Original Article

Proficiency testing for admission to the postgraduate family medicine education

Birgitte Schoenmakers¹, Johan Wens²

¹Department of Public Health and Primary Care, University of Leuven, 3000 Leuven, ²Department of Public Health and Primary Care, University of Antwerp, Antwerp, Belgium

ABSTRACT

Theory: In Belgium, there are no family medicine admission requirements. A three-phase admission program is developed and implemented by the collaboration of four involved universities. **Hypotheses:** A pilot testing of phase 2, comprising actual proficiency test, is designed as answer to two research questions: What is the validity and reliability of a multicomponent proficiency test? How does a multicomponent proficiency test proportionate to the final grades of family medicine master candidates? **Methods:** The population consisted of all last master-phase students applying for family medicine education in Flanders. Students completed a machine-assisted test on knowledge and situational judgment skills and evidence-based medicine appraisal. **Results:** In total, 322 students completed the test. A regression analysis measuring the relationship between the master grades and the test score revealed an odds ratio of 1.1. Analysis of variance showed that the differences were significant between the upper quartile and the lowest quartile of the test results. A qualitative appraisal of the test results showed that the highest and lowest quartiles of the full-test score included the students who were, respectively, known as "very good" or "very poor." **Conclusion:** The test scores were in agreement with the performance and profiling of the participating students. The test succeeded in identifying poor-performing students and in confirming competences of the average- and high-performing students. In the future, retesting will add to the statements on reliability and will refine the test construction. Follow-up will address validity.

Keywords: Admission, family medicine, medical education, proficiency testing

Introduction

In the past two decades, family medicine transformed into a full discipline and plays a major role in health care and in medical education.[1,2] Family medicine also remains a discipline dealing with its very own particularities. Nevertheless, admission to advanced master in family medicine is not commonly regulated.[3,4] Most traineeships are still mainly organized in a hospital setting. Students, therefore, hardly come in contact with family medicine, and they are not familiar with the required skills and competencies in this discipline.[5]

Due to evolutions in health research and to a changing societal reality (graying and silvering, rise of informal care, etc.), the provision of health care changed in a substantial way.[6] Nowadays, the health-care system puts more emphasis on primary care and on the role of the family physician. This medical professional is more than other care professionals confronted with the complex reality and consequences of chronic care.[7] Almost, related to medical and societal developments, patients demand more participation and sharing of decision-making.[8] These evolutions require a well-trained family physician skilled in more than only the medical expert competences.

The (worldwide) innovative family medicine education program attracts an increasing number of students.[9] The emphasis on

Address for correspondence: Prof. Birgitte Schoenmakers, Department of Public Health and Primary Care, University of Leuven, Kapucijnenvoer 33 Box 7001, 3000 Leuven, Belgium. E-mail: irgitte.schoenmakers@kuleuven.be

How to cite this article: Schoenmakers B, Wens J. Proficiency testing for admission to the postgraduate family medicine education. J Family Med Prim Care 2018;7:58-63.

workplace-based learning, comprehensive communication and vocational training, and adaptation of evidence-based guidelines to the primary care setting favors family medicine education. Family medicine evolved from a drop-out discipline to an opt-in specialty: more students explicitly opt for family medicine.[10]

However, with an increased number of family medicine students, there is a need for an adequate and reliable admission procedure[11-13] As in other disciplines, the aim of admission testing is to include the most suitable candidates.[14,15] A second aim is to give students insight into the family medicine discipline before they apply for admission. Third, in the near future, family medicine will be confronted with an overabundance of candidates in many Western European countries.[16]

In Belgium, admission regulation is not a public but a university matter. In collaboration between the four Flemish Universities, a three-phase admission procedure was developed and implemented. Phase 1 comprises the formal admission requirements (master degree in medicine, introduction course in family medicine, motivation letter, and language skills), phase 2 comprises a multicomponent, machine-assisted test, and phase 3 refers to the structured jury examination addressing students who failed the multicomponent test. Students passing phase 1 and succeeding in the multicomponent test are admitted to the advanced master in family medicine.

