Hindawi Publishing Corporation
ISRN Education
Volume 2013, Article ID 840627, 6 pages
http://dx.doi.org/10.1155/2013/840627

Research Article

Can We Share Multiple Choice Questions across Borders? Validation of the Dutch Knowledge Assessment in Family Medicine in Flanders

Lynn Ryssaert,[1,2] Johan Wens,[3] and Birgitte Schoenmakers[4,5,6]

[1] *Interuniversity Centre for Education in General Practice, Kapucijnenvoer 33, Block J, P.O. Box 7001, Belgium*
[2] *Department of Family Medicine and Primary Health Care, Ghent University, De Pintelaan 185-6K3, Belgium*
[3] *University of Antwerp, Campus Drie Eiken, D.R. 315, Universiteitsplein 1, 2610 Wilrijk, Belgium*
[4] *Department of Public Health and Primary Care, Academic Centre of General Practice, University of Leuven, Kapucijnenvoer 33, Block J, P.O. Box 7001, 3000 Leuven, Belgium*
[5] *Department of Public Health and Primary Care, Academic Centre of General Practice, Catholic University Leuven, Leuven, Belgium*
[6] *Department of Public Health and Primary Care, Academic Centre of General Practice, Academic Teaching Practice, Leuven, Belgium*

Correspondence should be addressed to Birgitte Schoenmakers; birgitte.schoenmakers@med.kuleuven.be

Received 21 August 2013; Accepted 30 September 2013

Academic Editors: K. Kingsley and R. Pasnak

Background. One of the methods to test knowledge of Family Medicine trainees is a written exam composed of multiple choice questions. Creating high-quality multiple choice questions requires a lot of experience, knowledge, and time. This study explores the opportunity to run the Dutch knowledge assessment in Flanders as well, the use of this test for formative purposes. *Methods*. The study test was performed in a Flemish sample of postgraduate Family Medicine (FM) trainees and FM trainers. The Dutch test, adjusted to the Flemish context, was analyzed according to the classical test theory: difficulty factor and discriminating power of the items and reliability of the test. *Results*. 82 of the 154 items well divided the group into two equal parts of correct and incorrect responders. The distribution of the discrimination index, of the items with an acceptable difficulty factor, was [−0.012–0.530]. The item-test-correlation shows that 52 items do not fit, and 87 items need revision in varying degrees. The test reliability was 0.917. *Conclusion*. The test was highly reliable, but many MC questions appeared to be too easy and poorly discriminative. Therefore, we question the test validity and recommend reconsideration of the items based on difficulty before it is applied and used as a mandatory formative test.

1. Introduction

To test the knowledge of medical students, a set of evaluation methods are used: oral exam, Objective Structured Clinical Examination (OSCE), written exam, and so forth. The written exam can differ according to the type of questions used: the most commonly used tests consist of open-ended or multiple choice questions.

Multiple choice question exams (MCQ) have been proved to be an adequate way to evaluate knowledge on different educational aspects in the medical curriculum [1, 2]: purely theoretically, knowledge of facts, knowledge of adaptation, and clear understanding of mechanisms [3]. There are two major types of MCQ: one in which the responders need to mark the answers that are appropriate (type true/false). The other type consists of questions in which the examinee has to select one single best answer [2, 4].

The most commonly used MCQ exams consist of questions with only one correct alternative, questions where the beginning of the sentence needs to be completed with one of the alternatives (both type 1) or questions where one alternative presents as the best answer (type 2) [5].

Twenty-five years ago the true/false (MTF) questions were widely used in summative examination. Since then

many doubts on both concept and content have been uttered. The most frequent shortcomings of true/false multiple choice questions are (1) high chance of guessing the answer (2) marks not awarding for knowing the correct answer, but for knowing the incorrect answer, (3) weakness in discriminating high and low performers, (4) absolute true or false items leading to the assessment of trivial knowledge, (5) difficulty to test higher-order thinking, (6) difficulty to write flawlessly, (7) not encouraging learning around items, and (8) not assessing what it purports to assess [6]. Some literature even suggests that MTF exams only giving a limited insight into the intellectual capacities of the responders [6, 7]. This makes MTF questions best suited for formative testing and self-assessment but not for situations where test results are directly translated into a mark [8].

