Journal of Educational Evaluation for Health Professions

J Educ Eval Health Prof 2015, 12: 11 • http://dx.doi.org/10.3352/jeehp.2015.12.11

Open Access

eISSN: 1975-5937

TECHNICAL REPORT

Best-fit model of exploratory and confirmatory factor analysis of the 2010 Medical Council of Canada Qualifying Examination Part I clinical decision-making cases

André F. De Champlain*

Research and Development, Medical Council of Canada, Ottawa, Ontario, Canada

Abstract

Purpose: This study aims to assess the fit of a number of exploratory and confirmatory factor analysis models to the 2010 Medical Council of Canada Qualifying Examination Part I (MCCQE1) clinical decision-making (CDM) cases. The outcomes of this study have important implications for a range of domains, including scoring and test development. **Methods:** The examinees included all first-time Canadian medical graduates and international medical graduates who took the MCCQE1 in spring or fall 2010. The fit of one- to five-factor exploratory models was assessed for the item response matrix of the 2010 CDM cases. Five confirmatory factor analytic models were also examined with the same CDM response matrix. The structural equation modeling software program Mplus was used for all analyses. **Results:** Out of the five exploratory factor analytic models that were evaluated, a three-factor model provided the best fit. Factor 1 loaded on three medicine cases, two obstetrics and gynecology cases, and two orthopedic surgery cases. Factor 2 corresponded to pediatrics, and the third factor loaded on psychiatry cases. Among the five confirmatory factor analysis models examined in this study, three- and four-factor lifespan period models and the five-factor discipline models provided the best fit. **Conclusion:** The results suggest that knowledge of broad disciplinary domains best account for performance on CDM cases. In test development, particular effort should be placed on developing CDM cases according to broad discipline and patient age domains; CDM testlets should be assembled largely using the criteria of discipline and age.

Key Words: *Educational measurement; Medical licensure; Statistical factor analysis; Canada*

INTRODUCTION

The Medical Council of Canada Qualifying Examination Part I (MCCQE1) is a two-part computer-based examination that assesses the knowledge, skills, and attitudes judged essential for entry into supervised post-graduate medical training according to the specific statement of objectives of the Medical Council of Canada [1]. The first part of the examination includes 196 five-option, single-best-answer (A-type) multiple choice items. These 196 multiple-choice questions are distributed into seven sections that contain 28 items apiece. The second part of the MCCQE1 is composed of approximately 60 clinical decision-making (CDM) cases. Each CDM case includes one to five questions, for a total of approximately 80 questions. CDM cases included in the MCCQE1 provide a measure of the problem-solving and decision-making skills of candidates as they pertain to specific clinical scenarios. The MCCQE1 is administered in two multi-week windows at over a dozen test sites located throughout Canada. The examination is internet-delivered at dedicated secure sites located largely in Canadian medical schools. Candidates have up to 3.5 hours to complete the multiple-choice question portion of the MCCQE1, whereas up to four hours are allocated for completing the CDM cases. This study aims to compare the fit of a

*Corresponding email: adechamplain@mcc.ca

Received: April 12, 2015; Accepted: April 15, 2015; Published: April 15, 2015

This article is available from: http://jeehp.org/

number of exploratory and confirmatory factor analysis models to the 2010 combined spring and fall MCCQE1 CDM item response matrix using the Mplus software package (Muthen & Muthen, Los Angeles, CA, USA) [2]. The results of this investigation will provide information relevant to a range of psychometric analyses related to CDM cases, including how to best estimate scores and calibrate this component of the MCCQE1, and will also help develop the unidimensionality test for estimating the item parameters for CDM cases.

METHODS

MCCQE1 cohort

The present investigation focused on the combined spring and fall 2010 MCCQE1 examinee cohorts. The spring administration population is composed primarily of first-time Canadian medical graduates (CMGs), whereas international medical graduates (IMGs) comprise the bulk of the fall testing cohort. Analyses were centered on all first-time test takers for both the spring and fall 2010 MCCQE1 administrations. A breakdown of the cohort by training (i.e., CMG vs. IMG) and test administration is provided in Table 1. The majority of the 2010 combined cohort was composed of CMGs (2,429, 60.2%) and does conform to expected cyclical patterns; that is, CMGs made up the majority of the spring 2010 MCCQE1 administration, whereas IMGs largely took the test in the fall administration window.

