Detection of Differential Test Functioning (DTF) and Differential Item Functioning (DIF) in MCCQE Part II Using Logistic Models

Jin Gong
University of Iowa
June, 2012

Background

The Medical Council of Canada Qualifying Examination Part II (MCCQE Part II) is an Objective Structured Clinical Examination (OCSE) taken by medical doctors who have completed at least 12 months of postgraduate education. Passing the MCCQE Part II is a prerequisite for a medical licensure in Canada.

The MCCQE Part II consists of 12 scored stations, which are either 10-minute encounters with standardized patients or 5-minute encounters with standardized patients followed by 5 minute post-encounter-probe which is a series of questions the candidate has to answer. Candidates are observed and evaluated as they go through these stations by physician examiners. In each station, examiners rate candidate performance using a combination of checklist scores, rating scales, and global rating scores. The examination cut score of MCCQE Part II is the sum of the 12 station cut scores plus one SEM, and the station cut scores are established using a modified borderline method. The candidates have to meet two criteria to pass the examination:

1) Obtain a total score higher than or equal to the examination cut score; and
2) Pass at least 8 out of the 12 scored stations.

Although the raters are asked to base their global rating scores purely upon the clinical competence that the candidate presents in the exam, due to the subjective nature of personal judgment, the stations global ratings made by the raters could be biased. When such biases are associated with the "group" the candidates are in, the item, or exam station, would be labeled with "Differential Item Functioning" (DIF). In other words, if candidates with equal ability, but from different groups, have an unequal probability of

passing a certain test station, DIF exits at this station. The presence of DIF can lead to an unfair advantage or disadvantage for certain subgroups in the test.

DIF does not mean simply that an item is harder for candidates in one group than for another group. If candidates in one group tend to be more capable than candidates in the other group, they tend to perform better on all the test items. A criterion is used to match the candidates from either group to ensure that they are at the same ability level. For this reason, the criterion in a DIF analysis is often referred to as a matching criterion. The most commonly used matching criterion is the total score of the candidate in the exam, which is used as an estimate of the candidate's ability.

There are two different types of DIF- uniform and non-uniform DIF (Mellenbergh, 1982). Uniform DIF exists when the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability: There is no interaction between ability level and group membership. Non-uniform DIF exists when differences in the probability of answering the item correctly varies across all levels of ability for any group: There is interaction between ability level and group membership.

Differential Test Functioning, or DTF, is the analogous procedure to DIF, however it is used at the exam level rather than the test item level. DTF is arguably more important because it impacts the final decision on the candidate's performance (DIF on one item may be significant, but might not have too much practical impact on test results).

For a high-stake licensure test as MCCQE Part II, the test results should be based upon evidence relevant to the target clinical knowledge and skills that the candidate presented during the exam, but not upon things such as demographic profile, educational

background, and exam history of the candidate. It is crucial to minimize the impact of DIF and DTF in the scoring and decision making process to help ensure the fairness of MCCQE Part II.

Purpose of Research and Data

This study includes two parts. Part I, the detection of potential DTF at the exam level of MCCQE Part II; Part II, the detection of DIF at the level of exam station components of MCCQE Part II. Both non-uniform and uniform DIF/DTF were examined in the study.

The data used in the study were collected from the records of 2,644 candidates who took a recent Fall administration of the MCCQE Part II. The pass/fail statuses of the whole examination, and the pass/fail status of stations, were used as outputs of the rater's "subjective rating". As argued by Mazor, Kanjee, and Clauser (1995), more accurate matching of candidates on ability could be achieved by using multiple relevant ability estimates, and thus multidimensional item impact was not mistakenly identified as DIF. In this study, multiple sub-scores instead of one total score were used to match the candidates from different groups, for both DTF and DIF parts of this study.,. The standardized sub-scores, including the patient interaction score (STDPI), the data acquisition score (STDDA), the problem solving score (STDPS), and the CLEO score (STDCLEO) were used as matching criterion for the DTF part of this study. The station component scores (the checklist portion and the written portion for a couplet station, and the communication component and the checklist component for a ten-minute station) were used as the matching criterion for the study of DIF for couplet stations and ten-minute stations, respectively.

