Evaluation of The Medical Board Exam in Portugal

Avaliação do Exame Final do Internato Médico em Portugal

Tiago Reis MARQUES✉*[1], Inês LAÍNS*[2,3,4], Maria João MARTINS[5,6], Francisco GOIANA-DA-SILVA[7], Filipa SAMPAIO[8,9], Inês PESSANHA[10], Diogo Hipólito FERNANDES[11], Mariana BRANDÃO[12,13], Pedro PINTO TEIXEIRA[14], Manuel de OLIVEIRA SANTOS[4,15], João Carlos SILVA[16], João Carlos RIBEIRO[4,17,18]

Acta Med Port 2018 Nov;31(11):670-679 ▪ https://doi.org/10.20344/amp.10646

ABSTRACT

Introduction: There is a high heterogeneity in the structure of postgraduate medical training evaluation worldwide. However, in contrast to other countries, there have been no scientific studies of the final medical board examination, in Portugal. The present study aimed to evaluate the adequacy of the medical board examination including its validity as measured by its association with medical school grade average and national seriation examination.

Material and Methods: Cross-sectional, observational study. We analyzed the final results on the medical board examination of 2439 physicians, across 47 specialties, who completed their training in 2016 and 2017, using measures of central tendency and variability. We assessed the association between these grades and the national exam to initiate residency, and the grade average in Medical School.

Results: Measures of central tendency and variability, and consequent shape measures, revealed that the distribution of the scores of the final medical board exam is extremely negatively asymmetric and leptokurtic. A positive association was also found between the results in this exam and the score on national exam to initiate residency, and the grade average in Medical School.

Conclusion: Although the medical board examination was, in general, positively associated with scores on the national exam to initiate residency, and the mean final Medical School grades, thus indicating its potential validity, our results demonstrate that this exam presents no satisfactory discriminative capacity. Therefore, there is room to improve the actual postgraduate medical examination model, including changes in its classification system and potentially consider other assessment models.

Keywords: Education, Medical, Graduate; Educational Measurement; Internship and Residency; Models, Educational; Schools, Medical

RESUMO

Introdução: Existe uma elevada heterogeneidade na estrutura da avaliação da formação médica pós-graduada a nível mundial. No entanto, contrastando com outros países, não existem estudos científicos em Portugal que tenham avaliado o modelo da avaliação final da especialidade. O presente estudo pretendeu avaliar a adequação do exame do final da especialidade aos seus propósitos; aí incluída a sua validade enquanto consubstanciada na relação com a prova nacional de seriação e média final de curso de medicina.

Material e Métodos: Estudo transversal, observacional. Foram analisadas com recurso a medidas de tendência central e variabilidade, as notas na avaliação final da especialidade de 2439 médicos, de 47 especialidades, que terminaram a sua formação em 2016 e 2017. Tendo em vista a sua validação cruzada, foram também avaliadas as correlações com a média final de curso e a nota na prova nacional de seriação.

Resultados: Das medidas de tendência central e variabilidade, e consequentes medidas de formato, resulta que a distribuição das pontuações do exame final de especialidade se apresenta com uma forma manifestamente assimétrica negativa e leptocúrtica. No geral, verificou-se a existência de uma associação positiva entre a avaliação final da especialidade e a média de curso e a prova nacional de seriação.

Conclusão: Apesar de positivamente associado, no geral, com a média de curso e a prova nacional de seriação, o que indica a sua potencial validade, os resultados demonstram que a avaliação final de especialidade não apresenta uma capacidade discriminativa

* Co-Primeiros Autores

1. Department of Psychosis Studies. Institute of Psychiatry, Psychology and Neuroscience. King´s College. London. United Kingdom.
2. Massachusetts Eye and Ear. Harvard Medical School. Boston. United States.
3. Centro Hospitalar e Universitário de Coimbra. Coimbra. Portugal.
4. Faculdade de Medicina. Universidade de Coimbra. Coimbra. Portugal.
5. Centro de Investigação em Neuropsicologia e Intervenção Cognitivo-Comportamental. Faculdade de Psicologia e Ciências da Educação. Universidade de Coimbra. Coimbra. Portugal.
6. Instituto de Psicologia Médica. Faculdade de Medicina. Universidade de Coimbra. Coimbra. Portugal.
7. Department of Surgery and Cancer. Imperial College Medical School. London. United Kingdom.
8. Serviço de Oftalmologia. Hospital Pedro Hispano. Senhora da Hora. Portugal.
9. Unidade Local de Saúde de Matosinhos. Matosinhos. Portugal.
10. Serviço de Cirurgia Pediátrica. Centro Hospitalar e Universitário de Coimbra. Coimbra. Portugal.
11. Serviço de Oftalmologia. Centro Hospitalar de Lisboa Central. Lisboa. Portugal.
12. Institut Jules Bordet et L'Université Libre de Bruxelles (U.L.B.). Brussels. Belgium.
13. Instituto de Saúde Pública. Universidade do Porto. Porto. Portugal.
14. Serviço de Cardiologia. Centro Hospitalar de Lisboa Central. Lisboa. Portugal.
15. Serviço de Cardiologia. Centro Hospitalar e Universitário de Coimbra. Coimbra. Portugal.
16. Serviço de Gastroenterologia. Centro Hospitalar Vila Nova de Gaia/Espinho. Porto. Portugal.
17. Serviço de Otorrinolaringologia. Centro Hospitalar e Universitário de Coimbra. Coimbra. Portugal.
18. Coimbra Institute for Clinical and Biomedical Research (iCBR). Coimbra. Portugal.

