Assessing the Underlying Structure of the Medical Council of Canada's Qualifying Examination Part I Clinical Decision Making Cases: A Comparison of Exploratory and Confirmatory Factor Analysis Models

André F. De Champlain, Ph.D.

Report Prepared for the Medical Council of Canada

Tuesday, August 23rd, 2011

Objective

The objective of this report is to:

- Provide a comprehensive account of analyses aimed at comparing the fit of a number of exploratory and confirmatory factor analytic models to the 2010 combined spring and fall Medical Council of Canada Qualifying Examination Part I (MCCQEI) Clinical Decision Making (CDM) item response matrix using the M*plus*® software package (Muthén & Muthén, 2010)

The results of this investigation provide useful information to guide a number of psychometric efforts relating to CDM cases, including how to best score and calibrate this component of the MCCQEI.

Medical Council of Canada Qualifying Examination Part I (MCCQEI)

Overview of Examination

The MCCQEI is a two-part, computer-based examination, which assesses the knowledge, skills and attitudes judged essential for entry into supervised post-graduate medical training (c.f. MCC objectives for specific statement of objectives; Medical Council of Canada, 2011).

The first part of the examination includes 196 five-option, single-best-answer (A-type) multiple choice items. These 196 multiple-choice questions (MCQ) are further allocated into seven sections of 28 items. The second part of the MCCQEI is composed of about 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 MCCQEI provide a measure of problem-solving and decision-making skills of candidates as they pertain to the specific clinical scenario.

The MCCQEI 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 MCQ portion of the MCCQEI whereas up to four hours are allocated for the completion of CDM cases.

Methods

MCCQEI Cohort

The present investigation focused on the combined spring and fall 2010 MCCQEI 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 MCCQEI administrations. A breakdown of the cohort, by training (i.e. CMG vs. IMG) and test administration is provided in Table 1. The bulk of the 2010 combined cohort is composed of CMGs (2429 or 60.2%) and does conform to expected cyclical patterns, i.e., CMGs largely comprise the spring 2010 MCCQEI administration, whereas IMGs largely test in the fall test administration window.

MCCQEI Bank

The bank of multiple-choice (MCQ) items available for the combined 2010 MCCQEI administrations included 2781 items. One hundred and eight 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 are presented with case descriptions followed by one or more test questions that assess key clinical issues in the resolution of the case. The latter might entail eliciting clinical information, ordering diagnostic procedures, making diagnoses or prescribing therapy.

Examinee responses (i.e. decisions) reflect the management of an actual patient. 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 M*plus*®. Table 2 provides a breakdown of CDM cases by score response categories. As shown in Table 2, the majority of CDM cases had either two or three response categories (68 cases or 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 M*plu*s, final analyses were conducted on a set of 17 CDM cases, culled from the original bank of 108 cases. The cases were representative of the bank with respect to a number of classification variables. More detail on these CDM cases is provided in the next section of the report.

Analyses

All analyses were carried out using the structural equation modeling software program M*plus*® (Muthén & Muthén, 2010).

Initially, the fit of one- to five-factor exploratory 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 test statistic (that use a full weight matrix), was implemented (Muthén, du Toit & Spisic, 1997). The latter estimation method is appropriate with data that violate assumptions of more common methods (e.g.: normality assumption underlying maximum likelihood and generalized least-squares estimation).

The second set of analyses focused on fitting a number of confirmatory factor analytic (CFA) models 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: (1) a 3-F *Location/Setting* model; (2&3) 3-F and 4-F *Lifespan Period* models; (4) a 4-F *Clinical Situation* model; and (5) a 5-F *Discipline* model. Table 3 provides a breakdown of the 17 CDM cases as a function of these classifying variables.