The grouping factors of interests included the candidate's demographic characteristics, his/her educational background, etc. Table 1 shows a summary of the names, definition and possible values of the grouping variables as well as the dependent variables and the matching criterion for the DTF and DIF studies.

Methods

Many methods for detecting DIF have been proposed. The most commonly used methods for dichotomous outcomes are the Mantel- Haenszel (M-H) test, the logistic regression model, and the Item Response Theory (IRT) based method. The IRT-based method compares the item characteristic curves (ICC) of the item between different groups to detect DIF. It has the limitations of strict model assumptions (uni-dimensionality and local independence of items), and there is no statistical test associated with the area between two ICCs. Many simulation studies have been done to compare the M-H test and logistic regression methods. Swaminathan and Rogers (1990) found that the logistic regression procedure is more powerful than the Mantel- Haenszel procedure for detecting non-uniform DIF, and as powerful in detecting uniform DIF. Moses, Miao and Dorans (2010) compared four methods of detecting non-uniform DIF and concluded that logistic regression was the most recommended strategy in terms of the limiting bias and variability of its estimates. The logistic model was used in this study since we are interested in examining both uniform and non-uniform DIF and DTF.

To examine the extent of DTF (research question 1), a logistic regression model was built at the exam level, The pass/fail status of the exam was selected as the response variable in the logistic model. The four standardized sub-scores (patient interaction (STDPI), data acquisition (STDDA), problem solving (STDPS), and Considerations of

the Cultural-Communication, Legal, Ethical, and Organizational aspects of the practice of medicine (STDCLEO)) were used as matching criteria that represent evidence of candidate's performance on the exam.

To examine the extent of DIF (research question 2), logistic regression models were created for three couplet stations (01, 02 and 05), and six ten-minute stations (01, 03, 04, 05, 07 and 08). The station-level indicator of pass/fail status was used as the response variable in the model for a given station. The corresponding component scores (checklist portion and written portion for a couplet station, and communication component and checklist component for a ten-minute station) were used as the matching criteria that represent evidence of candidate's performance on a particular station.

The grouping indicators that were examined for both DTF and DIF analyses included the remaining factors of interest, e.g., the examinee's demographic information (gender, age), education background (country of post graduate training, first time or repeat taker, years since graduation from medical school), and information of the examination administration (language of the exam).

The evaluation of the logistic models included three parts, the overall model evaluation, statistical tests of individual predictors, and goodness-of-fit statistics. Agresti (2002) has provided a thorough discussion of these methods.

Model evaluation examines if the additional predictors in the logistic model demonstrates significant improved fit to the observed data over the intercept-only model (also called the null model). Three tests, the likelihood Ratio, the Wald test and the Score tests, were used to assess improvement. They test against the null hypothesis that at least one of the predictors' regression coefficients is not equal to zero in the model.

The parameters and corresponding odds ratios were estimated using the method of maximum likelihood (ML). All three test statistics are distributed as chi-squares with degrees of freedom equal to the number of predictors.

Individual parameter estimates were tested by the likelihood ratio test, the Wald test and the Score test, to evaluate the contribution of the individual independent variable to the variation of the dependent variable. An independent variable that is significant at .05 level suggests that the variable is significantly associated with the dependent variable. If a grouping indicator is significant, this suggests that corresponding uniform DTF (or DIF) may be present and a review of examiner training might be indicated.

First order interaction terms of significant grouping indicators paired with matching criteria were added to the model to test the non-uniform DTF (or DIF) effects. They were evaluated by the Wald Chi-square test. If the contribution brought in by an interaction term is statistically significant, then there is some evidence that such an interaction term may be important.

After the model is created, it can be used to predict the probability of a "passing" event by the set of predictors. The probabilities can be revalidated with the actual outcome to determine if high probabilities are associated with events ("passing") and low probabilities with nonevents ("fail"). The degree to which predicted probabilities match with actual outcomes is expressed as a measure of association, and can be measured by a set of statistics, e.g., Percent Concordant, Tau-α, Gamma, Somers' D and the *c* statistics. For a detailed summary of these statistics please refer to Peng and So (2002).