✉ Autor correspondente: Tiago Reis Marques. tiago.marques@kcl.ac.uk

Recebido: 11 de abril de 2018 - Aceite: 07 de novembro de 2018 | Copyright © Ordem dos Médicos 2018

satisfatória. Deste modo, existe oportunidade para melhoria do modelo atual, nomeadamente através da alteração ao seu sistema de classificação e considerando outros modelos de exame.

Palavras-chave: Avaliação Educacional; Educação Médica Pós-Graduada; Escolas Médicas; Internato Médico; Modelos Educacionais

INTRODUCTION

Widely varying structures of postgraduate medical training and its evaluation can be found worldwide. Overall, there is an increasing use of competency-based education,[1] as well as the need for frequent knowledge assessments aimed at the promotion of better retention of information (test-enhanced learning).[2] In Portugal, the assessment of the medical board exam (MBE) (*exame final do internato médico*) is aimed to serve both as a certification test for specialties as well as a seriation measure for the hiring of new consultants in calls for positions within the public healthcare system. Therefore, this test has a significant impact on the professional perspectives of physicians working in Portugal.

The final evaluation of the MBE is currently defined by the *Decreto-lei* No. 13/2018, regulated by the *Portaria* 79 of 16 Mar 2018. Each registrar is admitted to a final evaluation regarding the whole training process, upon submission to positive performance and knowledge evaluations at each residency. Three public and eliminatory tests are included in this evaluation: curriculum analysis, practical and theoretical tests. The curriculum analysis consists of the evaluation and discussion of the *curriculum vitae*. An evaluation grid for the curriculum evaluation by the jury is usually recommended by each specialty college within the Portuguese Medical Association (*Ordem dos Médicos*). Problem-solving skills and response to situations within the specialty are evaluated by the practical test, usually including the examination of a patient, medical history and discussion or case analysis, including a final report which is usually discussed. Knowledge evaluation of the candidate is obtained through the theoretical test, which is usually an oral test, even though it is replaced by a written or multiple-choice test in some specialties, in which case it is a national test and is completed by all the candidates at the same time. The final score corresponds to the arithmetic average of all the scores in the three tests, evaluated on a 0-20 scale. This examination currently takes place within two annual seasons and different juries are involved in each season.

An exam should ideally be aimed at knowledge evaluation and also at its correct practical application.[3] In fact, the format of the final assessment of each specialty has been widely debated by physicians and has been empirically criticized due to its subjectivity, to the fact that high marks were obtained by most candidates and to the almost complete absence of fails. However, no data regarding the validity or discriminatory ability of this examination were ever published.

The following should be available to the analysis of the results of a knowledge test:

a) The major measures of central tendency - mean, mode and median- providing information on data distribution, as well as the most representative or central score. It is argued that an evaluation knowledge test should follow a normal distribution (i.e. a bilaterally symmetrical curve, divided in two, in which each half contains 50% of data and mean, median and mode are together in the centre of the distribution)[4];
b) The measures of variability – range, variance, interquartile range and standard deviation, allowing for the description of data distribution by the possible values. Usefulness is assessed by variability and the higher the variability, the more ability of a test to distinguish between subjects. The evidence of the validity of a test can be measured considering different sources, one of which is the analysis of patterns of convergence and divergence. This evidence of validity is obtained by the analysis of the associations between the results of a test and those in tests aimed at the measurement of the same construct or similar constructs.[4]

Internationally, the evaluation of postgraduate medical training is usually developed according to these characteristics and results are regularly published and available for analysis.[5-11] The associations between marks obtained in residencies and those obtained in the MBE have also been analysed.[12] Within each specialty, the predictors of success for admission to the college have also been analysed[13] and the marks obtained in evaluations before the MBE are also included.[14-16]

However, only one Portuguese study was aimed at the definition of a relationship between different evaluations throughout medical training, showing positive associations between the grade point average (GPA) (*média final do curso de medicina*) and the score obtained in the national seriation examination (NSE) (*prova nacional de seriação*).[17] Therefore, there are no studies aimed at the analysis of the relationship between the score obtained in different evaluations throughout the medical training.

This study was aimed at the assessment of the accuracy/reliability of the marks obtained in the MBE, through the assessment of the mark's distribution, as well as its validity, through the association with other performance measures available for public consultation, namely with the marks obtained in the NSE and the GPA.

MATERIAL AND METHODS

This was a cross-sectional observational study.

Participants

Physicians having completed specialty training in Portugal in 2016 and 2017 and their marks in the MBE for each specialty, obtained from the lists officially published by the *Administração Central do Sistema de Saúde (ACSS)* were included in the study. These were the final scores, corresponding to the arithmetic average score in all the tests. Considering that these marks were only available from 2016 onwards, only physicians having completed the exam from that date onwards were included. The remaining data, including the GPA and the mark in the NSE (i.e. the examination giving access to the specialty) were also available and were obtained from the official lists in the site of the ACSS.

The participants with available information on the three variables (MBE, GPA and NSE marks) were included in the study.

Procedure and statistical analysis

SPSS version 24.0 software was used for the statistical analysis and a p-value < 0.05 was considered as statistically significant.