The 3-F *Location/Setting* model posited the following factor structure: Factor 1 (*Family Physician Office*) loaded on CDM cases 1,2,3,6,7,8,9,10,12,13; Factor 2 (*General Hospital*) loaded on CDM cases 4,11; and, Factor 3 (*Emergency Department*) loaded on CDM cases 5,14,15,16,17. The 4-F *Lifespan Period* model posited the following factor structure: Factor 1 (*Adult*) loaded on CDM cases 1,2,3,7,15,16,17; Factor 2 (*Pediatrics*) loaded on CDM cases 9, 10, 13; Factor 3 (*Adolescent*) loaded on CDM cases 6,8,12,14; and, Factor 4 (*Pregnancy/ Neonatal/Infant*) loaded on CDM cases 4,5,11. A 3-F modified version of the latter CFA model was also examined, where Factor 2 (Pediatrics/Pregnancy/Neonatal/Infant) loaded on CDM cases 4,5,9,10,11,13, based on exploratory correlational analyses (the remaining factor structure was identical to the 4-F *Lifespan Period* model). The 4-F *Clinical Situation* model posited the following factor structure: Factor 1 (*Undifferentiated Complaint*) loaded on CDM cases 1,7,10; Factor 2 (*Single Typical Pr*oblem) loaded on CDM cases 2,4,8,13,15,16; Factor 3 (*Preventive Care and Health Promotion*) loaded on CDM cases 6,8,12,14; and, Factor 4

(*Multiple problem or Multi-system Life-threatening Event*) loaded on CDM cases 3,6,9. Finally, the 5-F *Discipline* model posited the following factor structure: Factor 1 (*Medicine*) loaded on CDM cases 1,2,3; Factor 2 (*Obstetrics-Gynecology*) loaded on CDM cases 4,5,6,7; Factor 3 (*Pediatrics)* loaded on CDM cases 8,9,10,11,12; Factor 4 (*Psychiatry*) loaded on CDM cases 13,14; and, Factor 5 (*Surgery*) loaded on CDM cases 15,16,17.

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/indices: (1) chi-square test of model fit; (2) Comparative Fit Index (*CFI*); (3) Tucker-Lewis Index (*TLI*); (4) Root Mean Square Error of Approximation (*RMSEA*). 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 covariances among all input indicators are fixed to zero or no relationship among variables is posited, i.e., where the number of dependent variables is equal to the number of factors. The *TLI* additionally imposes a correction for overparameterization. *CFI* and *TLI* values range from 0 to 1 (though the *TLI* can exceed 1 with severe over-fitting), with values of .90 or above indicating "acceptable" fit (McDonald & Mok, 1995). Similarly, *RMSEA* of 0.06 or recommended cutoffs for "acceptable" model fit.

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 (McDonald, 1994). 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 4 provides fit statistic values for the five EFAs that were examined in this study. Based on these results, it appears that a 3-F EFA solution provided the best fit of the item response matrix, without over-fitting (which is clearly the case for the 4-F and 5-F models based on CFI and TLI values). This 3-F obliquely rotated factor loadings are provided in Table 5. Using a rough cutoff of .25 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 3 Medicine cases, 2 biomedically oriented OBGYN cases and 2 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 & 14. This 3-F EFA model appears to suggest that performance on CDM cases relates primarily to broad discipline groupings, i.e. *Medicine* (in the most general sense), *Pediatrics* and *Psychiatry*. It is interesting to note that this structure did not appear to adequately account for performance on CDM cases 6 (*Contraception-OBGYN*), 7 (*Infertility-OBGYN/PHELO*) and 15 (*Pneumothorax*). CDM 6 &7 cases 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 cold more formally test this hypothesis.

Finally, the correlations between the three factors were quite low, ranging from -.04 (F1-F3) to .09 (F2-F3), suggesting that distinct competencies are required to perform well on each type of CDM case.

Confirmatory Factor Analyses

Table 6 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 as well as inter-factor correlations for the 4-F *Lifespan Period* CFA model are provided in Tables 7 and 8. Most the (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 5-F *Discipline* CFA model are provided in Tables 9 and 10. 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 (*OBGYN*), 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 *OBGYN*).

Discussion

Assessing the underlying structure of any item response matrix is critical to both test development and psychometric efforts. From a test development standpoint, the latter analyses can provide substantiating evidence both with respect to blueprinting and test design activities. From a psychometric perspective, the use of advanced modeling techniques (e.g. item response theory) is predicated on a clear understanding of the data structure that is being analyzed. While common IRT models assume unidimensionality of the underlying latent ability space, research has shown that the latter are robust to departures from this assumption, as long as the composite of proficiencies is comparable across test forms (Gessaroli & De Champlain, 2005). From a scoring standpoint, factor analysis might also inform how to best weight CDM cases to yield a composite that most aptly reflects the structure of MCCQEI. Finally, from a score reporting perspective, a better understanding of the underlying structure of the CDM component of the MCCQEI might also better support current feedback provision mechanisms.

