Education for Primary Care

ISSN: 1473-9879 (Print) 1475-990X (Online) Journal homepage: http://www.tandfonline.com/loi/tepc20

Has there been a change in the knowledge of GP registrars between 2011 and 2016 as measured by performance on common items in the Applied Knowledge Test?

Catherine A. Neden, Claire Parkin, Carol Blow & Aloysius Niroshan Siriwardena

To cite this article: Catherine A. Neden, Claire Parkin, Carol Blow & Aloysius Niroshan Siriwardena (2018): Has there been a change in the knowledge of GP registrars between 2011 and 2016 as measured by performance on common items in the Applied Knowledge Test?, Education for Primary Care, DOI: 10.1080/14739879.2018.1467737

To link to this article: https://doi.org/10.1080/14739879.2018.1467737

Published online: 08 May 2018.

EDUCATION FOR PRIMARY CARE, 2018
https://doi.org/10.1080/14739879.2018.1467737

Taylor & Francis
Taylor & Francis Group

Has there been a change in the knowledge of GP registrars between 2011 and 2016 as measured by performance on common items in the Applied Knowledge Test?

Catherine A. Neden[a], Claire Parkin[b], Carol Blow[c] and Aloysius Niroshan Siriwardena[d]

[a]East Cliff Practice, Ramsgate, UK; [b]Centre for Professional Practice (Room M3-22), University of Kent, Kent, UK; [c]Stonehaven Medical Practice, Stonehaven, UK; [d]Primary and Prehospital Health Care, Community and Health Research Unit, Brayford Campus, University of Lincoln, Lincoln, UK

ABSTRACT

The aim of this study was to assess whether the absolute standard of candidates sitting the MRCGP Applied Knowledge Test (AKT) between 2011 and 2016 had changed. It is a descriptive study comparing the performance on marker questions of a reference group of UK graduates taking the AKT for the first time between 2011 and 2016. Using aggregated examination data, the performance of individual 'marker' questions was compared using Pearson's chi-squared tests and trend-line analysis. Binary logistic regression was used to analyse changes in performance over the study period. Changes in performance of individual marker questions using Pearson's chi-squared test showed statistically significant differences in 32 of the 49 questions included in the study. Trend line analysis showed a positive trend in 29 questions and a negative trend in the remaining 23. The magnitude of change was small. Logistic regression did not demonstrate any evidence for a change in the performance of the question set over the study period. However, candidates were more likely to get items on administration wrong compared with clinical medicine or research. There was no evidence of a change in performance of the question set as a whole.

ARTICLE HISTORY
Received 22 January 2018
Revised 12 April 2018
Accepted 18 April 2018

KEYWORDS
Medical education; assessment; Applied Knowledge Test; multiple choice test; standard

Introduction

Within general practice training, there is controversy about the significant difference in pass rates between cohorts of candidates [1]. This has raised questions about training and assessment processes, as well as the standards, as assessed by the performance in selection tests, of candidates commencing training [2,3].

Various methods are used to assess competence in postgraduate medical assessments according to the model described by Miller [4]. The multiple choice format is widely used in medical examinations as it is deemed to be valid, reliable and efficient [5]. In the context of general practice, the Applied Knowledge Test (AKT) is the mandatory high stakes computer based knowledge test component of the Membership of the Royal College of General Practitioners (MRCGP) examination [6]. Institutions awarding qualifications such as the RCGP must be confident in the validity of their assessment processes [7].

One function of the MRCGP is to set standards of practice and to provide assurance that doctors have the knowledge required to work as independent general practitioners in the UK. For the AKT, the standard is set using Angoff's method [8]. This method sets a criterion-based standard which is maintained by a process of linear equating until the next Angoff meeting (usually held triennially), according to methods described by Bandaranayake [9].

Previous work on the American Board of Internal Medicine examination between 1983 and 1988 compared performance on pairs of common items [10]. The authors noted a cumulative decline in the performance of US graduates from US medical schools over this time and an improvement in performance of non-US citizens from foreign medical schools - although this group was smaller and heterogeneous. In the UK, considering the Part 1 (the multiple-choice component) of the Membership of the Royal College of Physicians (MRCP), McManus found a decline in performance on marker items in the test between 1985 and 2002. A separate study looking specifically at the papers from 1996 and 2001 showed a significant reduction in performance [11].