MCCQE1 bank

The bank of multiple-choice items available for the combined 2010 MCCQE1 administrations included several thousand items. Over 100 CDM cases were also available for use in the 2010 bank. CDM cases are developed to target problem-solving and clinical decision-making skills. Examinees were presented with case descriptions, followed by one or more test questions that assessed key clinical issues in the resolution of the case. Questions could relate to eliciting clinical information, ordering diagnostic procedures, making diagnoses, or prescribing therapy.

Table 1. Breakdown of the 2010 Medical Council of Canada Qualifying Examination part I first-time examinee population by training type and test administration timing

Training	Test administration		
	Spring	Fall	Total
Canadian medical graduates	2,407 (59.6)	22 (0.6)	2,429 (60.2)
International medical graduates	790 (19.6)	817 (20.2)	1,607 (39.8)
Total	3,197 (79.2)	839 (20.8)	4,036 (100.0)

Values are presented as number (%).

Analyzed cases

Examinee responses reflect decisions made in the management of actual patients. CDM cases include both short-menu and write-in item formats, and they are polytomously scored on a proportion-correct scale. For the purposes of this study, these proportion-correct case scores were integerized (i.e., transformed to whole numbers) to enable analyses using Mplus. The majority of CDM cases had either two or three response categories (63% of the bank). Given the very sparse nature of the CDM case matrix and the challenges that this poses from a covariance coverage perspective in Mplus, the final analyses were conducted on a set of 17 CDM cases that were culled from the original set of cases. The cases were representative of the bank with respect to a number of classification variables.

Analyses

All analyses were carried out using the structural equation modeling software program Mplus [2]. Initially, the fit of one- to five-factor exploratory models (exploratory factor analytic models [EFAs]) was assessed for the combined 2010 CDM item response matrix. Given the non-normal nature of CDM case score distributions, weighted least-squares parameter estimation, using a diagonal weight matrix with standard errors and mean- and variance- adjusted chi-square tests (using a full weight matrix), was implemented [3]. The latter estimation method is appropriate for data that violate the assumptions of more common methods (such as the normality assumption underlying the maximum likelihood and the generalized least-squares estimations).

The second set of analyses focused on fitting a number of confirmatory factor analytic models (CFA) to the same 2010 item response matrix based on substantive considerations identified through a review of the current CDM blueprint. Specifically, the following five CFA models were examined: first, a three-factor 'location/setting' model; second, a three-factor 'lifespan period' model; third, a four-factor 'lifespan period' model; fourth, a four-factor 'clinical situation' model; and fifth, a five-factor 'discipline' model. Table 2 provides a breakdown of the 17 CDM cases as a function of these classifying variables. The three-factor 'location/setting' model posited the following factor structure: factor 1 (family physician office) loaded on CDM cases 1, 2, 3, 6, 7, 8, 9, 10, 12, and 13; factor 2 (general hospital) loaded on CDM cases 4 and 11; and factor 3 (emergency department) loaded on CDM cases 5, 14, 15, 16, and 17. The four-factor 'lifespan period' model posited the following factor structure: factor 1 (adult) loaded on CDM cases 1, 2, 3, 7, 15, 16, and 17; factor 2 (pediatrics) loaded on CDM cases 9, 10, and 13; factor 3 (adolescent) loaded on CDM cases 6, 8, 12, and 14; and factor 4 (pregnancy/neonatal/infant) loaded on CDM cases 4, 5, and 11. A three-factor modified version of

Table 2. Seventeen clinical decision-making cases by location/setting, lifespan period, clinical situation, and discipline on the 2010 Medical Council of Canada Qualifying Examination part I

Case	Location/Setting	Lifespan period	Clinical situation	Discipline
1	Family physician office	Adult	Undifferentiated complication	Medicine
2	Family physician office	Adult	Single typical problem	Medicine
3	Family physician office	Adult	Preventive care/health problem	Medicine/PHELO
4	General hospital	Pregnancy/neonatal	Single typical problem	Obstetrics/gynecology
5	Emergency department	Pregnancy/neonatal	Multiple-problem event	Obstetrics/gynecology
6	Family physician office	Adolescence	Preventive care/health problem	Obstetrics/gynecology
7	Family physician office	Adult	Undifferentiated complication	Obstetrics/gynecology/PHELO
8	Family physician office	Adolescence	Single typical problem	Pediatrics
9	Family physician office	Pediatric	Preventive care/health problem	Pediatrics
10	Family physician office	Pediatric	Undifferentiated complication	Pediatrics
11	General hospital	Pregnancy/neonatal	Multiple-problem event	Pediatrics
12	Family physician office	Adolescence	Multiple-problem event	Pediatrics
13	Family physician office	Pediatric	Single typical problem	Psychiatry
14	Emergency department	Adolescence	Multiple-problem event	Psychiatry
15	Emergency department	Adult	Single typical problem	Surgery
16	Emergency department	Adult	Single typical problem	Surgery
17	Emergency department	Adult	Multiple-problem event	Surgery