To address the criticism on MCQ examination, the use of extended matching questions (EMQ) has been advocated as an alternative for knowledge assessment [1, 9]. EMQ are always composed of four components: a theme or problem, a list of options or answering possibilities, a lead-in statement, and item stems [5, 9–11]. The advantages of using EMQ are as follows: (1) the format of themes makes it easy to structure the examination, (2) the questions are written according to theme allowing the author to produce many options that can be shared afterwards, (3) the questions are designed to assess application of knowledge, (4) the approach of writing these questions is systematic and determined in procedure, (5) a list of all relevant options limits the guessing option, and (6) the extended list of options has proved to be more discriminating than two- and five-option versions of the same item [9].

Whatever format of knowledge testing is preferred, creating multiple choice questions assumes having expertise in the topic and in the learning outcome [3, 12]. Above all, creating these questions is very time-consuming [13]. The feasibility and validity of the examination strongly depend on high-quality questions [5]. For all the above considerations, young academic assistants are not the best partners to produce questions (Berkel and Bax in [12]). Optimizing the workload and guaranteeing high-standard exams might benefit from collaboration between countries using the same language.

The Belgian medical education is divided into three trajectories: bachelor, master, and postgraduate specialization. In this project, we focused on the specialization of Family Medicine as a master-after-master trajectory of three years (years 7, 8, and 9 of medical education). The actual knowledge test, that is used in this phase of medical education, consists of 150 EMQ without guess correction. In The Netherlands, the medical education is comparable in structure but not in content or in assessment. The knowledge test used in The Netherlands is a true/false multiple choice examination composed of 155 questions (LHK1103) and guess correction is included (+1 and −1). Both test formats are thoroughly validated and composed according to the International Classification of Primary Care (ICPC) and one particular "family-practice-in-theory" chapter. The questions are very diverse and related to different age groups, chronic and acute illnesses, and different aspects of vocational skills.

Our interests were to find out if this Dutch knowledge test could be run in Flanders as well as using it for formative reasons and to examine the test validity when performed in a Flemish sample of medical students.

The following research questions need to be answered. (1) How reliable are the different items of the test? (2) How reliable is the test in general?

2. Method

2.1. Test Construction Procedure. First, a pilot study was performed which revealed that medical knowledge and the implementation of it were similar for both Flemish and Dutch students. This pilot study was necessary since some guidelines and topics related to the health care system differ between both countries. Second, although both languages, Flemish (spoken in the northern part of Belgium) and Dutch (spoken in The Netherlands), are practically the same, small difference in vocabulary occurs. Therefore, two researchers, independently from each other, analyzed the questions on linguistic usage and on adherence to guidelines. At the end of this procedure, 26 questions of the original knowledge test used in The Netherlands were deleted and 21 questions were newly developed with respect to the blueprint of the original test (guaranteeing that test content did not change). The final dataset contained 154 MTF questions which all were addressed in the pilot test [14].

2.2. Organization of the Test. Students from the three-year master-after-master training were invited to complete the test. This test was offered voluntarily and presented as a progress evaluation which students could easily complete at home within a limited time span. Inherent to the fact that the first year of the Family Medicine specialization is organized on university level, not all universities offered the test under the same condition. One university rewarded passing the test by exemption from another test. The students at the other universities completed the test on a mere voluntary base without a rewarding incentive [14].