In this report, the results of the study performed on phase 2, comprising the actual proficiency test (the multicomponent test), are presented in two research questions: What is the validity and reliability of a multicomponent proficiency test in an admission procedure? How does a multicomponent proficiency test proportionate to the final grades (master score) of family medicine master candidates?

Methods

Research questions and outcome measures

The first research question was what is the validity and reliability of a multicomponent proficiency test? Validity and reliability of the individual components of the test are approached here.

The second research question was how does a multicomponent proficiency test proportionate to the final grades (master score) of family medicine master candidates? The option to use the final grades is defended by the assumption that these marks do not reliably predict the capability of each individual student to become a family physician.

Design

To design the multicomponent test, a step-wise, structured procedure was followed. The first step consisted of a comprehensive literature and field study and the consulting of experts to retrieve information on proficiency testing in medical education. The second step was the structural design

of a test composed of three components: knowledge testing, testing of skills in evidence-based medicine (EBM) test, and situational judgment testing (SJT). The absolute contribution assigned to each component was different: the SJT counted a maximum of 281 points, the knowledge test 114 points, and the evidence based medicine (EBM) test 11 points. The final, relative weight or real contribution of each component to the total test score was calculated after psychometric analysis of the test scores. The third step comprised the building of content of the test components. All test components were constructed in a multiple-choice (MC) format. The knowledge test was designed as a true-false test and contained 114 questions. Students had to mark the degree of certainty for each answer (on a scale from 0 to 100), and there was no correction for guessing. The test content was developed considering that students in this phase were not (or merely) instructed or familiar with family medicine skills and competences. This implies that the questions rather addressed the capability to reason as a primary care physician than the correct use and application of (primary care) guidelines. The EBM test was based on four articles on clinically relevant research topics (e.g., use of prostate-specific antigen to screen prostate cancer, impact of food supplements on cholesterol levels) and assessed by 11 MC questions with 4–6 answer options (referring to the articles). The SJT was based on 20 realistic cases addressing competencies in ethical and moral considerations, in decision-making, in professional attitude, and in collaboration. The answer options were offered in two formats: ranking of best options and choosing the three best options. The answer options were ranked and validated by a team of field (GP's) and academic (teachers and fellows) experts and supported by the literature on the topics involved. The fourth step consisted of the pilot testing of the multicomponent test. This pilot study addressed the face and construct validity of the composite test and of the individual components. After the pilot, adjustments were made: in particular, the reformulating of knowledge questions to improve understanding.

The multicomponent test was offered as a machine-assisted test. The digital platform was built in close collaboration with the main author of this report.

Population

The population consisted of all (333) final master-phase students applying for the admission to the advanced master in family medicine in Flanders. Participation in the admission test was obligatory, but the results were not binding in this study setting. Students were offered feedback on the test results together with an advice for further orientation. The test took place at the end of the examination period. Since all students were recruited from the same master phase, no demographics were added to the analyses.

The test was developed and organized by the four Flemish universities: University of Ghent, Leuven, Brussels, and Antwerp.

Analyses

In the first step, descriptive analyses were made of each individual component of the test (univariate analysis). In the second step, reliability and validity testing of each component was performed. Reliability was approached by measuring the internal consistency with Cronbach's alpha. This approach was only applied for the knowledge test. The answer options in the SJT are heterogeneous (both ranking and best options) and the number of questions of both the SJT and the EBM test was too low to obtain a reliable Cronbach's alpha. Reliability assessment of these components was therefore based on the psychometric features as Gauss distribution and variance coefficient. The contribution of the scores on the individual components to the total test score was approached by a repeated measures regression analysis. This approach was preferred above correlation analyses since constructs of all three test components were completely different. In this step, the component scores were reduced to a score on 100 (%-score) to equally balance the contribution of each component to the final test score.