Both exploratory and confirmatory factor analyses that were examined in this investigation 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 of the CDM case response matrix. It is also important to underscore that models based on *Clinical Situation* or *Location/Setting* provided substantially worse fit than competing structures.

It was also interesting to note that misfit tended to be associated with the same 2-3 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 previously noted in this report. Future analyses could be aimed at more formally testing this hypothesis.

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 (Linn & Burton, 2005). The latter effect is common with performance assessments in general and can severely impact reliability, especially with shorter examinations (which *de facto* performance assessments are, in comparison to MCQs). Our findings seem to suggest that broader discipline/patient age categories best account for performances on CDM cases. As such, these findings are consistent with similar conclusions drawn with standardized patient cases (De Champlain & Klass, 1998).

The results of this study also largely confirm the basic tenets of clinical decision making cases via key features, i.e., the importance of the clinical presentation and problem in formulating most appropriate decisions for a given scenario (Medical Council of Canada, 2010).

The practical (test development) implications of these results are twofold: (1) particular effort should be placed on developing CDM cases according to broad discipline and patient age domains with significantly less attention to setting and clinical situation; (2) CDM testlets should be assembled largely using these two constraints (again, discipline and patient age).

Similarly, from calibration and scoring perspectives, the use of common IRT models with CDM case scores appears reasonable if there is a concerted effort to assemble CDM forms to balance discipline and patient age categories.

Although informative, our results need to be interpreted in light of an important caveat, namely that both the EFAs and CFAs were based on a restricted (17) CDM case set. Nonetheless, these cases generally reflect levels of the various domains and there is thus little reason to believe that findings would differ drastically across 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 (along with those from additional studies) could provide 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 MCCQEI examination.

References

1. De Champlain, A.F. & Klass, D.J. (1997). Assessing the underlying structure of a nationally administered standardized patient examination. *Academic Medicine*, *72*, s88-s90.
2. Gessaroli, M.E. & De Champlain, A.F. (2005). Test dimensionality: Assessment of. *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ: John Wiley & Sons.
3. *Guidelines for the Development of Key Features Problems and Test Cases*. (2010). Ottawa, ONT: Medical Council of Canada.
4. Linn, R.L., & Burton, E. (2005). Performance-based assessments: implications of task specificity. *Educational Measurement: Issues and Practice*, *13*, 5-8.
5. McDonald, R.P. (1994). Testing for approximate dimensionality. In Laveault D, Zumbo BD, Gessaroli, ME, Boss MW (eds). *Modern theories in measurement: Problems and is*sues. Ottawa: Edumetrics Research Group, 1994:63–86.
6. McDonald, R.P. & Mok, M. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30, 23–40.
7. Muthén, B.O. & Muthén, L.K. (2010). *Mplus®* - Statistical analysis with latent variables – User's guide. Los Angeles, CA: Muthén & Muthén.
8. Muthén, B.O., du Toit, S.H.C. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report.
9. *Objectives for the qualifying examination*. Ottawa, ON: The Medical Council of Canada, 2011.

Table 1

Breakdown of 2010 MCCQEI First-time Examinee Population by Training and Test Administration

Training	Test Administration		
	Spring	Fall	Total
CMGs	2407 (59.6%)	22 (0.6%)	2429 (60.2%)
IMGs	790 (19.6%)	817 (20.2%)	1607 (39.8%)
Total	3197 (79.2%)	839 (20.8%)	4036 (100%)

Table 2

One Hundred and Eight MCCQEI CDM cases by Response Categories for the Spring and Fall 2010 Test Administrations

Response Categories	Frequency of CDM Cases
2	38 (35.2%)
3	30 (27.8%)
4	20 (18.5%)
5	14 (13.0%)
6	5 (4.6%)
7	0 (0.0%)
8	1 (0.9%)

