Using Statistical Measures of Differential Item Functioning to Identify Items that Elicit Group Differences on the Medical Council of Canada Qualifying Exam Part I

Mark J. Gierl
Hollis Lai

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Submitted to:
Dr. Krista Breithaupt
Director, Research and Development
Medical Council of Canada

September 14, 2011

INTRODUCTION

The authors of the *Standards for Educational and Psychological Testing* (1999) describe test fairness as a multi-faceted concept that has no single technical definition. Instead, they present four characterizations of test fairness. First, fair tests must be free from bias. Bias occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups. Second, test fairness requires that examinees receive just and equal treatment in the testing process. To achieve this outcome, both the test and the testing context must be considered when scores are interpreted for each examinee and for groups of examinees. Third, test fairness requires equity in the outcomes of testing. That is, examinees must be given the chance to demonstrate their knowledge and skills on the construct the test is designed to measure. Fourth, test fairness implies that examinees in the achievement domain have had the opportunity to learn the content covered on the tests, particularly when test scores are used to make decisions about the examinees. Thus, fairness requires that examinees have an equal opportunity to learn the test material.

The first and third characterizations are important for the study of group differences using differential item functioning (DIF) methods because they are consistent with a view held by many educational measurement specialists that examinees with the *same* standing on the construct the test is intended to measure should, on average, receive the same test score. An item is biased when it yields different test scores or promotes different test score interpretations for members of different groups (e.g., groups with racial, ethnic, language, cultural, gender, disability, or socio-economic status differences). Bias, according to the authors of the *Standards* (1999), is attributed to construct-irrelevant components that differentially affect the test scores for specific groups of examinees. Bias may be content related. For example, if the scoring rubric for a constructed-response item provides the highest score for testwise examinees who provide more information than was actually requested, then less testwise examinees who follow instructions, thereby providing less information, would earn

a lower score. In this case, testwiseness is a construct-irrelevant component. Bias may be response related. For example, if a quantitative reasoning test contains many context-based reading passages, then examinees with weak verbal ability would earn a lower score. In this case, verbal ability is a construct-irrelevant component.

DIF studies are designed to identify and interpret these construct-related components using a combination of statistical and substantive analyses. To conduct DIF analyses, examinees are first divided into two groups, a reference and focal group. Typically, the statistical analysis involves administering the test, matching members of the reference and focal group on a common measure of ability derived from that test, and using statistical procedures to quantify the differences between groups for each test item. An item exhibits DIF when examinees from the reference and focal groups differ in the probability of answering that item correctly, after controlling for the measure of ability derived from the test. Then, the substantive analysis builds on the results of the statistical analysis. With items identified to possess a quantitative bias, DIF items are often scrutinized by expert reviewers (e.g., test developers or content specialists) in an attempt to identify the construct-related components that produce group differences. A DIF item is considered biased when reviewers identify some component, deemed to be irrelevant to the construct measured by the test, that places one group of examinees at a disadvantage.

PURPOSE OF THE PRESENT STUDY

The purpose of the present study is to use four different statistical criteria to identify items that function differentially across five grouping variables that characterize examinees who write the Medical Council of Canada Qualifying Examination Part I (MCC QE Part I). The MCC QE Part I is a credentialing exam, required for the Licentiate. It is written by all medical students seeking entry into supervised clinical practice in postgraduate training programs in Canada. It is a one-day, fixed-length, multi-stage computer adaptive test used to assess the knowledge, clinical skills, and attitudes

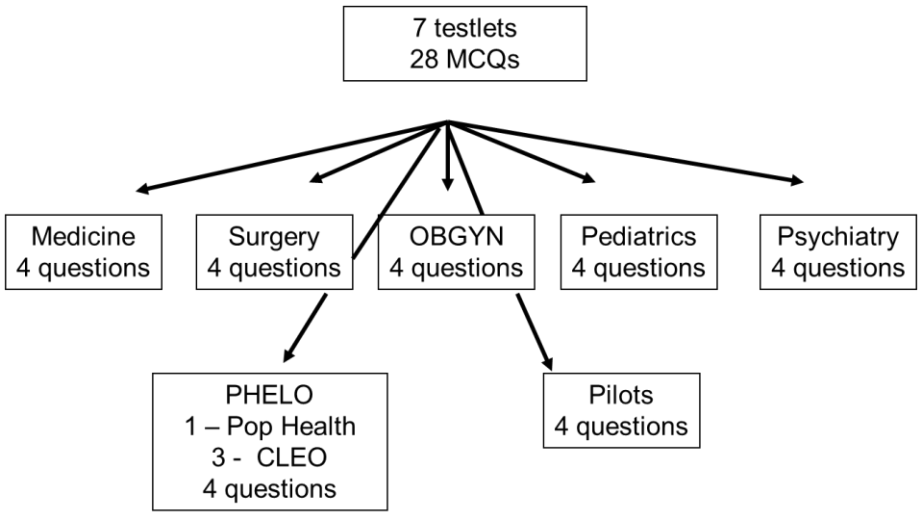
specified by the Medical Council as key objectives and competencies for medical training. The MCC QE Part I is divided into two sections. Section 1 is a 3.5-hour test containing 196 multiple-choice items administered adaptively by computer across six content areas (Internal Medicine; Surgery; Obstetrics and Gynecology; Pediatrics; Psychiatry; and Population Health, including considerations of the legal, ethical, and organizational aspects of the practice of medicine). Section 2 is a 4-hour test containing clinical decision-making items. Our study will focus on DIF detection with a sub-pool of 2806 multiple-choice items representative of the active item bank used with the MCC QE Part I. The focus of the present study is to identify DIF items using statistical analyses. There will be no attempt to interpret the DIF items substantively or account for why DIF occurs.