The following were used, in order to estimate the accuracy/reliability of MBE marks: a) measures of central tendency (mean, median and mode); b) measures of variability (range, standard deviation, variance and interquartile range) and c) measures of shape (kurtosis and skewness). Field's criteria (2005) were considered for the interpretation of kurtosis and skewness[18] in which data distribution follows a normal distribution when z values (obtained through the values of skewness and kurtosis divided by the standard error values) range between -1.96 and 1.96. Kolmogorov–Smirnov test, allowing for the identification on whether the data distribution is statistically different from a normal distribution, was also obtained. Histograms with the normality curve are shown for the total sample and by specialty (Appendix 1: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/10646/Apendice_01.pdf). Diagrams of extremes and quartiles (boxplot) are also shown, corresponding to the variability of the marks by specialty (Appendix 2: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/10646/Apendice_02.pdf).

The validity of the marks obtained in the MBE was assessed through the correlation of these with GPA and NSE marks by using Spearman's correlation coefficient (considering that the variables that were analysed did not follow a normal distribution), with the whole sample and by specialty. Considering the variability in the number of participants by specialty and in order to ensure statistical accuracy and the representativeness of the results, only those specialties in which data on 10 or more participants were available were analysed. Cohen's criterion (1988) was used for the evaluation of the dimension of correlations (1988),[19] in which a value of 0.10 corresponds to a small effect, 0.30 to a medium and 0.50 to a large effect. Therefore, the following ranges were considered: 0.10 (inclusive) to 0.30 (exclusive) corresponds to a weak correlation, 0.30 (inclusive) to 0.50 (exclusive) to a moderate correlation and > 0.50 (inclusive) to a strong correlation. Moderate to strong positive correlations between MBE and previous marks (GPA and NSE) were looked for in order to assess the validity of MBE marks. Dispersion charts of these relations are shown as complementary material (total sample).

RESULTS

Group of participants

Data regarding 2,439 physicians were included in the study and their distribution is shown in Table 1.

Variability in marks obtained in the MBE

MBE marks ranged between 10.8 and 20 with an 18.9 median score (Fig. 1). The distribution curve with the marks of all the participants regardless of the specialty was an asymmetrical curve, with most participants having obtained high marks, i.e. with a leptokurtic and negative skewed distribution. However, when looking at the distribution curves by specialty (complementary material) some specialties have shown an approximately symmetrical distribution curve (such as those representing the marks regarding Gastroenterology, Neurology and Public Health), even though this pattern was shown by most participants,

The distribution and normality of these data is shown in Table 2. The measures of central tendency (mean = 18.5; mode = 19.5 and median =18.9) have shown values close to the maximum score (20). The measures of variability (standard deviation = 1.3; variance = 1.6) have corresponded to a low variability of the results. The measures of shape and the analysis of normality have shown that the distribution of MBE marks does not follow a Gaussian/normal distribution (p < 0.001).

The diagrams of extremes and quartiles (boxplot) are shown in Fig. 2, corresponding to the variation of MBE marks by specialty (only those specialties with over ten participants were analysed).

Relationship between GPA, NSE and MBE marks

A significant positive and moderate correlation (r_s = 0.42, p < 0.001) between GPA and MBE marks has been found when the whole group of participants was considered. In addition, a significant, positive and strong correlation

Table 1 – Number and percentage of physicians, by specialty (n = 2,439)

	2016 n	2017 n	Total n	Total %*
Pathology	11	18	29	1.2
Anaesthesiology	65	60	125	5.1
Cardiology	25	23	48	2.0
Paediatric Cardiology	3	3	6	0.2
General Surgery	33	59	92	3.8
Oral & Maxillofacial Surgery	5	3	8	0.3
Paediatric Surgery	4	2	6	0.2
Plastic Surgery	7	10	17	0.7
Cardiothoracic Surgery	7	7	14	0.6
Vascular Surgery	7	8	15	0.6
Dermatology	9	5	14	0.6
Endocrinology & Diabetes	13	13	26	1.1
Dental & Oral Medicine	6	5	11	0.5
Gastroenterology	22	25	47	1.9
Medical Genetics	2	3	5	0.2
Gynaecology / Obstetrics	42	54	96	3.9
Haematology	10	15	25	1.0
Transfusion Medicine	5	12	17	0.7
Allergy Medicine	1	9	10	0.4
Infectious Diseases	10	14	24	1.0
Sport & Exercise Medicine	0	2	2	0.1
Occupational Health	0	4	4	0.2
Rehabilitation Medicine	24	3	27	1.1
Family Medicine	356	362	718	29.4
Internal Medicine	147	156	303	12.4
Forensic & Legal Medicine	5	6	11	0.5
Nuclear Medicine	2	3	5	0.2
Renal Medicine	13	18	31	1.3
Neurosurgery	6	8	14	0.6
Neurology	21	12	33	1.4
Neuroradiology	7	9	16	0.7
Ophthalmology	33	20	53	2.2
Medical Oncology	30	30	60	2.5
Orthopaedics	28	41	69	2.8
Audiovestibular Medicine	18	26	44	1.8
Clinical Pathology	11	13	24	1.0
Paediatrics	71	45	116	4.8
Child & Adolescent Psychiatry	9	15	24	1.0
Respiratory Medicine	16	16	32	1..3
Psychiatry	45	36	81	3.3
Radiology	24	25	49	2.0
Radiotherapy	9	4	13	0.5
Rheumatology	10	11	21	0.9
Public Health	10	19	29	1.2
Urology	12	13	25	1.0
Total	1,194	1,245	2439	

* This percentage refers to physicians included in the study (having completed the MBE in 2016 and 2017). Reliable national data regarding the percentage of physicians in each specialty are currently not available.