CONTACT Catherine A. Neden neden@binternet.com

Selection into General Practice specialty training uses a competency based approach and selection scores correlate with end point examination scores [12–14]. In considering AKT scores, factors such as changes in selection or the inherent popularity of a specialty, may be confounding factors when considering changes in scores. Differential attainment in medical examinations is a widespread but poorly understood phenomenon [15–17]. Differential performance in the Clinical Skills Assessment (CSA) of the MRCGP was the subject of a judicial review in 2014. This differential was most marked between UK medical graduates and international medical graduates [1].

Although a decline in performance of candidates has been demonstrated in the American Board of Internal Medicine and the UK MRCP examinations, this has not previously been considered in the context of UK General Practice and the MRCGP examination. The aim of this study was to assess whether the absolute standard of candidates sitting the MRCGP Applied Knowledge Test (AKT) has changed over time.

Methods

Design

This was a descriptive longitudinal study comparing performance on marker items of a reference group of UK graduates taking the AKT for the first time at one of the 14 sittings of the examination between October 2011 and January 2016 (labelled as AKT13–AKT26). The AKT is a 200-item machine-marked test. Routinely collected data about candidates includes the stage of training and place of primary medical qualification as well as gender, age and number of previous attempts. The items are selected from a question bank and sample across the GP curriculum. Each test included several marker items which have been used on two or more occasions, with the question and answer unaltered, and with a point biserial on psychometric analysis was >0.2.

The chosen reference group of UK graduates taking the test for the first time was the largest cohort being relatively homogeneous in terms of previous undergraduate training. This approach was consistent with that taken by other researchers in the field [3]. International medical graduates were excluded as they represented a relatively heterogeneous group in terms of training background and performance. Candidates resitting the test were excluded as their performance differs for various reasons [18–20].

Of the marker questions used since AKT 13, those that had been used on four or five occasions were included. Those used three times or less were excluded as they were less likely to give reliable data given the inherent variability of the data.

Data collection

Aggregate performance data for the reference group of candidates for each of the marker items were extracted from the RCGP examination database by the psychometric adviser supporting the AKT. This was entered into IBM SPSS statistics for analysis. An independent researcher checked a sample of the data entered for accuracy.

Data analysis

The data for each question were in a binary format (correct or incorrect) for each time of administration. Each question was classified according to the curriculum area being tested. Given that the data were categorical, Pearson's chi-squared test was used to look for statistical differences between times of administration [21]. Data for each question were viewed separately to obtain a trend-line analysis using Microsoft Excel v14. Binary logistic regression, using IBM SPSS statistics, of the whole data-set was used to test performance of questions overall, and the three categories of question (administration, research or clinical medicine) over time [22].

Results and analysis

Details of the total candidature for each sitting of the examination are included in Table 1. The study group varied between 559 candidates in AKT14 and 959 in AKT 24. The AKT question bank contains over 3000 questions, 222 of which have been used as marker questions. Over the study period, four of these were used on five occasions and 45 on four occasions. Data from these 49 questions were used in this study. Of these, 10 tested administration areas of the GP curriculum, six tested evidence based medicine and the remainder clinical medicine.

Analysis of the performance of individual marker questions

The value of Pearson's chi-squared was calculated for each of the questions. This compared the frequency of a correct response for the item for each of the sittings of the AKT in which it had been used. The chi-squared value showed statistically significant differences between the different sittings of the AKT for 32 of the 49 questions studied. There was no obvious difference in question category between the questions that demonstrated a difference from those that did not (Table 2). However, the Pearson's chi-squared test considers differences between groups but does not test for a chronological trend. It does not take account the order of administration of questions at different AKT sittings either.

Table 1 Details of the candidature for AKT13- AKT26 (October 2011–January 2016).