COMPUTER-ADAPTIVE TESTING AND THE MCC QE PART I

The MCC QE Part I is administered in the spring and fall of each year using a fixed-length, multi-stage, computer adaptive testing process. The MCC QE Part I is composed of six different content areas: Internal Medicine, Obstetrics and Gynecology, Pediatrics, Psychiatry, Surgery, and Population Health. Items in each area are developed by a panel of content experts specializing in each respective content area. Prior to administration, items are classified into four levels of difficulty based on their psychometric properties, and then assembled in four-question testlets according to the item difficulty estimates. In total, 196 items are selected and administered to each examinee across the six content areas using the multi-stage computer adaptive testing process.

The administration requires seven stages of adaptation. For each stage in the adaptive administration process, each examinee writes four-item testlets in six content areas across seven stages, plus one testlet per content area of pilot questions (see Figure 1) for a total of 196 items per examinee per administration.

Figure 1. The teslet structure for one stage of the MCC QE Part I, Section 1.



To adapt the item difficulty level for each examinee, testlets are administered based on the examinees' performance to the previous testlet in the same content area. An example of the administration process is outlined as follows: In the first stage, the six content-area specific testlets each containing four items, one from each of the four difficulty levels (where 1 contains the easiest items and 4 contains the most difficult items), are administered to each examinee. Based on their item-level performance, each testlet is scored and the results are used to determine the difficulty level of the testlet presented in the next stage for each content area. Different rules of adaptation exist for the first stage relative to the second through seventh (see Figure 2a and 2b), but the principle across all seven stages remains the same: stronger examinee performance leads to the administration of testlets with higher difficulty levels, whereas weaker examinee performance leads to the administration of testlets with lower difficulty levels. Because the adaptive process occurs independently for each content area, it is possible that an examinee may receive a more difficult testlet in one content area and an easier testlet in another content area, even within the same stage. Once the items within the testlets across the seven stages are completed, examinees' item responses are scored and scaled using an ability estimate derived from item response theory (IRT). A cut-score is

also set by a panel of content experts to determine the IRT ability level that is deemed satisfactory, and a pass/fail decision is then made for each examinee.

Figure 2a. An illustration of the adaptive process for the content area of Medicine at Stage 1.