PHELO, population health, ethical, legal, and organizational aspects of medicine.

the latter CFA model was also examined, in which factor 2 (pediatrics/pregnancy/neonatal/infant) loaded on CDM cases 4, 5, 9, 10, 11, and 13, based on exploratory correlational analyses. The remaining factor structure was identical to the four-factor 'lifespan period' model. The four-factor 'clinical situation' model posited the following factor structure: factor 1 (undifferentiated complaint) loaded on CDM cases 1, 7, and 10; factor 2 (single typical problem) loaded on CDM cases 2, 4, 8, 13, 15, and 16; factor 3 (preventive care and health promotion) loaded on CDM cases 6, 8, 12, and 14; and factor 4 (multiple problem or multi-system life-threatening event) loaded on CDM cases 3, 6, and 9. Finally, the five-factor 'discipline' model posited the following factor structure: factor 1 (medicine) loaded on CDM cases 1, 2, and 3; factor 2 (obstetrics/gynecology) loaded on CDM cases 4, 5, 6, and 7; factor 3 (pediatrics) loaded on CDM cases 8, 9, 10, 11, and 12; factor 4 (psychiatry) loaded on CDM cases 13 and 14; and factor 5 (surgery) loaded on CDM cases 15, 16, and 17. As was the case with the EFAs, a diagonal weight matrix-based estimation procedure was used in all CFAs.

The fit of all models was assessed via the following statistics and indices: the chi-square test of model fit, the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation. Both the CFI and TLI evaluate the fit of a user-specified solution in relation to a more restricted nested baseline model, in which the covariance among all input indicators is fixed to zero or no relationship among the variables that are posited; in other words, the number of dependent variables is equal to the number of factors. The TLI additionally imposes a correction for over-parameterization. CFI and TLI values range from 0 to 1, though the TLI can exceed 1 with severe over-fitting, with values of 0.90 or above indicating acceptable fit [4]. It is important, however, to underscore that the relative fit of the five-factor models will be compared as opposed to the absolute fit of any given solution. Practically speaking, it is of greater interest to compare the relative fit of the five alternative models previously outlined rather than attempting to identify an 'optimal' configuration from a statistical point of view. Adopting this relative approach is also congruent with views espoused by several factor analysts who maintain that no restrictive model fits the population and that all restrictive models are merely approximations [5]. Consequently, our analyses were aimed at identifying the best-fitting model among those under study, all of which were posited based on substantive considerations, rather than attempting to accept or reject an *a priori* false hypothesis.

RESULTS

Exploratory factor analyses

Table 3 provides fit values for the five EFAs that were examined in this study. Based on these results, it appears that a three-factor EFA solution provided the best fit to the item response matrix, without over-fitting, which was clearly observed in the four- and five-factor models based on CFI and TLI values. The three-factor obliquely rotated factor loadings are provided in Table 4. Using a rough cutoff of 0.25 in order to better define the nature of the factor structure, it appears as though factor 1

could generally be described as reflecting 'biomedical/medicine' CDM cases. Factor 1 loads on three 'medicine' cases, two biomedically oriented 'obstetrics and gynecology' cases, and two 'orthopedic surgery' cases. Factor 2, which loads more heavily on CDM cases 8, 9, 10, and 14, appears to reflect a 'pediatrics' factor. Finally, factor 3 could be labeled as a 'psychiatry' factor, with heavier loadings on CDM cases 13 and 14. Finally, the correlations between the three factors were quite low, ranging from -0.04 (F1-F3) to 0.09 (F2-F3), suggesting that distinct competencies are required to perform well on each type of CDM case.