AKT sitting	Date of exam	Total number of candidates	First time takers	UK Graduate first time takers	First time takers in ST2 year of training
13	Oct-11	1514	1115	798	22.40%
14	Jan-12	1204	847	559	57.10%
15	May-12	1472	1146	919	70.40%
16	Oct-12	1681	1199	861	28.40%
17	Jan-13	1248	881	683	60.50%
18	May-13	1670	1255	956	68.70%
19	Oct-13	1472	1000	822	31.90%
20	Jan-14	1284	1004	819	71%
21	Apr-14	1430	1089	869	70.80%
22	Oct-14	1264	913	756	39%
23	Jan-15	1196	956	792	69.80%
24	Apr-15	1487	1150	959	71.10%
25	Oct-15	1332	968	816	38.70%
26	Jan-16	1086	848	707	70.80%

Table 2 Significant differences (p <0.05) in Pearson's chi-squared according to question category.

Question category	Question sub-category	No. questions in the study	No. with significant difference
Clinical medicine	Disease	3	3
	Symptoms	11	4
	Investigation	6	5
	Management	13	10
Administration		10	8
Research		6	5

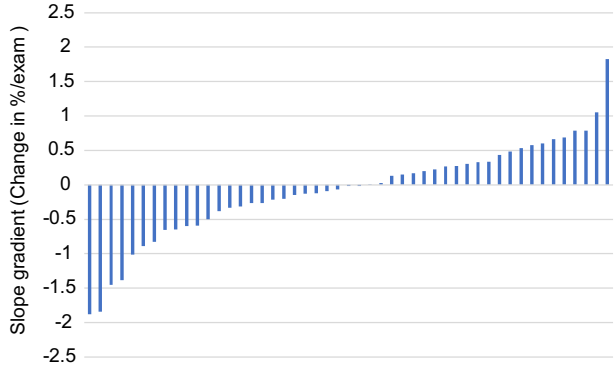

Figure 1. Trendline gradients for individual questions.

Table 3 Results of logistic regression (testing effect of AKT diet) reporting odds ratio.

	β	SE β	OR	95% CI	
				Lower	Upper
AKT diet (time)	0.10	0.001	1.010	1.007	1.013
Constant	−1.387	0.028	0.250		
Classification accuracy	76.7%				
Nagelkerke R^2	0.000				
Hosmer & Lemeshow Test	p<0.05				
-2Log Likelihood	174832.462				

The data for each question were examined using Microsoft Excel® v14.7.1, adding a trend line to obtain a graphical representation of the data trends for each question. To estimate the relationship between the correct response rate and time, the equation of this trend line was calculated using Microsoft® Excel® v14.7.1. This is determined using the least squares method. R^2 was calculated as a measure of the 'fit' of this line to the data. The value of R^2 lies between 0 and 1, and is the proportion of the variation

Table 4. Results of logistic regression (testing effect of question category) reporting odds ratio.

	β	SE β	OR	95% CI	
				Lower	Upper
Clinical	Reference				
Admin	0.122	0.014	1.130	1.098	1.162
Research	−0.155	0.019	0.857	0.825	0.889
Constant	−1.201	0.007	0.301		
Classification accuracy	76.7%				
Nagelkerke R^2	0.002				
Hosmer & Lemeshow Test	*p* 1.000				
-2Log Likelihood	174710.092				

Table 5. Results of logistic regression (testing effect of AKT diet, question category and interaction) reporting odds ratio.

	β	SE β	OR	95% CI	
				Lower	Upper
Clinical	Reference				
Admin	0.530	0.068	1.698	1.485	1.942
Research	-0.316	0.090	0.729	0.612	0.869
AKT diet (time)	0.13	0.002	1.013	1.010	1.017
AKT diet and Clinical	Reference				
AKT diet and Admin	−0.021	0.003	0.979	0.973	0.986
AKT diet and Research	0.009	0.005	1.009	1.000	1.018
Constant	−1.455	0.034	0.233		
Classification accuracy	76.7%				
Nagelkerke R^2	0.002				
Hosmer & Lemeshow Test	*p* <0.05				
-2Log Likelihood	174617.675				

that can be explained by the linear relationship [23]. The trend line had a positive value for 23 of the marker questions and a negative value for 26. This is shown in Figure 1. However, in the majority of cases, actual values were small and the R^2 value was greater than 0.8 in only seven cases. This suggests that the accuracy of the line of best fit was limited for the majority of questions.