In a summary, the data analyses used to investigate the presence of DTF and DIF in this study followed three steps :

Step 1. A initial logistic regression model was created with all grouping indicators and matching criteria, but no interaction terms. Significance tests for both the overall model and individual predictor variable were performed. Estimate of parameters and odds ratios were obtained.

Step 2. First-order interaction terms of any significant grouping indicator identified in step 1 paired with every matching criteria were added to the initial model to test non-uniform DTF (or DIF). Significance tests of individual interaction term were performed.

Step 3. If there is any significant interaction term found in Step 2., then new estimates of parameters and odds ratios were obtained from the model with these interaction terms. Otherwise the estimates of parameters and odds ratios obtained in Step 1 would be kept.

Results

Part I: DTF Analysis for the exam

The logistic regression model for DTF analysis was performed using the SAS LOGISTIC procedure.

Table 2 shows the descriptive statistics for the continuous independent variable crossed with the dependent variable (exam status). Table 3 shows the frequency distribution of the categorical independent variables in the model crossed with the dependent variable (exam status).

The overall logistic regression model for DTF was significant at the 5% level, as assessed by all three tests (Table 4). This indicates that including all the independent variables in the model improved the data fit, compared to the model with the intercept only.

Table 5 shows the estimated individual parameters and corresponding *p*-values of each predictor in the model. At 0.05 confidence level, age, language, and the four sub-scores are all significant predictors of examination status, while gender, first time or repeat taker, country of post graduate training, the number of years since graduation from medical school, and country of medical degree are not. The parameter estimates for each of the four standard sub-scores are positive, indicating that, after controlling for other factors in the model, the higher the standardized score, are the more likely the candidate would pass the exam. The estimated parameter for the age predictor is positive. This indicates that older candidates are more likely to pass the exam compared to younger candidates, controlling all other factors in the model. The variable language (English vs. French) has a negative estimated parameter, which indicates that candidates who took the exam in English are less likely to pass it compared to candidates who took the exam in French, after controlling the confounding effects of all other predictors in the model. The remaining factors (gender, first-time candidate or not, PGMT country, years since graduation, and university country) did not significantly predict examination status.

Table 6 shows the estimated odds ratio and corresponding 95% confidence limits for each factor. Note that each of the non-significant factors includes 1 within the confidence limits. The units for interpreting the odd ratio point estimates for continuous variables are ratios to 1. For example, a point estimate of 1.09 for "age" means that the

odds of a candidate passing the exam is 1.09 times as large as the odds of a candidate who is 1 year younger, given that all other factors in the model are the same. A point estimate of .36 for "language" means that the odds of passing the exam for a candidate who takes the exam in English is .36 times smaller than for a candidate who takes it in French, given all other factors. Table 7 lists the point estimates of odds ratios when the units are set to 5 for age, and 10 for the standardized scores.

Table 8 shows the statistics of association between the predicted probabilities using the model and the observed responses. Values indicate that the predictive ability of the model is adequate. The percentage of concordant pairs is high (98.3%). The Gamma statistic is .98, which is interpreted as 98% fewer errors made in predicting if the candidate would pass the exam by utilizing the estimated probabilities from this model, compared to chance alone. The statistic "c" is the area under the Receive Operating Characteristic (ROC) curve, which ranges from 0.5 to 1, where a 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response. The c statistic for this model is .99, which is adequately high.

To investigate if the presence of DTF related to language and candidate's age was uniform or variable along the ability construct, interaction terms for every sub-score and exam language or candidate's age were added to the model. Table 9 shows the Wald Chi-square test of these interaction terms. None of them were significant at 0.05 level. There is no evidence that the degree of DTF from either exam language or age depends on the candidate's ability.

In summary, some uniform DTF from exam language and age of candidate was detected in the sample data. Controlling all the four sub-scores, those who took the exam in English are less likely to pass the exam, compared to those who took it in French, and older candidates are more likely to pass it, compared to those who are younger. There is no evidence that these patterns depend on ability.