Confirmatory factor analyses

Table 5 provides fit statistic values for the five CFA models that were examined in this study. Based on these results, it appears that both the 'lifespan period' and 'discipline' models provided the best fit amongst the five CFA models examined in this study. Factor loadings and inter-factor correlations for the four-factor 'lifespan period' CFA model are provided in Tables 6 and 7. Most prescribed loadings were statistically significant. However, some of the factors did not load on their assigned CDM cases. With regard to factor 1 (adult), CDM case 7 (infertility) was poorly associated with the domain. Similarly, factor 3 (adolescent) poorly loaded on CDM case 12 (life-threatening asthma). Finally, factor 4 (pregnancy/neonatal/infant) poorly loaded on CDM case 5 (diabetic pregnancy). In regard to factor correlations, values ranged from -0.04 (between 'pediatrics' and 'pregnancy/neonatal/infant') to 0.92 (between 'adult' and 'pregnancy/neonatal/infant'). Factor loadings as well as inter-factor correlations for the five-factor 'discipline' CFA model are provided in Tables 8 and 9. Again, the vast majority of pre-specified loadings were statistically significant. However, as was the case with the previous model, some of the factors did not load on their prescribed CDM cases. With regard to factor 2 (obstetrics/gynecology), CDM cases 5 (diabetic pregnancy) and 7 (infertility) were poorly associated with the domain. Similarly, factor 3 (pediatrics) poorly loaded on CDM case 12 (life-threatening asthma). Finally, factor 4 (psychiatry) was heavily defined by CDM case 14 (threatened suicide). In regard to factor correlations, values ranged from 0.07 (between 'medicine' and 'psychiatry') to 0.91 (between 'medicine' and 'obstetrics/gynecology').

Table 3. Goodness-of-fit statistics for five exploratory clinical decision-making factor analysis models of the 2010 Medical Council of Canada Qualifying Examination part I

Model	Chi-square	Goodness of fit	Tucker-Lewis index	Root mean square error of approximation
One-factor	260.23, P < 0.001	0.65	0.60	0.07
Two-factor	138.56, P = 0.01	0.91	0.88	0.05
Three-factor	88.83, P = 0.46	1.00	1.00	0.05
Four-factor	64.08, P = 0.79	1.00	1.05	0.04
Five-factor	44.03, P = 0.95	1.00	1.10	0.04

Table 4. Obliquely rotated factor loadings for the three-factor clinical decision-making exploratory solution of the 2010 Medical Council of Canada Qualifying Examination part I

Clinical decision-making case	Factor 1	Factor 2	Factor 3
1	0.43	0.18	-0.01
2	0.39	0.09	-0.08
3	0.25	0.00	-0.19
4	0.34	0.00	-0.07
5	0.28	-0.07	-0.07
6	0.16	0.11	0.19
7	0.08	0.08	0.03
8	0.00	0.48	0.02
9	0.00	0.39	-0.39
10	-0.07	0.51	0.00
11	0.27	0.03	-0.01
12	0.29	-0.06	0.13
13	-0.02	0.08	0.36
14	0.00	0.43	0.36
15	0.20	0.13	0.05
16	0.31	0.04	0.09
17	0.40	-0.05	0.13

Table 5. Goodness-of-fit statistics for five confirmatory clinical decision-making factor analytic models of the 2010 Medical Council of Canada Qualifying Examination part I

Model	Chi-square	Goodness of fit	Tucker-Lewis index	Root mean square error of approximation
Three-factor 'location/setting'	259.53, P < 0.001	0.64	0.58	0.02
Four-factor 'lifespan'	185.86, P < 0.001	0.82	0.79	0.01
Three-factor 'modified lifespan'	221.07, P < 0.001	0.74	0.69	0.02
Four-factor 'clinical situation'	302.21, P < 0.001	0.53	0.44	0.02
Five-factor 'discipline'	180.44, P < 0.001	0.83	0.78	0.02

Table 6. Factor loadings for the four-factor clinical decision-making lifespan confirmatory factor analytic model of the 2010 Medical Council of Canada Qualifying Examination part I

Clinical decision-making case	Factor 1 (adult), factor 2 (pediatrics), factor 3 (adolescent)	Factor 2 (pediatrics)	Factor 3 (adolescent)	Factor 4 (pregnancy/ neonatal)
1	0.47[a]			
2	0.36[a]			
3	0.20[a]			
4				0.35[a]
5				0.21
6			0.27[a]	
7	0.09			
8			0.44[a]	
9		0.14[a]		
10		0.49[a]		
11				0.30[a]
12			0.10	
13		0.23[a]		
14			0.52[a]	
15	0.27[a]			
16	0.33[a]			
17	0.38[a]			

[a]P < 0.02.