Analysis of the combined question data-set

Each AKT paper includes a defined selection of items from the question bank and decisions were based upon performance on these items as a whole. Thus, it is considered appropriate to consider performance upon a group of items in addressing the research question. Given that the data have a binary outcome variable (correct–incorrect), logistic regression (using SPSS) was used to test the following hypotheses:

Null hypothesisH_0 1 = There is no difference in the standard of candidates for MRCGP as assessed by scores on AKT marker items between October 2011 and April 2016.

Null hypothesis

H_0 2 = There is no difference between different major question categories (administration, evidence-based medicine and clinical medicine) in the standard of candidates for MRCGP, as assessed by scores on AKT marker items, between October 2011 and April 2016.

For this analysis, the questions were grouped according to the categories used for setting the test and reporting the results (clinical medicine, administration and evidence-based medicine). Clinical medicine was used as the baseline category for the analysis as described by Kirkwood and Sterne [24]. As the questions were coded in SPSS with correct = 0 and incorrect = 1, the odds ratios were expressed in terms of comparison with the baseline group of candidates answering the item correctly. Given that the AKT diet term is an ordered, categorical variable, logistic regression was used to estimate the most likely value in the increase in the log odds for each sitting of the test. This tested for a linear association, with a constant increase in the log odds per unit increase in the exposure variable, between sittings of the AKT. The results of the logistic regression are summarised in Tables 3–5. The logistic regression output was evaluated according to the methods described by Peng, Lee and Ingersoll [25]. A logistic regression model is said to provide a better fit to the data if it demonstrates an improvement over the intercept (constant) only model.

Primary outcome

Logistic regression found no difference in terms of candidate performance with time between the different sittings of the AKT (OR = 1.010 95% CI 1.007–1.013). Thus, the null hypothesis (H_0 1*)* is accepted. The significant value of Wald statistic suggested that both the constant and the effect of time should be included in the model. However, the inferential goodness of fit statistic (Hosmer–Lemeshow) was significant ($p < 0.05$) suggesting that there was not a good model fit of data. The low value of Nagelkerke's R^2 (a descriptive measure of goodness of fit) suggested that the variation was not explained by the predictors used in the model. This was supported by the high value of the -2 Log Likelihood (-2LL) ratio, suggesting that the accuracy of the model was limited. Thus, the results must be treated with caution.

Secondary outcome

Logistic regression found strong evidence that the odds of failing an administration category marker question was 1.13 times that of failing a clinical medicine question (OR = 1.13 95% CI 1.098- 1.162). However, the odds of failing an evidence-based medicine question were significantly lower (OR = 0.857 95% CE 0.825–0.889).

Evaluating this model according to the principles described above, the Wald statistic was significant for each value of β and the constant. The Hosmer–Lemeshow test was not significant, suggesting the model fitted the data. However, Nagelkerke's R^2 suggested that only 2% of the variation was accounted for by this model, and this was supported by the high value of the -2LL ratio. In summary, there is little variation over time and it is presumed that most of this relates to other factors which might include candidate ability and learning.

Discussion

Main findings

This study aimed to test whether there had been a significant change in the standard of knowledge as measured by the score on AKT marker items between October 2011 and January 2016. Whilst statistically significant differences in the correct response rate (facility) were noted for 32 of the 49 questions, in the majority of cases, whilst there was test-to-test variation there was no consistent pattern. Significant and non-significant differences were seen for each of the categories of question used in the test.

This was in accord with the trend line analysis of the marker questions. There was no consistent trend in the performance of candidates on individual marker questions over the study period. The largest increase (steepest trend line) was noted for an administration question relating to prescribing regulations, whilst the largest falls were in two clinical medicine questions relating to the

management of long-term conditions and another relating to research terminology.

Logistic regression found no significant difference in candidate performance on the marker items considered as a whole over the duration of the study period. There were significant differences between the question categories with significantly higher odds of failing an administration question and lower odds of failing an evidence-based medicine question (compared to clinical medicine questions). This was in accord with Esmail and Roberts' [2] findings in the independent review of candidate performance in the MRCGP between 2010 and 2012. The authors reported that the mean score of UK graduates at their first sitting of the AKT was lower for administration questions compared to those of clinical medicine.