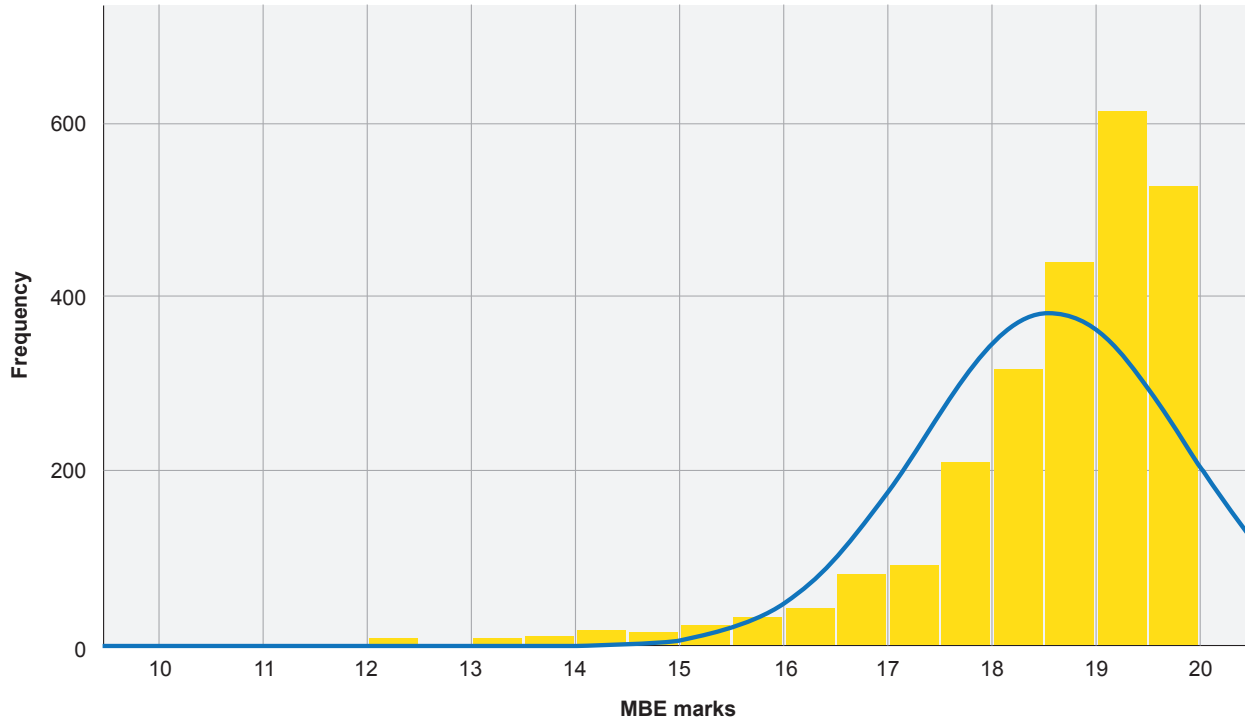

Figure 1 – MBE marks regarding the whole sample (n = 2,439)

between NSE and MBE marks has been found (r_s = 0.59, p < 0.001). The same analysis was made by specialty and is shown in Table 3. Widely varying scores were found by specialty. Significant correlations in both analyses were found in some specialties, such as Pathology, Vascular Surgery or Gynaecology/Obstetrics, while significant correlations were only found in the analysis of the relationship between MBE and GPA marks in Anaesthesiology, Paediatrics and Psychiatry. Significant correlations between GPA and NSE marks were found in Cardiology, General Surgery or Pathology. No significant correlations in any of the associations were found in Plastic Surgery, Cardiothoracic Surgery, Dermatology or Forensic and Legal Medicine. These were the specialties with the smallest groups of participants (n < 30, except Audiovestibular Medicine and Respiratory Medicine). Even considering specialties showing significant associations within the same analyses, a widely varying dimension of these correlations has been found. For instance, weak (Paediatrics), moderate (Internal Medicine) and strong associations between GPA and MBE marks (Vascular Surgery, for instance) have been found, in line with what was found between MBE and NSE marks, showing a small (ex.: Orthopaedics), medium (Internal Medicine) and large effect size (Transfusion Medicine). Larger effect sizes have been found as regards this relationship and stronger correlations have generally been found as regards the relationship between MBE and NSE marks.