Table 7. Inter-factor correlation matrix for the four-factor clinical decision-making lifespan confirmatory factor analytic model of the 2010 Medical Council of Canada Qualifying Examination part I

	Adult	Pediatrics	Adolescent	Pregnancy/ neonatal
Adult	1.00	0.16	0.32[a]	0.92[a]
Pediatrics		1.00	0.90[b]	-0.04
Adolescent			1.00	0.22
Pregnancy/neonatal				1.00

[a]P < 0.01. [b]Correlation was fixed at 0.90 due to estimation difficulties (multicolinearity).

Table 8. Factor loadings for the five-factor clinical decision-making discipline confirmatory factor analytic model of the 2010 Medical Council of Canada Qualifying Examination part I

Clinical decision-making case	Factor 1 (medicine)	Factor 2 (obstetrics/ gynecology)	Factor 3 (pediatrics)	Factor 4 (psychiatry)	Factor 5 (surgery)
1	0.57[a]				
2	0.41[a]				
3	0.20[a]				
4		0.34[a]			
5		0.11			
6		0.24[a]			
7		0.13			
8			0.54[a]		
9			0.22[a]		
10			0.54[a]		
11			0.20[a]		
12			0.05		
13				0.20[a]	
14				0.90[b]	
15					0.30[a]
16					0.42[a]
17					0.40[a]

[a]P < 0.02. [b]Loading was fixed at 0.90 due to estimation difficulty.

Table 9. Inter-factor correlation matrix for the five-factor clinical decision-making discipline confirmatory factor analytics model of the 2010 Medical Council of Canada Qualifying Examination part I

	Medicine	Obstetrics/ gynecology	Pediatrics	Psychiatry	Surgery
Medicine	1.00	0.91[a]	0.39[a]	0.07	0.61[a]
Obstetrics/ gynecology		1.00	0.19	0.35[a]	0.59[a]
Pediatrics			1.00	0.41[a]	0.19
Psychiatry				1.00	0.14
Surgery					1.00

[a]P < 0.02.

DISCUSSION

Assessing the underlying structure of any item response matrix is critical to both test development and psychometric efforts. From a test development standpoint, such analyses can provide substantiating evidence with respect to both blueprinting and test design activities. From a psychometric perspective, the use of advanced modeling techniques, such as item response theory, is predicated on a clear understanding of the data structure that is being analyzed. While common item response theory models assume unidimensionality of the underlying latent ability, research has shown that the underlying latent ability is robust against departures from this assumption, as long as the composite of proficiencies is comparable across test forms [6]. From a scoring standpoint, factor analysis might also inform how to best weight CDM cases in order to yield a composite score that most closely reflects the structure of the MCCQE1. Finally, from a score reporting perspective, a better understanding of the underlying structure of the CDM component of the MCCQE1 might also better support current feedback provision mechanisms.

Both the exploratory and confirmatory factor analyses examined in this study suggest that broad discipline domains best account for performance on CDM cases. While a 'lifespan

period'-based CFA model did yield the best comparative fit of the 17-case CDM matrix, it is important to underscore that the latter categorizations are heavily nested within broad disciplines. For example, 'pediatric' and 'adolescent' CDM cases ('lifespan period' categories) are virtually identical to those classified as 'pediatric' cases (within the 'discipline' codes). Similarly, 'medicine' CDM cases are exclusively associated with 'adult' cases. It is consequently not surprising to note that the 'discipline' and 'lifespan period' CFA models provided a similar level of fit to the CDM case response matrix. It is also important to underscore that the models based on 'clinical situation' or 'location/setting' provided substantially worse fit than the competing structures.

The three-factor EFA model appears to suggest that performance on CDM cases relates primarily to broad disciplinary groupings, such as 'medicine,' 'pediatrics,' and 'psychiatry.' It is interesting to note that this structure did not appear to adequately account for performance on CDM cases 6 (contraception/obstetrics-gynecology), 7 (infertility/obstetrics-gynecology/population health and ethical, legal, and organizational aspects of medicine) and 15 (pneumothorax). CDM cases 6 and 7 could be categorized as 'women's health' cases, while case 15 could be conceived as an 'emergency medicine' scenario. Additional analyses with larger case sets could more formally test this hypothesis.