In the third step, to test validity, the total test scores of the students were ranked in four categories and compared to the students' master grades (expressed as %-scores) using a repeated measures regression technique. These master grades were the final marks students obtain when finishing their master graduate. In this step, the total test score was reduced to a score on 100 (%-score) to correctly compare with master grades (also expressed as %-score). To assess the relation between the individual test components and the master grades, a multiple regression analysis was performed. By taking this step, a more accurate profiling of the students was possible: it was expected that the master grade was in particular positively related to the EBM and the knowledge test score.

All quantitative analyses were performed with SAS 9.4 (SAS Institute Inc, USA).

The face validity of the test was qualitatively approached by a manual screening of the students by score quartile. In this step, teachers and head of departments were asked to identify, screen, and appreciate (interpret) the scores and ranking of their students. The assessment of the test scores of the students was made by a Likert scale: from fully unexpected result to fully expected result. A particular attention was paid to the students in the highest and the lowest quartile and to students scoring significantly better or worse on a single component of the test. Above, the scores of students known as "problem learners" were also assessed with particular attention.

Ethical approval

According to the Belgian legislature, no ethical approval is required when no patients are involved. Permission to perform the research was obtained from the deans, program directors, heads of department, appointed student representatives, and departmental staff (teachers and fellows). The full procedure was also subjected to the legal requirements of admission and selection of all four universities and in agreement with the federal legislation.

Results

In total, 322 out of 333 students completed the test. The descriptive analytics are presented in Table 1. All test scores were normally distributed.

In the upper quartile of scores on the full test, the mean students' master grade was 74%; in the second and third quartiles of scores, the mean master grade was 73%; and in the lowest quartile, the mean master grade was 69% [Table 2]. A regression analysis to express the relation between the master grades as a dependent variable and the quartiles of the test scores as independent variable showed a significant difference between the master grades over the quartiles ($F = 5.3$, $P = 0.01$). An analysis of variance showed that the differences were significant between the three upper quartiles and the lowest quartile [Table 2]. The other between-level differences were not significant.

A regression analysis measuring the relationship between the master grades and the test score revealed an odds ratio of 1.1, (95% CI 1.027–1.13). This means that an increase of 1 unit in the "master grade score" is followed by a 1.1-fold increase in the test score.

The relation of each individual component to the full test score was approached by a multiple regression analysis using the full test score as dependent variable. All three components, SJT, knowledge, and EBM, were positively and significantly related to the final test score (parameters estimate of 2.8; 1.1; 0.1, respectively, $P < 0.0001$) [Table 3].

The prediction of the scores on the individual test components by the master grades was also addressed by multiple regression analysis. The master grades score was considered as the dependent variable. Only the score on the knowledge test was significantly predicting the master grade (parameter estimate of 0.2, $P < 0.0001$) [Table 3].

A qualitative screening and appraisal of the test scores revealed

Table 1: Descriptive analysis of the test scores				
Analysis	Full test	SJT	Knowledge	EBM
Mean/total	288/406	214/281	68/114	6.5/11
Median	294	218	70	6
Min/max score	168/330	135/242	23/90	3/10
SD	25	16	12	1.7
SE of the mean	14	1	0.6	0.1
Coefficient of variance	8.7	7.6	17	26
Percentile 10/P95	261/316	198/232	51/81	4/9
Cronbach's alpha	NA	NA	0.6	NA

SJT: Situational judgment testing; EBM: Evidence-based medicine; NA: Not available; SD: Standard deviation; SE: Standard error

Table 2: Distribution of master grades by quartiles (from low to high) of test score and difference between master grade means (significance when*)

Quartile of test score	Number of students	Mean master grade score	SD	Minimum master grade	Maximum master grade
0	81	69.2	13.1	55.7	85.5
1	77	72.9	5.3	61.2	90.0
2	81	72.9	4.8	61.2	85.2
3	83	73.9	5.0	59.7	87.0
Quartile comparison	**Difference between master grade means**	**Simultaneous 95% confidence limits ($P<0.005$)**			
3-0	4.725	1.4-7.9*			
2-0	3.780	0.5-7.0			
1-0	3.748	0.4-7.1			

