidence in support of validity assumption 6 outlined above: course participants ('experts') were more competent in solving test tasks than the non-participants (non-expert groups), and the expertise levels could also be extrapolated as expected, i.e. experts outperformed medical non-experts who in turn outperformed other non-experts. We found that information seeking was not a sufficiently influential cause of Construct Irrelevant Variance (CIV) or 'noise' (Downing & Haladyna, 2004), to make the assessment results uninterpretable in this open book/web multiple-choice test, meaning that validity assumption 7 could not be rejected. Others have previously reported examinees' test scores in open-book tests to be the same as their scores in closed book tests

(Kalish, 1958; Krarup, Naeraa, & Olsen, 1974), or significantly lower although student ranking was almost the same (Heijne-Penninga et al., 2008).

Decisions

As internationally recognized standard setting procedures apparently are at odds with Danish laws on examinations, and because test administrators of the original exam were obliged to choose a cut score before exams, an arbitrarily chosen pass score of >50% correct was imposed in the exam context. This arbitrarily chosen cut score appeared to be a greater potential threat to test validity, as such a sizable proportion (51%) of medical non-experts who had not embarked on the Inflammation course were able to pass the test with the cut score in operation (table 2). However, in practice only three of the 178 medical expert students (1.7%) who passed the original exam in June 2013 appeared to be no more competent, than a medical student who had never taken the course. Some of the best medical non-experts might well have had an excellent pre-existing knowledge base in physiology etc., which may have made it possible for them to deduce some answers, even though they had not taken the Inflammation course yet. A higher cut score could have secured the failure of all our non-experts as well as the outliers observed in the expert group (see fig. 1). Recognized *standard setting methods* (such as Angoff's, Ebel's, Hofstee's, borderline, or contrasting groups methods etc.) are generally recommended for the purpose of trying to reach the most defensible cut scores possible (Downing & Yudkowsky, 2009). However, challenging any cut score in operation (whichever way it was decided) with a reality check - as we did with this study - is recommended in all test settings (Downing & Haladyna, 2004; Livingston & Zieky, 1982), as indefensible cut scores may end up undermining test validity (Downing & Haladyna, 2004). The evidence in relation to validity assumption 8 indicated that test takers with *no* expertise level (other non-experts) were unlikely to pass the test, whereas test takers with *lower* expertise levels (medical non-experts) were relative likely to pass the test, although in reality it appeared to be a relatively rare occurrence (the expert outliers).

Interestingly, the presence of factual items did not seem to make it easier for the non-experts to pass the test (tables 2 and 3). One explanation of this result could be that they tended to be extensive information seekers irrespective of whether a factual or applied knowledge item was encountered. However, it could also be a coincidental finding.

In summary: we found an indefensible pass score, and *not* the 'off-target' or 'non-primary' construct of information seeking to be the largest threat to the test.

Limitations

Some limitations to the interpretation of results need to be mentioned. Firstly, we cannot rule out that group differences in test scores could have been even larger

under closed-book test conditions, i.e. we cannot rule out some negative effects on test validity because of the open book/web conditions. However, our immediate concern was whether or not information seeking was sufficiently influential to make assessment results uninterpretable.

Despite the high completion rates of our non-experts, the expert group's desire to do well in the test may well have been larger than that of our non-expert volunteers, which may have counted towards the group differences in scores observed.

We assumed that generic information seeking skills (e.g. using the search function in the electronic textbook and in google etc.) would be relatively equally developed in the three groups we compared, as all participants had around three years of university experience. However, we cannot be sure. If these generic or subject independent information-seeking skills were very different in the three groups, it may have biased results. Also, we cannot prove that our samples of non-experts were representative of their respective background populations on other important variables (e.g. intelligence, general academic skills), which may restrict the external validity of the results. However, voluntary participants generally tend to be more intelligent and better performing as students than non-volunteers (Callahan, Hojat, & Gonnella, 2007; Rosenthal & Rosnow, 2009), so it is more likely that our non-experts may have represented relatively capable challengers of the test validity.

Finally, it is still wise to adapt a cautious attitude, as it is also clear from the literature, that we need more studies from different contexts examining the threats to validity in open-book tests of medical knowledge before we can be more certain of any general tendencies (Durning et al., 2016). The next natural step in a validation programme would be to challenge whether test scores obtained under open-book/web condition also predict post-graduate performance.

Conclusion

No other previously published study of open-book/web assessment challenged validity assumptions in operation by examining the test performance of known non-experts. We found that free access to look up information did not undermine test validity to such an extent, that it was impossible to discriminate between known groups' hypothesized performances in a 79-item open book/web test of medical knowledge with 1.5 minutes of testing time per item. In contrast, an indefensible pass score was found to be the largest potential threat to test validity. Being allowed to use a recognized standard setting method, which takes into account the free access to information seems to be a more defensible approach for the future. These results add original and relevant knowledge to the limited existing body of studies examining the validity of open-book tests in medical education (Durning et al., 2016).

Acknowledgements

The authors would like to thank all the students who participated in the study.