DISCUSSION

Postgraduate medical training is crucial and a quality assurance. A correct knowledge evaluation is a major element of this process and scientific evidence-based examinations and evaluation procedures are therefore necessary. Tests with strong psychometric properties are valid considering different factors: content (representativeness of the items regarding the domain to be tested), relationship with other variables (convergent, divergent and predictive validity),

Table 2 – Measure of central tendency, variability, shape and analysis of normality of the distribution of 'specialty evaluation' variable

Measures of central tendency			Measures of variability				Measures of shape				Normality analysis
Mean	Mode	Median	*SD*	Range	Variance	Interquartile range (25; 50; 75)	Skewness (SE)	z	Kurtosis (SE)	z	K-S
18.52	19.50	18.90	1.27	10.8 - 20	1.61	18.10 18.90 19.40	-2.047 (0.05)	-40.94	5550 (0.10)	55.5	0.144***

SD: standard deviation; SE: standard error; K-S: Kolmogorov-Smirnov test; ***p < 0.001

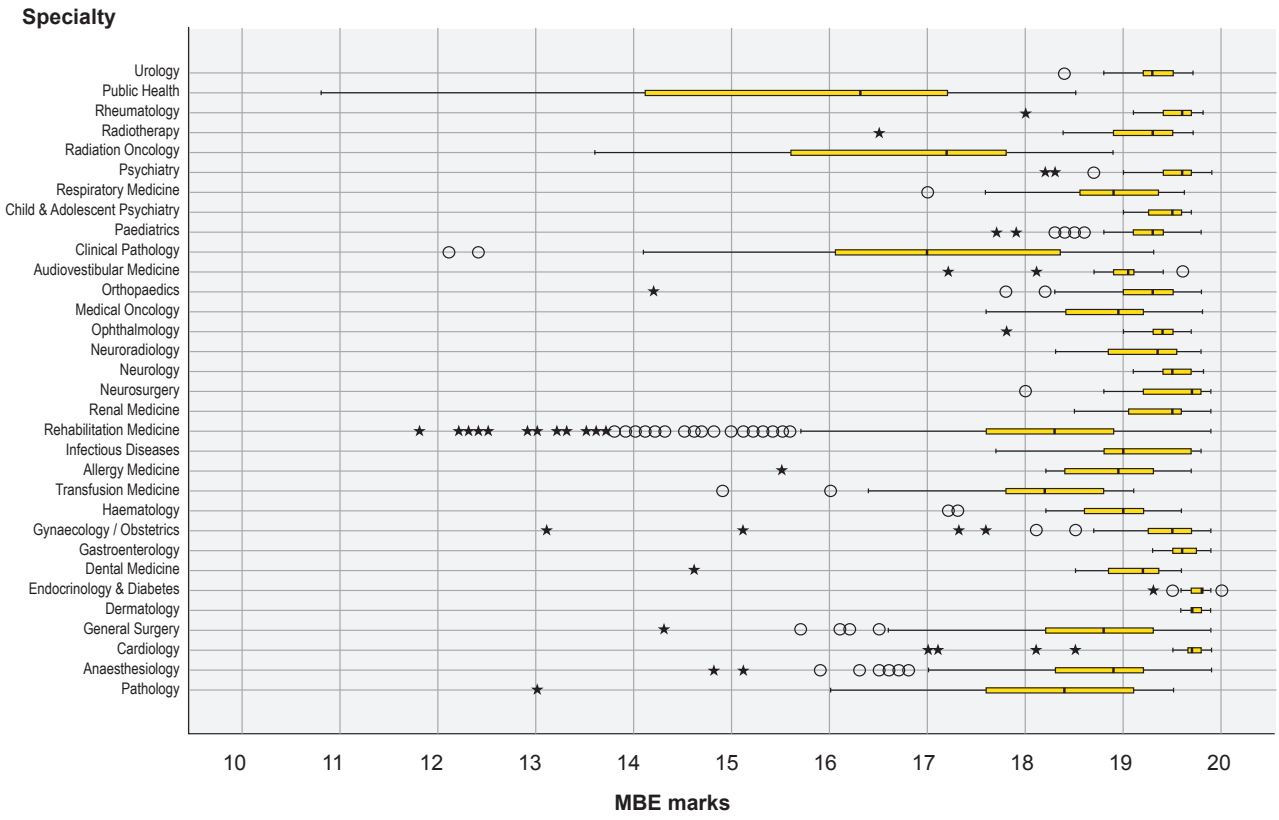

Figure 2 – Diagrams of extremes and quartiles regarding MBE marks, by specialty

among others.[20] Reliability is also a major characteristic of the tests with good psychometric properties and internal coherence is used (assessing whether the different items of a test are measuring the same construct). However, multiple assessment data are needed throughout medical training for the full assessment of these variables (validity and reliability) (multi-item questionnaires, for instance). This study was aimed at the evaluation of the reliability of MBE through the analysis of the distribution of the marks, as well as its validity through the analysis of the association between the results in this test with other performance measures (GPA and NSE marks).

The use of measures of central tendency and variability of score distribution, as described, is involved in the accuracy of an evaluation test. Our results have shown that MBE marks did not show a normal distribution and measures of shape have shown a negatively skewed distribution, i.e. most participants have obtained high marks. The measures of central tendency (mean, mode and median) were very close to the maximum possible score (20 points), with an extremely low standard deviation, showing that marks were very close to the mean (18.5).