2.3. Statistical Analysis: Classical Test Theory. The item difficulty is expressed by the difficulty factor as the amount of correct answers on the total amount of examinees, so the maximum this value can adopt is 1 [12, 15]. It gives a first indication on how well the item splits a group up into two parts: correct responders and incorrect responders. It is preferable to retain items that have a difficulty value between 0.3 and 0.7 [15]. Because the item difficulty value does not give any information on the direction of difficulty discrimination, an additional analysis is required. An upper-lower analysis and corresponding index of discrimination, together with the item-test-correlation, are necessary to achieve a more complete image of the contribution of the items. The first analysis checks how well the global score "predicts" the item score. The second method starts from the item score and checks how well this score correlates with the global score on the test [12].

The items of a knowledge test should discriminate between low- and high-ability candidates [15]. This means that for each item the proportion of high-performance

students answering correctly is higher than the proportion of low-performance candidates answering correctly. If this condition is fulfilled, the test can be generalized [12].

To check whether the observed score reflects the true score, a reliability coefficient is calculated. This value estimates the level of precision of the examination since every exam is subjected to different circumstances [15]. Circumstances influencing test outcome may be intrinsic; for example, the exam could have been composed of totally different questions but may also be extrinsic if, for example, a student has a bad day [12]. Cronbach's alpha was used to verify the reliability. The maximum value Cronbach's alpha can adopt is 1 which means that both groups present with identically the same output.

All analyses were done by using SPSS Statistics 20.

2.4. Ethical Approval. Ethical approval is granted by the Medical Ethical Advisory Board of all four universities for educational research in the ongoing academic year. Following the national legislation, an informed consent of the study population is only required when patients are involved.

3. Results

In total 353 students responded to our invitation to participate in this experiment. 110 were 7th year students (first year of master in Family Medicine) while 157 were 8th year students (second year of master in Family Medicine), and 86 were 9th year students (last phase before graduation). These numbers correspond to a response rate of, respectively, 48.2%, 93.4%, and 59.7% of the total number of students for each year [14]. To have a complete image of the validity of the test, also experienced family physicians, who all were involved in the medical training, took part. Of those invited, 11%, or 38 doctors, completed the test [14].

Fourteen examinees pulled out before the end of the trial; most of them were 7th year students. Table 1 visualizes where they quit the trial examination.

3.1. Difficulty Factor. Figure 1, three items have a difficulty factor lower than 0.3, while on 69 items 70% or more of the persons who took part answered correctly.

3.2. Upper-Lower Analysis and Index of Discrimination. The low- and high-ability examinees (Figure 2), in the upper-lower analysis, are defined as the first and fourth quartiles of the total score on the trial test.

Table 2 shows the composition of those two quartiles. In the 1st quartile, most 7th year students are situated. However, looking at the total response of the different subgroups, the final scores of more than one-third of the family physician trainers who took part in the trial are situated in the first quartile. In the fourth quartile, the 8th year students are relatively most represented. But globally, when the spread of all the obtained scores is studied in detail, most of the 9th year students are situated in the last quartile.

The highest value of the index of discrimination is 0.530. This means that from the total upper and lower scorers 53% more of the upper scorers answered the item correctly in comparison with the low scorers (last quartile). Two items have a negative index of discrimination, so more examinees who obtained the lowest global score on the test answered the item correctly compared to the high scorers. The distribution of the discrimination index, of the items with an acceptable difficulty factor (indicated by the rectangle), was [−0.012–0.530].

3.3. Item-Test-Correlation. The point-biserial correlation in Figure 3 gives an indication on how well every single item fits in the test.

52 items do not fit the test features at all with an $R_{it} \leq 0.19$ and should be eliminated or completely revised. 55 items with $0.20 \leq R_{it} \leq 0.29$ need revision because the item only measures in a small way the ability (global score) of the student. For 32 items, a small adjustment is required and only 15 items do not need any revision at all ($R_{it} \geq 0.04$) (Ebel, 1972, in [16]).

Table 1: Dropout.