SD: Standard deviation, * level of significance 0.005

Table 3: Relationship between test components and full-test score and master grade (significance when*)

Test component score	Parameter estimates for full test score	P	Parameter estimates for master grade	P
SJT	2.8	<0.0001*	0.1	0.2
Knowledge	1.1	<0.0001*	0.2	<0.0001*
EBM	0.1	<0.0001*	0.1	0.02

SJT: Situational judgment testing; EBM: Evidence-based medicine, * level of significance 0.0001

no unexpected observations. The highest and lowest quartile of the total test score included all students who were respectively known as "very good" or "very poor." When students scored remarkably better on one component of the test, it particularly concerned knowledge test and SJT. The test results of five students, identified with high master examination grades but known with poorer social skills, showed a discrepancy between a high knowledge score and a low SJT score. The inverse was also perceived: a group of socially committed students scored high on the SJT component but lower on the knowledge test. In the first and third quartile of the test score, these observations are confirmed by two extreme examples: respectively, a high-performing student (overall high curricular grades with a master grade of 90) did not score well on the SJT while a poorer-performing student (master grade of 59.7) scored very well on the SJT and therefore ended up in the upper quartile of the total test score [Table 2].

Discussion

This study reported the results of validity and reliability analysis of a multicomponent proficiency test for admission to the advanced master in family medicine in Flanders, Belgium. The actual test is part of a three-phase inclusion procedure were succeeding on the multicomponent test (here reported) is rewarded with an admission to the advanced master.

The option to compose an admission test with three different assessment components was the result of a consensus reached by the education staff of the four departments of family medicine in Flanders.[12,17,18] This consensus was supported by a literature review, expert consulting, and the AMEE guideline on assessment. The primary objective was to develop a valid, reliable but also acceptable and feasible test. It is known that master grades are not reliable enough to orientate and select students for further medical specialization.[3,19] Hence, an admission test should go beyond the traditional competence assessment.[11,20] Indeed, according to the CANMED roles, a doctor is not only a medical expert. In particular family doctors have to master social competences to prove their talent in collaboration and in healthcare advocacy. The multicomponent admission test therefore contains besides a knowledge test also an EBM critical appraisal test and an SJT.[18,21] The EBM test assesses the clinical reasoning and the ability to search for evidence online on the spot. The SJT addresses moral, ethical, and psychosocial critical situations and assesses the ability of future doctors to react with empathy, professionalism and in agreement with the social context. Students' knowledge is comprehensively approached during the under- and post-graduate education phases but in a rather unidimensional way. In this proficiency test, knowledge was assessed by decision making in realistic case. Indeed, knowledge skills remain a sensitive and critical filter in high-stake decisions.[22]

To keep testing time low and to increase students' acceptability, the number of questions was set at the critical minimum. In a relatively understaffed department, the feasibility of the test setting was ensured by offering a machine-assisted test. Students passing this test were theoretically (except in this, not binding pilot) admitted to further education. Students who failed were redirected to a structured jury examination. This strategy reconciles the conflict between the large numbers of applicants and the low number of staff members with high-stake decisions: students will not be rejected only based upon a machine-assisted test result. Second, it is reassuring for both staff and students that the decision on admission was not made by a computer but finalized through a jury defense.

Reliability was not tested in a conventional way but estimated by the observation of a normal Gaussian distribution of the total test score and the individual component scores.[23] Only the knowledge test contained enough questions to obtain a reliable Cronbach's alpha. The decision to include a limited set of questions in each component was driven by concerns about the