These results have a major impact on the way MBE marks are understood and on possible further changes. Considering the results, namely the poor discriminatory ability of the current examination, this need to be adapted so that most results will remain within the centre of the distribution, namely its system of classification should be changed or other models of examination should be considered. This is particularly important considering that the progression of physicians through their professional career within the Portuguese National Health System (SNS) is currently highly dependent on the marks obtained in this examination. These are in fact one of the major parameters that is considered in public tenders for those willing to work in the SNS as consultants.[20]

Good practice in use in other countries are reflected in the presence of structures aimed at the orientation and certification of internship programs, including evaluation methods (Accreditation Council for Graduate Medical Education in the United States of America; General Medical Council in the United Kingdom). These have been widely discussed in literature.[21-23] According to these policies, different recommendations would improve the evaluation of the Portuguese postgraduate training. Selected-response tests (multiple-choice items, for instance) have been described as an efficient, versatile and direct way of knowledge evaluation due to their advantages in terms of objectivity regarding its scoring and psychometric properties (validity and internal coherence, for instance).[24] However, major limitations regarding the evaluation of practical knowledge can exist. Therefore, the use of examinations with standardised clinical cases can be necessary, common to all the candidates within the same specialty.[25] A possible change in the

Table 3 – Mean, standard deviation, median, percentiles (25th and 75th) and evaluation of the correlation between GPA and NSE marks with MBE marks

	n	M (SD)	Median	P25	P75	GPA *vs.* MBE r_s	NSE *vs.* MBE r_s
Pathology	29	18.01 (1.37)	18.40	17.45	19.10	**0.44***	**0.69****
Anaesthesiology	125	18.61 (0.95)	18.90	18.30	19.20	**0.23***	0.05 ns
Cardiology	48	19.56 (0.61)	19.70	19.63	19.80	0.14 ns	**0.49****
General Surgery	92	18.68 (0.89)	18.85	18.20	19.30	0.14 ns	**0.25***
Plastic Surgery	17	18.61 (0.67)	18.50	18.00	19.40	-0.11 ns	0.18 ns
Cardiothoracic Surgery	14	18.13 (1.05)	18.30	17.80	18.83	0.47 ns	0.33 ns
Vascular Surgery	15	19.17 (0.49)	19.30	19.00	19.50	**0.77****	**0.61***
Dermatology	14	19.74 (0.09)	10.70	19.70	19.80	-0.05 ns	0.05 ns
Endocrinology & Diabetes	26	19.77 (0.15)	19.80	19.70	19.83	0.04 ns	0.23 ns
Dental Medicine	11	18.75 (1.42)	19.20	18.50	19.50	0.45 ns	**0.68***
Gastroenterology	47	19.64 (0.15)	19.60	19.50	19.80	0.11 ns	**0.32***
Gynaecology / Obstetrics	96	19.30 (0.89)	19.50	19.23	19.70	**0.30****	**0.33****
Haematology	25	18.84 (0.61)	19.00	18.50	19.30	0.17 ns	**0.40***
Transfusion Medicine	17	17.94 (1.16)	18.20	17.60	18.80	**0.69****	**0.78*****
Allergy Medicine	10	18.63 (1.20)	18.95	18.35	19.33	-0.07 ns	**-0.69***
Infectious Diseases	24	19.07 (0.63)	19.00	18.80	19.70	0.39 ns	0.27 ns
Rehabilitation Medicine	27	18.94 (0.58)	19.10	18.60	19.30	**0.52****	**0.49***
Family Medicine	718	17.83 (1.28)	18.20	17.50	18.60	**0.48*****	**0.54*****
Family Medicine	303	18.49 (1.13)	18.80	18.00	19.30	**0.35*****	**0.45*****
Forensic & Legal Medicine	11	17.25 (1.29)	17.80	16.50	18.10	-0.16 ns	0.33 ns
Renal Medicine	31	19.32 (0.44)	19.50	19.00	19.60	**0.56****	**0.46****
Neurosurgery	14	19.45 (0.56)	19.70	19.10	19.83	0.39 ns	**0.59***
Neurology	33	19.50 (0.19)	19.50	19.40	19.70	**0.45****	**0.51****
Neuroradiology	16	19.23 (0.48)	19.35	19.78	19.58	0.24 ns	0.01 ns
Ophthalmology	53	19.33 (0.34)	19.40	19.30	19.50	**0.40****	**0.31***
Medical Oncology	60	18.84 (0.57)	18.95	18.40	19.20	0.25 ns	**0.50*****
Orthopaedics	69	19.10 (0.73)	19.30	19.00	19.50	0.08 ns	**0.27***
Audiovestibular Medicine	44	18.97 (0.38)	19.05	18.90	19.10	0.09 ns	-0.08 ns
Clinical Pathology	24	16.86 (1.97)	17.00	16.03	18.38	0.32 ns	**0.52****
Paediatrics	116	19.22 (0.35)	19.30	19.10	19.40	**0.22***	0.17 ns
Child & Adolescent Psychiatry	24	19.41 (0.23)	19.50	19.23	19.60	0.16 ns	0.24 ns
Respiratory Medicine	32	18.86 (0.61)	18.90	18.53	19.38	0.21 ns	0.31 ns
Psychiatry	81	19.49 (0.35)	19.60	19.40	19.70	**0.33****	0.09 ns
Radiology	49	16.79 (1.46)	17.20	15.45	17.85	0.26 ns	0.10 ns
Radiation Oncology	13	19.00 (0.84)	19.30	18.85	19.50	**0.66***	0.39 ns
Rheumatology	21	19.49 (0.39)	19.60	19.40	19.70	0.36 ns	0.15 ns
Public Health	29	15.56 (2.13)	16.30	14.05	17.30	0.32 ns	0.20 ns
Urology	25	19.29 (0.28)	19.30	19.20	19.50	0.23 ns	0.32 ns