	7th year students	8th year students	9th year students	Fam. Phys.
Item 3	—	—	x	—
Item 13	—	—	—	x
Item 23	x	—	x	—
Item 30	—	x	—	—
Item 31	—	—	—	x
Item 37	x	—	—	—
Item 45	—	—	—	x
Item 53	x	—	—	—
Item 86	x	—	—	—
Item 93	x	—	—	—
Item 101	—	—	—	x
Item 110	—	x	—	—
Item 111	x	—	—	—
Total	6	2	2	4

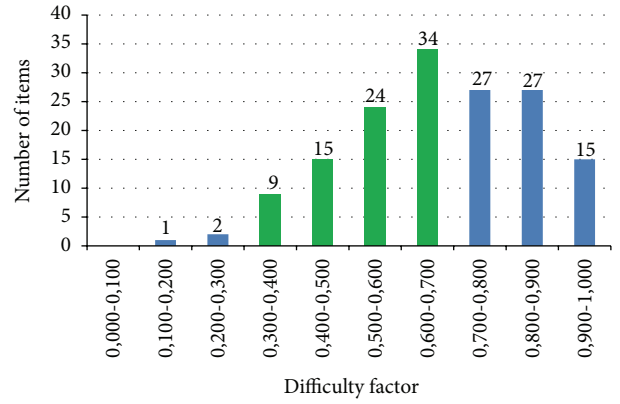

Figure 1: Number of items and corresponding difficulty factor, split up in categories of spread 0.100.

Table 2: The composition of the first and fourth quartiles of the final score of the trial examination.

	7th year students	8th year students	9th year students	Fam. Phys.	Total
1st quartile	33	30	6	14	83
In correspondence with the total amount of examinees belonging to that subgroup	31.1%	20.7%	19.4%	36.8%	25.9%
In correspondence with the total 1st quartile	39.8%	36.1%	7.2%	16.9%	100.0%
4th quartile	21	44	12	6	83
In correspondence with the total amount of examinees belonging to that subgroup	19.8%	30.3%	38.7%	15.8%	25.9%
In correspondence with the total 4th quartile	25.3%	53.0%	14.5%	7.2%	100.0%
Total	106	145	31	38	320

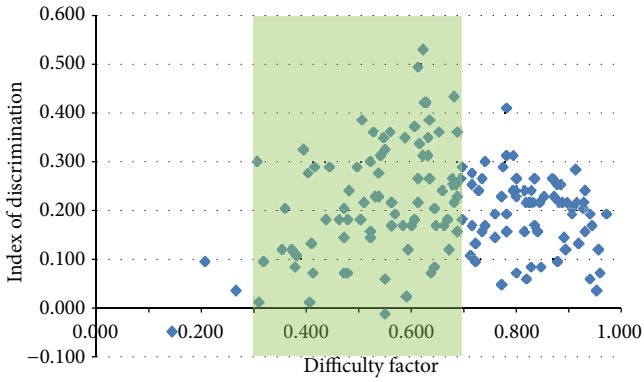

Figure 2: Index of discrimination and corresponding difficulty factor.

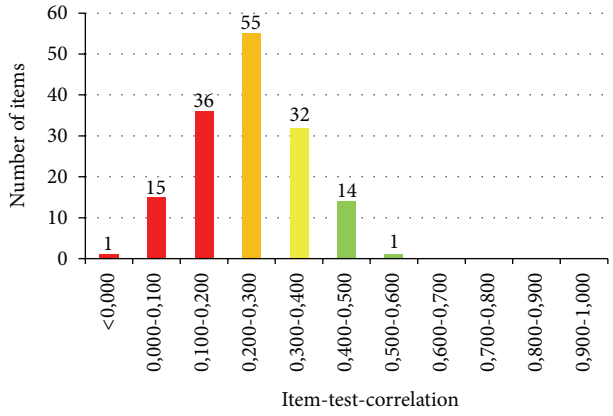

Figure 3: Number of items and corresponding item-test-correlation, split up in categories of spread 0,100.