testing time.[24] The total testing was set on 2 h to avoid fatigue and loss of concentration.[12] Multivariate analyses demonstrated that all three test components significantly contributed to the total test score. This is an important observation since it proves that no test component was favored over the other components. All students therefore started unbiased and without a prior benefit. Students with a brilliant curriculum started at the same level as the average or the poorer students. High curricular scores were therefore not indicative or beneficial. The second proof of reliability was obtained by comparing the test data to the master grades of the students. In the contemporary curriculum, the master grade score is composed of scores on structured jury examinations, knowledge testing, and workplace-based assessment. In particular, the first two components are highly awarded in the medical curriculum. It seems reasonable to assume that the students with higher test scores (upper three quartiles) performed with higher master grades and that the variance of grades in these quartiles was lower than in the lowest quartile.[25] The differences between the master grades of students were significant between the three upper quartiles and the lowest quartile. Furthermore, as expected, only the knowledge test component significantly correlated with the master grade of the students.[3] These observations emphasize the knowledge-based assessment in graduate medical education and stresses the need to explicitly test beyond this competence. Since the aim of an admission test is not to confirm earlier academic achievement, this admission test included situational judgment- and EBM-skills evaluation.[18] In particular, the best-performing medical students (thus with the highest master grades), scored well on all three components, while variance in the individual component test scores increased with decreasing master grades. A similar relation was observed in the group of underperforming students: they scored poorly in all three components. This means that the multicomponent test is capable to identify and confirm the highest and the poorest performing students but also to highlight competences as professionalism, attitude, empathy, reasoning, etc. Identifying gaps and shortcomings in these competences can be the base of a future learning agenda for the individual student. Finally, in the recalculations, it was decided to redistribute the weight of the scores on the components: 50% SJT, 30% knowledge, and 20% EBM.

The interesting part of this pilot study was that all students included, on a voluntarily but compelling base, were known by the teaching staff of the four involved departments. A manual screening of the test results was an opportunity to qualitatively assess and test the validity of the scores. In the lowest percentile of test scores, no unexpected results were observed. All students in this percentile were recognized as poor performing, but they already passed their master graduate examinations. The teachers confirmed that all students who were previously (on a formal base, during the graduate curriculum) identified as "at risk" were included in a rehearsal program or given particular attention by the head of the department. Further, during deliberation, there were recurring examples of mediocre-performing students, in terms of master grades, but with remarkable social skills

or other (noncognitive) favorable competences who ended up in quartile one or two (of four from 0 to 3). In the upper quartile, all the best-performing students were recognized. This means that the format of the admission test is particularly in favor of the students who have learning potential and succeeds in detecting the poorest performers where the master graduate (and undergraduate) exams fail in this purpose.

The weakness of this pilot study is mainly attributed to the apparently poor test psychometrics. Indeed, as mentioned above, the test reliability was estimated by the observation of a normal distribution of the test scores. Validity was approached by multivariate regression analyses and by extrapolation to the qualitative analyses. This regression technique was earlier successfully applied to an analysis of a complex OSCE.[23]

The strength of the research is that the results of the quantitative and the qualitative analyses mutually accorded. Although the researchers are aware of the risks of self-fulfilling prophecies (hoping that these poor-performing students also fail on the admission test), they believe that the screening and appraisal of the test scores were thoughtful and objectively performed and discussed. Second, the strength of the admission test lies in the reconciliation of feasibility and validity. Indeed, by offering a machine-assisted test followed by a structured jury examination for the students who failed, the burden on the education staff is reduced to the minimum.

In the future, this pilot needs both retesting and follow-up. Retesting will add to the reliability and will refine the test construction. Follow-up will address validity. Moreover, the students who did not pass the test in this pilot only received a negative advice. Most of them finally decided to continue the family medicine. This means that the researchers will be able to follow both groups over time.

Conclusion

In this pilot, a multicomponent machine-assisted admission test to family medicine education was successfully studied. The emphasis of assessment lied on situational judgment, knowledge and EBM appraisal. The overall and component test scores were in agreement with the performance and profiling of the participating students. The test succeeded in identifying the poor-performing students and in confirming and revealing the competences of the average- and high-performing students.

Conflicts of interest

There are no conflicts of interest.

References

1. Howe AC, Stokes-Lampard H, Rafi I, Baker M. Academic general practice: Supported by the RCGP. Br J Gen Pract 2016;66:14.