Part II: DIF Analysis for stations

Similar logistic models to those used in Part I of this study were built up for all stations with complete data to investigate the presence of DIF. For these analyses, station status (pass/fail) is used as the dependent variable of the logistic models. The checklist portion score and the written portion score of a couplet station were used as matching criteria for couplet stations. The communication component score and the checklist component score of a ten-minute station were used as the matching criteria in models for the ten-minute station. The grouping factors examined for DIF in the model are the same as the factors in the model for the exam status.

The results of the model fitting for all the 9 stations are summarized in table 10. The station type, station number, significant grouping indicators of the station (if any), significant interaction term (if any), type of DIF detected, estimated parameters of the significant grouping indicators, corresponding estimated odds ratios and the 95% confidence limits of the odds ratios are presented in table 10.

Among the three couplet stations, DIF was detected in station 1 and 5, but not in 2. For couplet station 1, candidates who took the exam in English were less likely to pass this station, controlling after all other factors examined in the study. In other words,

station 1 seemed to be more difficult to the candidates who took it in English, even after adjusting for differences on component scores.. However, for couplet station 5, the DIF effect of exam language was in the opposite direction. Candidates who took the exam in English were more likely to pass the station, controlling after all other actors in the study. In addition, this station seemed to be easier to pass to the first-time candidates, compared to candidates who have taken it before, controlling after all other factors in the study.

Among the six ten-minute stations, DIF was detected in three of them. Both station 3 and 4 seemed to be easier to pass for candidates who took it in French, while station 4 also appeared to be more difficult for candidates who finished their post-graduate training in Canada. DIF due to age was detected in station 8. Elder candidates seemed to have higher likelihood of passing this station, even after controlling for differences in component scores for this station. Last but not least, these are no significant interaction terms in all the station-level models. This indicates that all the DIF effects detected are uniform rather than non-uniform.

Conclusion

The presence of both DTF and DIF were detected in MCCQE Part II sample data investigated. Although the evidence of DIF was found to vary by station and factor, exam language seems to be the most consistent. DIF was flagged in 5 of the 9 stations studied. Three stations appeared to be easier to pass for candidates who took the exam in French, compared to those who took it in English, even after controlling for differences in component scores for those stations. At the exam level, similar effects from exam language were detected. Another flag for DIF and DTF was related to candidate's age.

Older candidates appeared to have higher chance of passing both the ten-minute station 8 and the whole exam, even after controlling for differences in component scores for the station, or similar sub-scores at the exam level.

This study flagged possible sources of DIF and DTF for some test stations and total scores on the MCCQE Part II. Further review of these "DIF" stations by content experts is suggested to distingusish DIF from impact, to propose sources of DIF effects, and to suggest modification of cases or training of examiners so DIF can be re-evaluated on a future exam.

The author proposes continued monitoring of MCCQE Part II so that evidence of DIF can be collected periodically and ongoing improvement can be made to cases and rater training.

With regard to the two consistent flags for potential bias, it may be helpful to note that age and language bias were in the expected direction (i.e., the same as for other MCC exams). This study contributes to our confidence that demographic factors do not constitute an important indicator reflected in the pass or fail decisions made with the MCCQE Part II scores.

Limitations of this study

An alpha level of 0.05 was used for significance tests of estimated parameters in this study. Considering the fact that there were 7 grouping indicators being tested in each model, the corrected chance of finding a significant effect of grouping indicator by chance alone would be inflated from 5% to 30%. If an adjusted significance level, such as 0.01, is used instead of 0.05, none of the effects tested would be have been significant.

Eighty-six percent of the cases in the data have a "passing" status of the whole exam, whereas only 14% of the cases failed the exam. Such skewed data may have introduced some bias to the parameter estimates. A possible alternative model to accommodating this issue is the Exact Logistic model. Some data mining techniques such as bagging and boosting could be other solutions to consider.

The size of contrasting samples, and number of total examinees limited the detection of the potential source of DTF and DIF examined in this study. It might be useful to examine other sources of DTF and DIF in the MCCQE Part II that were not covered in this study (e.g., the rater's ID, the test center). Alternative methods for DIF flags might be appropriate. One option might use logistic model with raters as a random effect. Another possible model for DIF would combine the values of variables based on actual meanings to form fewer categories for factors in the logistic regression modeling. It could be a topic of the future study.