19 items have no significant relationship with the final test score. This means that those items are not measuring the same construct as the test with an R_{it} of [−0.047–0.109]. Those items also performed with the lowest index of discrimination.

3.4. Reliability of the Test. The item response possibilities, seen as dichotomous values (one answer is correct; the other is incorrect or unknown), generate a Cronbach's alpha of 0.917, which is extremely near to perfect, being 1.000.

4. Discussion

In this experiment, the validity and reliability of replacing the knowledge test composed of EMQ by a true-false multiple choice examination were studied. The trial test was taken by 353 students and 38 experienced family physicians involved in medical education.

If the test items were closely studied, 3 items (or 1,9%) were identified with a difficulty factor of lower than 0.3. Those items may be too difficult. On the other hand, 69 items, or nearly half of the questions, had a difficulty factor higher than 0.7, which might identify them as easy. A more thorough study of those items will clarify whether this latter observation occurs in questions with similar themes, in questions where the phrasing is completely different or just occurs by coincidence. It is known that MTF questions are often either too ambiguous or too easy [5]. 45% of the questions can be identified as too easy, but that does not necessarily mean that those questions need to be eliminated from the exam. Any exam needs to contain a balanced mix of difficulty levels. Here, the examining staff must reconsider how this proportion is composed and justified.

Also the discrimination index and the item-test-correlation demonstrated that some items need to be revised. The distribution of the discrimination index of the 82 items with a difficulty factor in between the acceptable borders was very wide [−0.012–0.530]. This means that an examinee with a high global score on the test is not necessarily more likely to answer a difficult question correctly compared to an examinee of the low-score group. When using the ranges, set by Ebel (1972 in [16]), categorizing the R_{it}, only 15 items functioned in a satisfactory way while 52 items should be eliminated from the test or completely revised.

In spite of the remarks, the test reliability was 0.917. This is very high and demonstrates that if someone would take the test for a second time, total test scores would be very similar. Nevertheless, the high test reliability combined with the difficulty factor and discrimination power indicates that the test might have been too easy. This presumption is confirmed when interpreting the final score (mean − SD, calculated for trainers and trainees separately) with 87% of the trainers and 89% of the trainees passing the test.

It was rather surprising that the test score of 36,8% of the trainers was found in the first quartile. Possibly

the experience of the trainers will explain this observation. The family physician does not simply adopt and follow guidelines when making a clinical decision. Other factors like patient preferences, clinical judgment, and experience are playing key roles in the diagnostic process and in decision making [17].

The results need to be interpreted carefully because the chosen statistical method is sample dependent. This type of analysis can give some indication on the difficulty level of an item, but it does not allow an interpretation on the competence of the candidates who completed the exam [15]. The Rasch model of Item Response Theory (IRT) which examines on the one hand the difficulty level of the item and on the other hand the competence of the candidates could offer an alternative way of analyzing [9]. Besides, in the method used, the questions are analyzed as if there were dichotomous response categories (true or false). But there was also a third answer possibility "do not know," which was considered as "false." Closer analysis on the difference between "false" and "do not know" may also give new information on how the questions are experienced. Hypothetically, the difficulty factor of two different groups on a certain item may be the same, but if in one group there are more nonresponders than in the other group, the interpretation of the real difficulty level of the item is totally different [15].

A significant limit of the test is the different ways in which the trial test was organized at university level. The fact that some students were rewarded for taking part in the trial test can distort the results of this study. Also the remarkable response rate of 93.4% of the 8th year students is probably related to the fact that one month after the trial test, the real exam, based on EMQ, was planned. Those students were preparing themselves for that test during the trial period which can explain why they were more interested in participating. The different level of motivation and preparation of the participants hinder the generalization of findings. Still the results presented in this paper might give a first indication.