These differences reduced when the effect of time was introduced to the logistic regression equation. It should be noted that in all of these cases, the classification accuracy of the model used remained at 76.7% and this, taken alongside the statistical evaluations, suggests that these models only explain a small proportion of the variation seen. Any conclusions drawn regarding these results should be treated with caution. The use of logistic regression for an ordered variable (such as AKT diet) assumes a linear relationship and that the change in log odds between each interval is similar. It does not test for a non-linear association.

Strengths and limitations

The study explored an area not previously investigated. The data-set included information from a large number of candidates in 14 consecutive sittings of the test over 4.5 years. It included 161,129 responses from candidates to individual questions. Each question had been used on four or five occasions during the study period, giving a number of data points for comparison. There was no missing data to account for. The current process for setting the pass mark for the AKT uses Angoff's methodology. The constancy of scoring in a group of marker questions over time affirms the continued use of linear equating.

There were several limitations to the study. The study used questions that had been administered at least four times. Although statistically this improved the accuracy of each question used, a consequence was that a relatively small number of questions were included in the analysis. There was heterogeneity in the nature of knowledge being tested by these questions, which limited the ability to consider sub-categories of question in the analysis, particularly in the clinical medicine domain. A further study considering a larger number of questions administered over a longer period would be valuable since the period adopted in this study may be too short for trends to emerge.

The data-set was in aggregated format. It was limited to a group of UK graduates taking the test for the first time, as it was considered that this group would be similar in terms of educational background and experience. This study was not able to take account of the known differences between ethnic groups, in terms of academic performance, in the data analysis. These differences have been described in the context of undergraduate and postgraduate examinations [1,15]. Differences in other demographic features such as age and gender, which are known to affect performance in the AKT, could not be included [27,28]. It was not possible to explore any linkage with any other data such as that from selection into GP training. Future work is needed to address and explore these differences further. Assessing whether there is differential attainment between any of these groups is important in demonstrating fairness in a licensing examination. Restricting the study group to those taking the test for the first time means that it is not possible to add to the debate upon the reuse of items in an MCQ test.

Comparison with existing literature

Publications from other medical colleges have demonstrated a decline in the standard of candidates [11,26]. This had not previously been considered in the context of the MRCGP and this study does not demonstrate any change over a period of 4.5 years. The study by Norcini et al. (1991) [26] tested for differences in the mean scores of items that were common to two pairs of specific test papers, and it may be that this design allowed for more precise analysis. However, these sets of questions were derived from an examination lasting two days and it would not have been possible or appropriate to adopt this method for items drawn from a 200-item, three-hour test.

Implications for policy, practice and research

To maintain the trust of the wider community, the medical profession needs to be open to scrutiny about self-regulation. Part of this is a requirement for transparency about how standards for licensing are set and implemented. Quality assurance of all elements of the examination is essential. This study appears to confirm the stability of performance on marker items and affirms their use in standard setting processes. It contributes to this by validating the use of linear equating in setting the pass mark for an individual AKT paper.

Conclusion

When differences between different times of administration of individual questions are considered, there are statistically significant differences in the rate these were correctly answered. These do not appear to follow a consistent trend since the raw data shows variation around the mean. When considered in the context of an examination, this is unlikely to be significant from the candidate's perspective. This study has not demonstrated any change in the knowledge of candidates for the AKT as measured by performance on a subset of marker questions between October 2011 and January 2016, using logistic regression.

Ethical approval

Ethical approval for the study was granted by the University of Kent (Centre for Professional Practice) Ethics committee. The RCGP examination department gave permission for the use of anonymised data.

Disclosure statement

The authors, aside from Claire Parkin, are members of the panel of examiners of the Royal College of General Practitioners and members or recent members of the Applied Knowledge Test (AKT) Development Group.

Declaration of interests

The authors state there are no other competing interests.

ORCID

Catherine A. Neden http://orcid.org/0000-0001-6293-2960

References

[1] Rendel S, Foreman P, Freeman A. Licensing exams and judicial review: the closing of one door and opening of others? Br J Gen Pract. 2015 Jan;65(630):8–9. DOI:10.3399/bjgp15X683029. PubMed PMID: 25548294; PubMed Central PMCID: PMCPMC4276001. eng.