Table 3

17 CDM Cases by Location/Setting, Lifespan Period, Clinical Situation and Discipline

Case	Location/Setting	Lifespan Period	Clinical Situation	Discipline
1	Family Physician Office	Adult	Undifferentiated Compl.	Medicine
2	Family Physician Office	Adult	Single Typical Problem	Medicine
3	Family Physician Office	Adult	Prevent. Care/Health Pr.	Medicine/PHELO
4	General Hospital	Pregnancy/Neonatal	Single Typical Problem	OBGYN
5	Emergency Department	Pregnancy/Neonatal	Multiple Problem Event	OBGYN
6	Family Physician Office	Adolescence	Prevent. Care/Health Pr.	OBGYN
7	Family Physician Office	Adult	Undifferentiated Compl.	OBGYN/PHELO
8	Family Physician Office	Adolescence	Single Typical Problem	Pediatrics
9	Family Physician Office	Pediatric	Prevent. Care/Health Pr.	Pediatrics
10	Family Physician Office	Pediatric	Undifferentiated Compl.	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

Table 4

Goodness of Fit Statistics for Five Exploratory CDM Factor Analytic Models

Model	X^2	GFI	TLI	RMSEA
1-F	260.23, $p<0.001$	0.65	0.60	0.07
2-F	138.56, $p=0.01$	0.91	0.88	0.05
3-F	88.83, $p=0.46$	1.00	1.00	0.05
4-F	64.08, $p=0.79$	1.00	1.05	0.04
5-F	44.03, $p=0.95$	1.00	1.10	0.04

Table 5

Obliquely Rotated Factor Loadings for 3-F CDM Exploratory Solution

CDM 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 6

Goodness of Fit Statistics for Five Confirmatory CDM Factor Analytic Models

Model	X^2	GFI	TLI	RMSEA
3-F *Location/Setting*	259.53, $p<0.001$	0.64	0.58	0.02
4-F *Lifespan*	185.86, $p<0.001$	0.82	0.79	0.01
3-F *Modified Lifespan*	221.07, $p<0.001$	0.74	0.69	0.02
4-F *Clinical Situation*	302.21, $p<0.001$	0.53	0.44	0.02
5-F *Discipline*	180.44, $p<0.001$	0.83	0.78	0.02

Table 7

Factor Loadings for 4-F CDM Lifespan CFA Model

CDM Case	F1 (Adult)	F2 (Pediatrics)	F3 (Adolescent)	F4 (Pregnancy/Neonatal)
1	0.47*			
2	0.36*			
3	0.20*			
4				0.35*
5				0.21
6			0.27*	
7	0.09			
8			0.44*	
9		0.14*		
10		0.49*		
11				0.30*
12			0.10	
13		0.23*		
14			0.52*	
15	0.27*			
16	0.33*			
17	0.38*			

* *p*<0.02

Table 8

Inter-factor Correlation Matrix for 4-F CDM Lifespan CFA Model

	Adult	Pediatrics	Adolescent	Pregnancy/Neonatal
Adult	1.00	0.16	0.32*	0.92*
Pediatrics		1.00	0.90**	-0.04
Adolescent			1.00	0.22
Pregnancy/Neonatal				1.00

* p<0.01; ** correlation was fixed at .90 based on estimation difficulties (multicollinearity)

Table 9

Factor Loadings for 5-F CDM Discipline CFA Model

CDM Case	F1 (Medicine)	F2 (OBGYN)	F3 (Pediatrics)	F4 (Psychiatry)	F5 (Surgery)
1	0.57*				
2	0.41*				
3	0.20*				
4		0.34*			
5		0.11			
6		0.24*			
7		0.13			
8			0.54*		
9			0.22*		
10			0.54*		
11			0.20*		
12			0.05		
13				0.20*	
14				0.90**	
15					0.30*
16					0.42*
17					0.40*

* p<0.02; ** loading was fixed at 0.9 due to estimation difficulty.

Table 10

Inter-factor Correlation Matrix for 5-F CDM Discipline CFA Model

	Medicine	OBGYN	Pediatrics	Psychiatry	Surgery
Medicine	1.00	0.91*	0.39*	0.07	0.61*
OBGYN		1.00	0.19	0.35*	0.59*
Pediatrics			1.00	0.41*	0.19
Psychiatry				1.00	0.14
Surgery					1.00

* p<0.02