Ethical approval

The project was exempt from ethics review by the regional ethics committee according to their policy on surveys, database studies and quality assurance studies. This committee supplied a response in English documenting this exemption. In addition, permission to handle the data in this project was sought at the Danish Data Protection Agency as required by law and granted in September 2014 (file number 2014-41-3417). The work has been carried out in accordance with the Declaration of Helsinki.

Declaration of interest

This study was funded by the Centre for Health Sciences Education and by the Institute for Clinical Medicine, both at Aarhus University, Denmark. The funding bodies had no role in study design, data collection, data analysis, data interpretation or manuscript preparation. The authors declare no conflicts of interests.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861-876.

American Educational Research Association, A. P. A., National Council of Measurement in Education. (2014). *Standards for educational and psychological testing.* Whashington (DC): American Educational Research Association.

Baillie, C., & Toohey, S. (1997). The 'power test': its impact on student learning in a materials science course for engineering students. *Assessment & Evaluation in Higher Education, 22*(1), 33-48.

Biggs, J., & Tang, C. (2007). Teaching for Quality Learning at University 3rd edition Open university Press: Maidenhead.

Biggs, J. B., & Collis, K. F. (1982). *Evaluation the quality of learning. The SOLO taxonomy (structure of the observed learning outcome).* New York: Academic Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives, handbook I: The cognitive domain: New York: David McKay Co Inc.

Brennan, R. L. (2001) *Generalizability Theory*. New York: Springer.

Brightwell, R., Daniel, J.-H., & Stewart, A. (2004). Evaluation: Is an open book examination easier? *Bioscience Education, 3*(1), 1-10.

Broyles, I. L., Cyr, P. R., & Korsen, N. (2005). Open book tests: assessment of academic learning in clerkships. *Medical Teacher, 27*(5), 456-462.

Callahan, C. A., Hojat, M., & Gonnella, J. S. (2007). Volunteer bias in medical education research: an empirical study of over three decades of longitudinal data. *Medical Education, 41*(8), 746-753.

Case, S., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia (PA): National Board of Medical Examiners.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education, 49*(6), 560-575.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*: Routledge.

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing Open-Book and Closed-Book Examinations: A Systematic Review. *Academic Medicine, 91*(4), 583-599.

Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: a contribution to improved learning? *Studies in Educational Evaluation, 26*(2), 91-103.

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation, 20*(2), 235-238.

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

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research, 2*(8), 469.

Harvill, L. M. (1991). Standard Error of Measurement. *Educational Measurement: Issues and Practice, 10*(2), 33-41. doi:10.1111/j.1745-3992.1991.tb00195.x

Heijne-Penninga, M., Kuks, J., Hofman, W., & Cohen-Schotanus, J. (2011). Directing students to profound open-book test preparation: The relationship between deep learning and open-book test time. *Medical Teacher, 33*(1), e16-e21.

Heijne-Penninga, M., Kuks, J., Hofman, W., Muijtjens, A., & Cohen-Schotanus, J. (2013). Influence of PBL with open-book tests on knowledge retention measured with progress tests. *Advances in Health Sciences Education, 18*(3), 485-495.

Heijne-Penninga, M., Kuks, J., Schönrock-Adema, J., Snijders, T., & Cohen-Schotanus, J. (2008). Open-book tests to complement assessment-programmes: analysis of open and closed-book tests. *Advances in Health Sciences Education, 13*(3), 263-273.

Kalish, R. A. (1958). An experimental evaluation of the open book examination. *Journal of Educational Psychology, 49*(4), 200.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64): ACE/Praeger.

Krarup, N., Naeraa, N., & Olsen, C. (1974). Open-book tests in a university course. *Higher Education, 3*(2), 157-164.

Krasne, S., Wimmers, P. F., Relan, A., & Drake, T. A. (2006). Differential effects of two types of formative assessment in predicting performance of first-year medical students. *Advances in Health Sciences Education, 11*(2), 155-171.

Livingston, S. A., & Zieky, M. J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.

Messick, S. (1987). VALIDITY. *ETS Research Report Series, 1987*(2), i-208. doi:10.1002/j.2330-8516.1987.tb00244.x

Rosenthal, R., & Rosnow, R. L. (2009). The volunteer subject *Artifacts in behavioral research* (pp. 48-92).

Schroeder, T., Schulze, S., Hilsted, J., & Aldershvile, J. (2012). *Basisbog i medicin og kirurgi*: Munksgaard.

Spetz, N. (1989). No right answer. *The History and Social Science Teacher, 24*, 73-75.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Theophilides, C., & Dionysiou, O. (1996). The major functions of the open-book examination at the university level: A factor analytic study. *Studies in Educational Evaluation, 22*(2), 157-170.

Theophilides, C., & Koutselini, M. (2000). Study behavior in the closed-book and the open-book examination: A comparative analysis. *Educational Research and Evaluation, 6*(4), 379-393.

Zoller, U., & Ben-Chaim, D. (1989). Interaction between examination type, anxiety state, and academic achievement in college science; an action-oriented research. *Journal of Research in Science Teaching, 26*(1), 65-77.