n: number; M: mean; SD: standard deviation: r_s: Spearman's correlation coefficient * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns: vnot significant; p-values < 0.05 are presented in bold

current time-based approach to skill acquisition towards an objective structured approach in the future will eventually produce improvements in general and individual learning curve of consultants.[26] This will have an impact on the average quality of postgraduate training, as well as on training capability. Additional options include (i) the development of an evaluation structure for the supervision of postgraduate medical evaluation; (ii) support documentation and training material; (iii) re-design of the method of evaluation and finally (iv) the development of a network of evaluators under continuous training and with skills aimed at the development of examinations. On the other hand, different skills well beyond the academic knowledge can be involved in a medical board exam. Therefore, current practical assessment, which is developed on a non-systematic basis, without an evaluation by a serial list, could not fully assess the quality of a candidate.[27] Minimal skills required to be completed by a junior consultant should be recommended by each specialty. According to these characteristics, different models of evaluation can be recommended and their effectiveness evaluated including standard patients and checklists.

In medicine, validity has been evaluated through the comparison with other examinations and/or with the results in clinical practice.[28] However, there were no studies in Portugal showing the association between different evaluation moments throughout the academic and professional career. The association between MBE marks and those previously obtained (GPA and NSE) has been analysed in this study and, considering the whole group of participants, positive and moderate associations have been found, corresponding to an adequate validity, on a preliminary basis. However, when the results by specialty were analysed, these were widely varying, in line with international studies.[29] Considering the analysis already made in terms of variability of the marks, the absence of significant correlations or the presence of weak correlations in some specialties can be due to the narrow range of variability of this classification, not necessarily corresponding to a real absence of correlation. Due to the nature of the current MBE and to the scarce published data, any evaluation on its reliability as well as on its validity was not possible in a complete way. Likewise, no comments regarding the ability of the exam to evaluate knowledge and skill acquisition throughout the internship can be expressed. However, the low failure rate is surprising when compared to other countries, in which these vary between 60 and 90%. Although no definitive conclusions can be drawn as regards the validity of the exam, the results of the correlational analysis have underlined its potential validity. Despite the wide variability between specialties, participants with higher GPA and NSE marks also seemed to obtain higher MBE marks.

Some limitations are worth mentioning. First of all, only marks obtained in MBE on two years previous to this study were considered. Therefore, any generalisation of these results can be questionable, even though there are no data suggesting significant changes in MBE marks during the study period. Therefore, global results are expected to overlap those obtained in the past. Secondly, the cross-sectional nature of the study does not allow for any causality from the associations that were found. Therefore, these associations can be due to co-variables which were not considered in this study. Thirdly, due to the limited size of the sample, no data analysis by venue of the MBE has been possible. The fact that different examining juries were involved at the same time, within the same specialty, could have an impact on a possible variability and is worth mentioning. The influence of inter-examiner variability in medicine, considering different evaluation methods, is well known and studied.[30] In fourth place, an additional bias can be induced by the fact that MBE is not precisely the same for all specialties. In addition, this study was aimed at the evaluation of MBE marks, showing its distribution (in general and by specialty) rather than the comparison of marks in different specialties with each other. Even though the comparison between specialties is important, further studies with a greater number of participants will certainly address this subject. In fifth place, the limitations due to the type of evaluation should always be considered. It is well known that exams with these characteristics do not measure human, social and even technical skills of the candidates.[31] In sixth place, the fact that NSE and GPA marks were used to study the validity of MBE marks is on its own a limitation as these are measures of constructs that can be related, although eventually different (and measured at different times). Finally, the fact that marks obtained in each of the three moments was not made available to the authors is worth mentioning as simply using the final mean score is on its own a limitation of the study. The reliability and validity of each component of evaluation can be analysed in further studies, providing for a more complete evaluation of the MBE as related aspects even though different are evaluated by these components.

In the future, the analysis of validity, reliability, cost-effectiveness and acceptability of the MBE will be relevant, whenever multiple assessment elements throughout postgraduate training are available. The recommendation regarding the lack of validity and discrimination regarding the final test should also be applied to pregraduate training, in which there is a wide diversity regarding the evaluation methods and even curricular units. To the best knowledge of the authors, a systematic analysis of the quality of internal evaluations is currently developed by the Faculty of Medicine of the *Universidade do Porto*, the *Universidade do Minho* and the *Universidade da Beira Interior*. The expected changes in the model of the NSE in the short-term could

eventually be reflected on the internal models of evaluation in each of the Portuguese medical schools, as well as on the model of the MBE.

CONCLUSION

Considering the low variability in MBE marks, no satisfactory discriminatory ability of the current model of MBE in Portugal seem to exist, based on the results of this study. However, in general as in some specialties, a significant association between the MBE marks and the GPA has been found, on one hand, as well as with the NSE, on the other. These results seem to show the potential validity of the MBE when its relationship with these two previous evaluations is considered.

The results of this study suggested that the current MBE needs to be reconsidered, through a re-design and/or the implementation of more objective evaluation methods, allowing for better knowledge evaluation and candidate seriation.

HUMAN AND ANIMAL PROTECTION

The authors declare that the followed procedures were according to regulations established by the Ethics and Clinical Research Committee and according to the Helsinki Declaration of the World Medical Association.