[2] Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. BMJ. 2013;347. DOI:10.1136/bmj.f5662

[3] Patterson F, Kerrin M, Baron H, et al. Exploring the relationship between general practice selection scores and MRCGP examination performance. Final Report to General Medical Council; 2015.

[4] Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-7. PubMed PMID: 2400509; eng.

[5] Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine?. BMC Med Edu. 2014;14(1):3. DOI:10.1186/s12909-014-0249-2.

[6] RCGP. MRCGP exam overview 2017 [cited 2017 Mar 5]. Available from: http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview.aspx

[7] Boursicot K, Fuller R, Kemp S, et al. Final report for the provision of identifying key principles for consistency and reliability in curricula and assessment frameworks. Singapore: HPAC; 2016.

[8] RCGP. AKT standard setting; 2015 [cited 2017 April 30; 2017 May 5]. Available from: http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview/mrcgp-applied-knowledge-test-akt.aspx

[9] Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach. 2008;30(9–10):836–845. DOI:10.1080/01421590802402247 . PubMed PMID: 41561997.

[10] Norcini JJ, Shea JA, Day SC, et al. Trends in the medical knowledge and clinical competence of graduates of internal medicine residency training programs. Acad Med. 1991;66(9):S58–60.

[11] McManus I, Mollon J, Duke O, et al. Changes in standard of candidates taking the MRCP(UK) part 1 examination, 1985 to 2002: analysis of marker questions. BMC Med. 2005;3(1):13. DOI:10.1186/1741-7015-3-13.

[12] Ahmed H, Rhydderch M, Matthews P. Do general practice selection scores predict success at MRCGP? An exploratory study. Educ Prim Care. 2012;23(2):95-100. PubMed PMID: 22449464; eng.

[13] Davison I, Burke S, Bedward J, et al. Do selection scores for general practice registrars correlate with end of training assessments? Edu Prim Care. 2006;17(5):473.

[14] Davison I, McManus IC, Taylor C. Evaluation of GP specialty training. Birmingham: University of Birmingham; 2016.

[15] Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ. 2011;342. DOI: 10.1136/bmj.d901.

[16] Regan de Bere S, Nunn S, Nasser M. Understanding differential attainment across medical training pathways: a rapid review of the literature; 2015. Available from: http://www.gmc-uk.org/GMC_Understanding_Differential_Attainment.pdf_63533431.pdf

[17] Woolf K, Rich A, Viney R, et al. Fair training pathways for all: understanding experiences of progression; 2016. Available from: http://www.gmc-uk.org/about/research/29485.asp

[18] Pell G, Boursicot K, Roberts T. The trouble with resits …. Ass Evalu Higher Edu. 2009;34(2):243–251.

[19] Ricketts C. A new look at resits: are they simply a second chance? Assess Eval Higher Educ. 2010;35(4):351–356.

[20] McManus IC, Ludka K. Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations [OriginalPaper]. BMC Med. 2012;10(1):1. Doi: 10.1186/1741-7015-10-60. En.

[21] Bryman A, Cramer D. Quantitative data analysis with SPSS: a guide for social scientists. Hove: Routledge; 2001.

[22] Field A. Discovering statistics using IBM SPSS statistics. 4th ed. London: Sage; 2014.

[23] Winston WL. Microsoft® Excel® 2010: data analysis and business modeling. 3rd ed. Pearson Education MUA; 2011.

[24] Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Oxford: Blackwell; 2003.
[25] Peng C-YJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Edu Res. 2002;96(1):3-14.
[26] Norcini JJ, Shea JA, Benson JA Jr. Changes in the medical knowledge of candidates for certification. Ann Int Med. 1991;114(1):33–35.
[27] Siriwardena AN, Irish B, Asghar ZB, et al. Comparing performance among male and female candidates in sex-specific clinical knowledge in the MRCGP. Br J Gen Pract. 2012;62(599):446–450.
[28] Wakeford R. International medical graduates' relative under-performance in the MRCGP AKT and CSA examinations. Educ Prim Care. 2012;23(3):148–152. PubMed PMID: 22762872; eng.