5. Conclusion

It is evident that statistical analysis on MCQ is only one part of the evaluation of a test or an exam. This mathematical approach may never lead to immediate conclusions on test validity. Clear and profound reasoning on how and why a certain result on an item could be obtained and explained is equally important [3]. No doubt that medical knowledge needs to be tested with high-quality questions, but this condition is more essential for summative assessments than for formative ones [18]. This knowledge test certainly needs reconsideration of the items based on difficulty level before it is applied and used as a mandatory formative test.

Conflict of Interests

The authors declare to be free of any competing interests.

Acknowledgments

The authors wish to thank Professor An De Sutter (University Gent) and Professor Jan Kartounian (University Brussels).

References

[1] S. P. Coderre, P. Harasym, H. Mandin, and G. Fick, "The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts," *BMC Medical Education*, vol. 4, article 23, 2004.

[2] L. W. T. Schuwirth and C. P. M. van der Vleuten, "Different written assessment methods: what can be said about their strengths and weakness?" *Medical Education*, vol. 38, no. 9, pp. 974–979, 2004.

[3] O. Geneeskunde, "Richtlijnen mbt het evalueren van meerkeuzevragen voor de opleiding geneeskunde," 2001.

[4] (NBME) NBoME, "Constructing Written Test Questions for the Basic and Clinical Sciences," 2002.

[5] D. E. Campbell, "How to write good multiple-choice questions," *Journal of Paediatrics and Child Health*, vol. 47, no. 6, pp. 322–325, 2011.

[6] M. Chandratilake, M. Davis, and G. Ponnamperuma, "Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions," *National Medical Journal of India*, vol. 24, no. 4, pp. 225–228, 2011.

[7] R. Richardson, "The multiple choice true/false question: what does it measure and what could it measure?" *Medical Teacher*, vol. 14, no. 2-3, pp. 201–204, 1992.

[8] J. Anderson, "Mutiple-choice questions revisited," *Medical Teacher*, vol. 26, no. 2, pp. 110–113, 2004.

[9] B. Bhakta, A. Tennant, M. Horton, G. Lawton, and D. Andrich, "Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education," *BMC Medical Education*, vol. 5, no. 9, 2005.

[10] J. Beullens, B. van Damme, H. Jaspaert, and P. J. Janssen, "Are extended-matching multiple-choice items appropriate for a final test in medical education?" *Medical Teacher*, vol. 24, no. 4, pp. 390–395, 2002.

[11] A. S. Baird, "The new Extended Matching Question (EMQ) paper of the MFSRH examination," *Journal of Family Planning and Reproductive Health Care*, vol. 36, no. 3, pp. 171–173, 2010.

[12] H. de Neve and P. Janssen, "Succesvol examineren in het hoger onderwijs," Acco, Leuven, Belgium, 1992.

[13] J. Collins, "Writing multiple-choice questions for continuing medical education activities and self-assessment modules," *Radiographics*, vol. 26, pp. 543–551, 2006.

[14] B. Schoenmakers, A. De Sutter, J. Kartounian, A. Stockmans, and J. Wens, "Project: Valideren van de Nederlandse Kennistoets in Vlaanderen".

[15] A. F. De Champlain, "A primer on classical test theory and item response theory for assessments in medical education," *Medical Education*, vol. 44, no. 1, pp. 109–117, 2010.

[16] A. Zumbairi and N. Kassim, "Classical and Rasch analyses of dichotomously scored reading comprehension test items," *Malaysian Journal of ELT Research*, vol. 2, pp. 1–20, 2006.

[17] C. S. Tracy, G. C. Dantas, and R. E. G. Upshur, "Evidence-based medicine in primary care: qualitative study of family physicians," *BMC Family Practice*, vol. 4, article 1, 2003.

[18] A. Vanderbilt, M. Feldman, and I. Wood, "Assessment in undergraduate medical education: a review of course exams," *Medical Education Online*, vol. 18, 2013.