It was also interesting to note that instances of misfit tended to be associated with the same CDM cases, regardless of the model that was examined. Specifically, CDM cases 5 (diabetic pregnancy), 7 (infertility) and 12 (life-threatening asthma) did not tend to be well accounted for by the various models under study, including, to a lesser extent, the EFA structures. It is plausible that the performance on CDM cases 5 and 7 would be better captured by a 'women's health counseling' factor, while CDM case 12 might correspond to an 'emergency medicine' factor, as mentioned above. Future analyses could potentially test this hypothesis more formally. While tempting, it is probably incorrect to wholly ascribe the results of this study to case or content specificity effects, which are reflected by very different performances from case to case due to the specific nature of the problem outlined in a given CDM case [7]. The latter effect is common with performance assessments in general and can severely impact reliability, especially with shorter examinations (which performance assessments are *de facto*, in comparison to multiple choice questions). This finding seems to suggest that broader discipline/patient-age categories best account for performances on CDM cases, which is consistent with similar conclusions drawn from the analysis of standardized patient cases [8].

The results of this study also largely confirm the basic tenets of CDM cases via key features, such as the importance of the clinical presentation and problem in formulating the most appropriate decisions in a given scenario [9]. The practical test development implications of these results are twofold: first, particular effort should be placed on developing CDM cases according to the broad domains of discipline and patient age, with significantly less attention paid to the setting and clinical situation; second, CDM testlets should be assembled largely using the criteria of discipline and patient age. Similarly, from the perspective of calibration and scoring, the use of common item response theory models with CDM case scores appears reasonable if a concerted effort is made to assemble CDM forms in a way that balances the discipline and patient-age categories.

Although informative, the above results need to be interpreted in light of an important caveat; namely, the EFAs and CFAs were based on a restricted set of 17 CDM cases. Nonetheless, these cases likely reflect the levels of the various domains present in the test bank. Furthermore, there is little reason to believe that these findings would differ drastically based on a larger set of cases. However, future analyses should be geared towards replicating the models that appeared to best fit the CDM case matrix in this study. Despite this limitation, the findings in this investigation provide useful initial information on what domains account for performance on CDM cases. These results could provide extremely valuable information in a number of arenas that could lead to the improvement of test assembly, scoring/calibrating, equating, and other processes. In turn, these could enhance the overall quality and defensibility of the MCCQE1 examination.

ORCID: André F. De Champlain: http://orcid.org/0000-0002-2472-798X

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

SUPPLEMENTARY MATERIAL

Audio recording of the abstract.

REFERENCES

1. The Medical Council of Canada. The Medical Council of Canada (MCC) Objectives for the Qualifying Examination [Internet]. Ottawa (ON): Medical Council of Canada; 2011 [cited 2015 Apr 12]. Available from: http://mcc.ca/examinations/objectives-overview/
2. Muthen LK, Muthen B. Mplus: statistical analysis with latent variables: user's guide. Los Angeles (CA): Muthen & Muthen; 2010.

3. Muthen B, du Toit SH, Spisic D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika. 1997;75:1-45.
4. McDonald RP, Mok M. Goodness of fit in item response models. Multivariate Behav Res. 1995;30:23-40. http://dx.doi.org/10.1207/s15327906mbr3001_2
5. McDonald RP. Testing for approximate dimensionality. In: Laveault D, Zumbo BD, Gessaroli ME, Boss MW, editors. Modern theories in measurement: problems and issues. Ottawa (ON): Edumetrics Research Group; 1994. p.63-86.
6. Gessaroli ME, de Champlain AF. Test dimensionality: assessment of. In: Everitt BS, Howell DC, editors. Encyclopedia of statistics in behavioral science. Hoboken (NJ): John Wiley & Sons; 2005. p.2014-2021.
7. Linn RL, Burton E. Performance-based assessments: implications of task specificity. Educ Meas: Issues Pract. 1994;13:5-8. http://dx.doi.org/10.1111/j.1745-3992.1994.tb00778.x
8. De Champlain AF, Klass DJ. Assessing the factor structure of a nationally administered standardized patient examination. Acad Med. 1997;72(10 Suppl 1):S88-S90. http://dx.doi.org/10.1097/00001888-199710000-00053
9. Medical Council of Canada. Guidelines for the development of key features problems and test cases [Internet]. Ottawa (ON): Medical Council of Canada; 2010 [cited 2015 Apr 12]. Available from: http://meds.queensu.ca/assets/CDM_Guidelines_e.pdf