DATA CONFIDENTIALITY

The authors declare that they have followed the protocols of their work centre on the publication of patient data.

CONFLICTS OF INTEREST

The authors declare that there were no conflicts of interest in writing this manuscript.

FINANCIAL SUPPORT

The authors declare that there was no financial support in writing this manuscript.

REFERENCES

1. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;27;32:676–82.
2. Larsen DP, Butler AC, Roediger HL. Test-enhanced learning in medical education. Med Educ. 2008;42:959–66.
3. Lineberry M, Matthew Ritter E. Psychometric properties of the Fundamentals of Endoscopic Surgery (FES) skills examination. Surg Endosc. 2017;10;31:5219–27.
4. Urbina S. Essentials of psychological testing. Vol. 53. New Jersey: John Wiley & Sons, Inc; 2014.
5. Unwin E, Potts HW, Dacre J, Elder A, Woolf K. Passing MRCP (UK) PACES: a cross-sectional study examining the performance of doctors by sex and country. BMC Med Educ. 2018;6;18:70.
6. Farooq S. High failure rate in postgraduate medical examinations - sign of a widespread disease? J Pak Med Assoc. 2005;55:214–7.
7. Bowhay AR, Watmough SD. An evaluation of the performance in the UK Royal College of Anaesthetists primary examination by UK medical school and gender. BMC Med Educ. 2009;29;9:38.
8. Dewhurst NG, McManus C, Mollon J, Dacre JE, Vale AJ. Performance in the MRCP(UK) Examination 2003–4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. BMC Med. 2007;3;5:8.
9. Rushd S, Landau AB, Lindow SW. An evaluation of the first time performance of international medical graduates in the MRCOG Part 1 and Part 2 written examinations. Eur J Obstet Gynecol Reprod Biol. 2013;166:124–6.
10. Membership of the Royal Colleges of Physicians of the United Kingdom. Pass rates for MRCP(UK) Diploma and Specialty Certificate Examinations. [consultado 2018 jun 04]. Disponível em: https://www.mrcpuk.org/mrcpuk-examinations/results/exam-pass-rates.
11. American Board of Internal Medicine. Residency Program Pass Rates 2015 – 2017. [consultado 2018 jun 04]. Disponível em: http://www.abim.org/~/media/ABIM Public/Files/pdf/statistics-data/residency-program-pass-rates.pdf.
12. Baker K, Sun H, Harman A, Poon KT, Rathmell JP. Clinical performance scores are independently associated with the American board of anesthesiology certification examination scores. Anesth Analg. 2016;122:1992–9.
13. Muensterer OJ, Bronstein ME, Mackenzie R, Snyder CW, Carachi R. Factors associated with passing the European Board of Paediatric Surgery Exam. Pediatr Surg Int. 2015;31:671–6.
14. Puscas L. Otolaryngology resident in-service examination scores predict passage of the written board examination. Otolaryngol - Head Neck Surg. 2012;147:256–60.
15. McClintock JC, Gravlee GP. Predicting success on the Certification Examinations of the American Board of Anesthesiology. Anesthesiology. 2010;J112:212–9.
16. Monteiro K, George P, Dollase R, Dumenco L. Predicting United States Medical Licensure Examination Step 2 clinical knowledge scores from previous academic indicators. Adv Med Educ Pract. 2017;8:385–91.
17. Martins IP. Admissão ao internato complementar em Portugal: análise dos resultados do exame nacional de seriação entre 2006 e 2011. Acta Med Port. 2013;26:569–77.
18. Field A. Discovering statistics using SPSS. London: Sage; 2005.
19. Cohen J. Statistical power analysis for the behavioral sciences. New Jersey: Lawrence Erlbaum; 1988.
20. Diário da República. Aviso n.o 1347/2017. [consultado 2018 jun 04]. Disponível em: http://www.acss.min-saude.pt//wp-content/uploads/2017/02/Aviso_1347_2017.pdf.
21. Durning SJ, Hemmer P, Pangaro LN. The structure of program evaluation: an approach for evaluating a course, clerkship, or components of a residency or fellowship training program. Teach Learn Med. 2007;19:308–18.
22. Musick D. A conceptual model for program evaluation in graduate medical education. Acad Med. 2006;81:1051–6.
23. Weggemans MM, van Dijk B, van Dooijeweert B, Veenendaal AG, Ten Cate O. The postgraduate medical education pathway: an international comparison. GMS J Med Educ. 2017;34:Doc63.
24. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37:830–7.
25. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten CP, Newble DI, Dolmans DH, Mann KV,

Rothman A, et al, editors. International handbook of research in medical education. Dordrecht: Springer Netherlands; 2002. p. 647–72.
26. Harden RM. Revisiting 'assessment of clinical competence using an objective structured clinical examination (OSCE).' Med Educ. 2016;50:376–9.
27. Pusic MV, Boutis K, Hatala R, Cook DA. Learning curves in health professions education. Acad Med. 2015;90:1034-42.
28. Taveira-Gomes I, Mota-Cardoso R, Figueiredo-Braga M. Communication skills in medical students – an exploratory study before and after clerkships. Porto Biomedical Journal. 2016;1:173-80.
29. Stempień KL. Predictive validity of the examination for the Membership of the Royal Colleges of Physicians of the United Kingdom. London: University College London; 2014.
30. Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 2007;71:540–4.
31. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50:36-60.