Another possible issue for this study is the existence of outliers, which should be identified and removed before the models were created. However, due to the limited sample size and sparse data structure in this study, diagnostic analysis of outliers was not performed. It could be an extension study when more data were collected in the future.

Table 1 Summary of variables

Role in model	Variable Name	Definition	Possible values
Dependent variable for DTF study	Exam Status	Candidate's pass/fail status on the exam	"pass", "fail"
Matching criterion for DTF study	STDSPI	Standardized Patient Interaction Score	50-950
	STDDA	Standardized Data Acquisition Score	50-950
	STDPSDM	Standardized Problem Solving Score	50-950
	STDCLEO	Standardized CLEO Score	50-950
Dependent variables for DIF study (couplet stations)	P_F_Couplet	The pass/fail status on a couplet station	"pass", "fail"
Matching criterion for DIF study (couplet stations)	SCO_AC	The score for the checklist portion of a couplet station	
	SCO_BC	The score for the written portion of a couplet station	
Dependent variables for DIF study (ten minute stations)	P_F_Ten minute	The pass/fail status on a ten minute station	"pass", "fail"
Matching criterion for DIF study (ten minute stations)	SCO_TC	The score for the communication portion of a ten-minute station	
	SCO_TT	The score for the checklist portion of a ten-minute station	
Grouping indicators for both DTF and DIF studies	Gender	Candidate's gender	"male","female"
	Age	Candidate's age when he/she took the exam	
	Language	The language that the candidate used to take the exam	"English", "French"
	UniversityCountry	Country of university where the candidate received his/her medical degree	"Canada", "other"
	PGMTCountry	Country where the candidate attended post-graduate training	"Canada", "other"
	YearAfterGrad	Years since candidate graduated from medical school	
	FT_RP	First time or repeat taker	"FT", "RP"

Table 2 Descriptive Statistics for Continuous Variables

Descriptive Statistics for Continuous Variables					
Variable	Exam Status	Mean	Standard Deviation	Minimum	Maximum
Age	fail	37.66	8.54	24	64
	pass	30.17	5.64	23	59
	Total	31.22	6.65	23	64
STDPI	fail	353.80	95.03	53	576
	pass	523.72	78.28	233	720
	Total	500.01	99.99	53	720
STDDA	fail	355.04	85.46	-22	573
	pass	523.52	80.45	269	770
	Total	500.00	99.98	-22	770
STDPS	fail	371.99	87.56	98	609
	pass	520.63	85.41	224	789
	Total	499.88	99.99	98	789
STDCLEO	fail	372.41	103.64	55	675
	pass	520.71	82.55	196	742
	Total	500.01	100.00	55	742
Year After Grad	fail	11.33	9.22	1	39
	pass	3.17	5.29	1	35
	Total	4.31	6.62	1	39

Table 3 Frequency Distribution of Class Variables

Class	Value	Exam Status		Total
		Fail	Pass	
Gender	female	115	1250	1365
	male	254	1025	1279
Language	English	354	1892	2246
	French	15	383	398
Repeat taker	First	245	2108	2353
	Repeat	124	167	291
PGMT	Canada	142	1969	2111
	other	227	306	533
University Country	Canada	90	1789	1879
	other	279	486	765

Table 4 Overall model fit evaluation

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1655.77	11	<.0001
Score	1301.53	11	<.0001
Wald	238.09	11	<.0001

Table 5 Maximum likelihood estimates of parameters in the logistic model for DTF analysis

Analysis of Maximum Likelihood Estimates						
Parameter	Contrast	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-35.63	2.68	176.80	<.0001
Gender	female vs. male	1	-0.05	0.12	0.19	0.6639
Age		1	0.09	0.04	5.60	0.018
Language	Eng vs. Fr	1	-0.51	0.24	4.45	0.0349
FT_RP	first vs. repeat	1	0.03	0.16	0.03	0.8522
PGMT	Canada vs. other	1	0.01	0.19	0.01	0.938
STDPI		1	0.02	0.00	57.28	<.0001
STDDA		1	0.03	0.00	142.16	<.0001
STDPS		1	0.02	0.00	150.51	<.0001
STDCLEO		1	0.01	0.00	55.44	<.0001
Year After Grad		1	-0.06	0.04	2.47	0.116
University Country	Canada vs. other	1	-0.25	0.20	1.66	0.1981