2. Gray DP. Academic general practice: A viewpoint on achievements and challenges. Br J Gen Pract 2015;65:e786-8.

3. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, *et al.* Assessment for selection for the health care professions and specialty training: Consensus Statement and Recommendations from the Ottawa 2010 Conference. Med Teach 2011;33:215-23.

4. Vermeulen MI, Kuyvenhoven MM, Zuithoff NP, Tromp F, van der Graaf Y, Pieters RH, *et al.* Selection for dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012;18:201-5.

5. Staten A. Getting the swagger back into general practice. Br J Gen Pract 2015;65:257.

6. van Oostrom SH, Picavet HS, de Bruin SR, Stirbu I, Korevaar JC, Schellevis FG, *et al.* Multimorbidity of Chronic diseases and health care utilization in general practice. BMC Fam Pract 2014;15:61.

7. Moth G, Vestergaard M, Vedsted P. Chronic care management in Danish general practice – A cross-sectional study of workload and multimorbidity. BMC Fam Pract 2012;13:52.

8. Mola E. Patient empowerment, an additional characteristic of the European definitions of general practice/family medicine. Eur J Gen Pract 2013;19:128-31.

9. Thistlethwaite JE, Kidd MR, Hudson JN. General practice: A leading provider of medical student education in the 21st century? Med J Aust 2007;187:124-8.

10. Myhre DL, Sherlock K, Williamson T, Pedersen JS. Effect of the discipline of formal faculty advisors on medical student experience and career interest. Can Fam Physician 2014;60:e607-12.

11. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: Implications for selection and recruitment. Med Educ 2008;42:1195-204.

12. Plint S, Patterson F. Identifying critical success factors for designing selection processes into postgraduate specialty training: The case of UK general practice. Postgrad Med J 2010;86:323-7.

13. Lacasse M, Théorêt J, Tessier S, Arsenault L. Expectations of clinical teachers and faculty regarding development of the CanMEDS-family medicine competencies: Laval developmental benchmarks scale for family medicine residency training. Teach Learn Med 2014;26:244-51.

14. Rhee SO. Factors determining the quality of physician performance in patient care. Med Care 1976;14:733-50.

15. Kreiter CD. A proposal for evaluating the validity of holistic-based admission processes. Teach Learn Med 2013;25:103-7.

16. Roberfroid DS, Camberlin C, Van de Voorde C, Vrijens F, Léonard C. Het Aanbod Van Artsen in België: Huidige Toestand en Uitdagingen. Health Services Research (HSR). Brussel: Federaal Kenniscentrum Voor de Gezondheidszorg (KCE); 2008. KCE Reports. 72A (D/2008/10.273/07). KCE Reports 47A; 2008.

17. Shumway JM, Harden RM; Association for Medical Education in Europe. AMEE Guide No 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 2003;25:569-84.

18. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ 2009;43:50-7.

19. Bodger O, Byrne A, Evans PA, Rees S, Jones G, Cowell C, *et al.* Graduate entry medicine: Selection criteria and student performance. PLoS One 2011;6:e27161.

20. Stern DT, Frohna AZ, Gruppen LD. The prediction of professional behaviour. Med Educ 2005;39:75-82.

21. Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M, *et al.* Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Med Educ 2012;46:399-408.

22. Eftekhar H, Labaf A, Anvari P, Jamali A, Sheybaee-Moghaddam F. Association of the pre-internship objective structured clinical examination in final year medical students with comprehensive written examinations. Med Educ Online 2012;17:15958 - http://dx.doi.org/10.3402/meo.v17i0.15958.

23. Schoenmakers B, Wens J. The objective structured clinical examination revisited for postgraduate trainees in general practice. Int J Med Educ 2014;5:45-50.

24. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: How should the components be combined to maximize reliability? Med Educ 2001;35:326-30.

25. Haldane T, Shehmar M, Macdougall CF, Price-Forbes A, Fraser I, Petersen S, *et al.* Predicting success in graduate entry medical students undertaking a graduate entry medical program. Med Teach 2012;34:659-64.