Table 6 Odds ratio estimates for effects in DTF analysis (without interaction term)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gender female vs. male	0.90	0.55	1.46
Age	1.09	1.02	1.18
Language English vs. French	0.36	0.14	0.93
FT_RP first vs. repeat	1.06	0.57	1.99
PGMT CA vs. Others	1.03	0.49	2.18
STDPI	1.02	1.01	1.02
STDDA	1.03	1.02	1.03
STDPS	1.03	1.02	1.03
STDCLEO	1.01	1.01	1.02
Year After Grad	0.94	0.87	1.02
University Country CA vs. Others	0.60	0.28	1.31

Table 7 Estimated odds ratios at customized units for Age and sub-scores

Odds Ratios		
Effect	Unit	Estimate
Age	5	1.56
STDPI	10	1.17
STDDA	10	1.32
STDPS	10	1.28
STDCLEO	10	1.14

Table 8 Evaluation of associations of predicted probabilities and observed responses

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	98.30	Somers' D	0.97
Percent Discordant	1.30	Gamma	0.98
Percent Tied	0.40	Tau-a	0.23
Pairs	839475	c	0.99

Table 9 Wald Chi-square tests of interaction terms for model of DTF study

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
STDCOMM*Language	Eng	1	-0.02	0.01	2.65	0.1033
STDDA*Language	Eng	1	-0.01	0.01	1.29	0.2556
STDPSDM*Language	Eng	1	-0.01	0.01	1.59	0.208
STDCLEO*Language	Eng	1	-0.01	0.01	2.61	0.1062
STDCOMM*age		1	0.00	0.00	0.14	0.7053
STDDA*age		1	0.00	0.00	3.20	0.0734
STDPSDM*age		1	0.00	0.00	1.53	0.2157
STDCLEO*age		1	0.00	0.00	2.90	0.0886

Table 10 Summary of models and DIF detection for stations

Station Type	Station	Significant Grouping Indicators	Significant Interaction	Type of DIF detected	Estimate of Parameter (Standard Error)	Estimate of Odds Ratio	95% Confidence Limits of Odds Ratio	
							Lower Limit	Upper Limit
Couplet	1	Language (E vs F)	None	Uniform	-.72(.25)	0.24	0.09	0.63
	2	None	N/A	None				
	5	Language ((E vs F)	None	Uniform	1.37 (.28)	15.55	5.191	46.60
		First time/repeat (First vs. Repeat)			.76 (.27)	4.56	1.56	13.37
Ten-minute	1	None	N/A	None				
	3	Language ((E vs F)	None	Uniform	-.82 (.29)	0.19	0.06	0.59
	4	Language (E vs F)	None	Uniform	-.64 (.22)	0.28	0.12	0.65
		Country of PGMT (Canada vs. other)			-.62 (.27)	0.29	0.10	0.84
	5	None	N/A	None				
	7	None	N/A	None				
	8	Age	None	Uniform	.35 (.16)	1.42	1.03	1.94

References

Agresti A., (2002). *Categorical Data Analysis* . 2nd ed. Hoboken, New Jersey: Wiley-Interscience.

Mazor, K. M., Kanjee, A., & Clauser, B. E., (1995). Using Logistic Regression and the Mantel-Haenszel with Multiple Ability Estimates to Detect Differential Item Functioning. *Journal of Educational Measurement, 32*(2), 131-144

Mellenbergh, G.J., (1982). Contingency Table Models for Assessing Item Bias. *Journal of Educational Statistics. 7* (2), 105-118

Moses, T., Miao, J., & Dorans, N. J., (2010). A Comparison of Strategies for Estimating Conditional DIF. *Journal of Educational and Behavioral Statistics*. *35* (6), 726-743

Swaminathan, H., & Rogers, H. J., (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*. *27* (4), 361-370