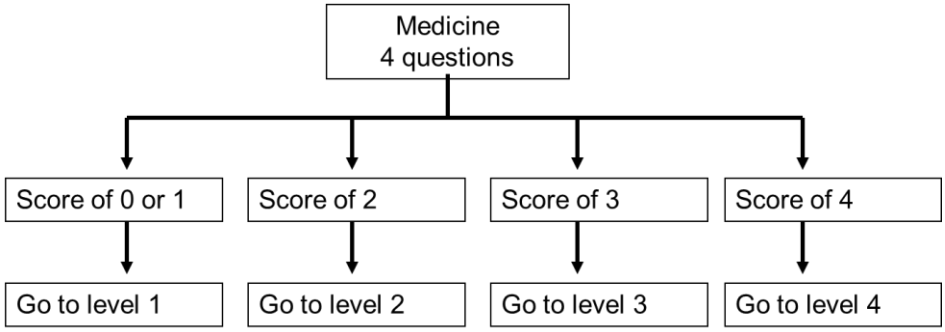
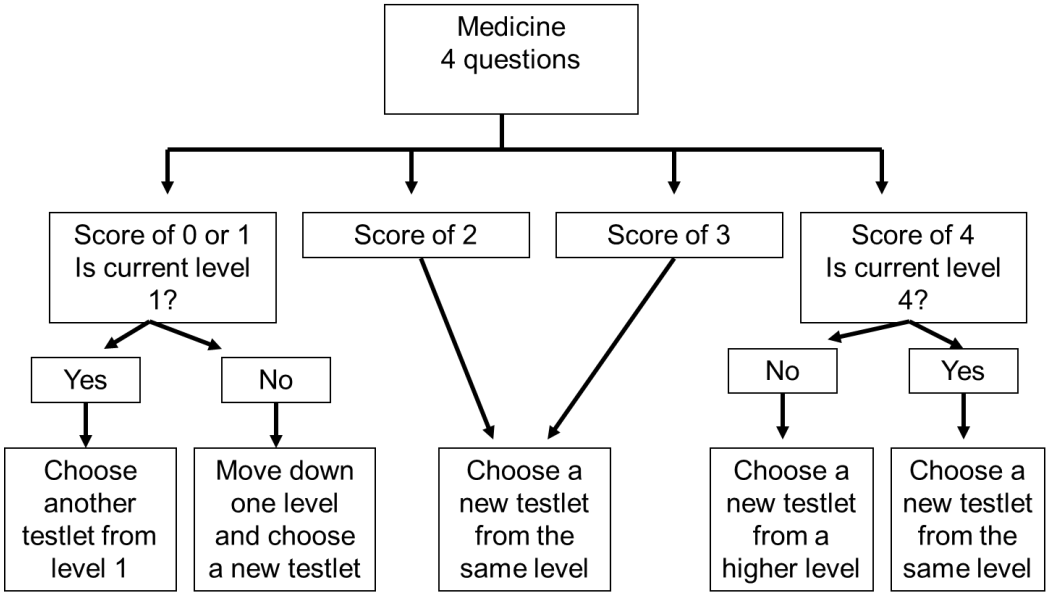


Figure 2b. An illustration of the adaptive process for the content area of Medicine at Stages 2 through 7.



DIF DETECTION AND COMPUTER-ADAPTIVE TESTING

It is particularly important to conduct DIF analyses with a computer adaptive test (CAT) because of the adaptive nature of the item selection process. Examinees write fewer items on an adaptive test, which is one important benefit of this test administration approach. However, each item contributes more to the final ability estimate because fewer item are administered. As a result, the presence of item bias could exert a stronger affect on the examinees' estimates of ability because fewer items are

administered compared to a tradition non-adaptive paper-based test. The presence of bias could also affect the order of item administration, given that the selection of items on an adaptive test is determined, in part, by the examinees' response to the previous item (see Figure 2b). Hence, bias should be minimized to ensure that construct-irrelevant variance does not adversely affect the item selection process which, in turn, could differentially affect the test score estimates for specific groups of examinees.

While it is important to ensure that items are DIF-free in CAT, it is also challenging to detect these types of problematic items on an adaptive assessment. CAT requires large numbers of items because banks are needed to permit continuous testing while, at the same time, minimizing item exposure. In this study, a sub-pool of 2806 previously administered items were analyzed. Because item exposure rates must be carefully controlled to promote test security, large item banks must first be developed to operationalize the testing process, and then continually replenished to minimize item exposure and maintain test security. As a result, the number of examinees who write any one item on an adaptive test is small, particularly when the item bank is large, relative to the items on a paper-based exam. Consequently, DIF methods designed to help monitor fairness in adaptive tests must function in diverse testing environments and, often, when the total number of items in the bank is large but the number of examinees who respond to any one of those items is relatively small.

Given that DIF detection is both an important and challenging undertaking with CAT, it comes as some surprise that little research has been conducted on this topic in the last decade. In 2000, Zwick published a seminal chapter on DIF in CAT as part of the book *Computer Adaptive Testing: Theory and Practice* (van der Linden & Glas, 2000). She reviewed the three DIF detection methods that, at the time, were considered the main CAT DIF procedures—the Zwick, Thayer, and Wingersky CAT DIF method, the CAT version of the empirical Bayes Mantel-Haenszel DIF method of Zwick, Thayer, and Lewis, and CAT for SIBTEST by Nandakumar and Roussos (CATSIB). Zwick provided a review of each

method in its original, non-adaptive version. Then, she described each method in its modified, adaptive version. Finally, she presented some empirical results from simulation studies to support each CAT DIF method.

A decade later, in 2010, *Computer Adaptive Testing: Theory and Practice* (van der Linden & Glas, 2000) was revised and updated, and published as *Elements of Adaptive Testing* (van der Linden & Glas, 2010). The new volume featured revised chapters from many of the original authors as well as some new chapters. Zwick's chapter on DIF in CAT was included in the 2010 volume. The most striking feature of Zwick's revised chapter was how *little* the area of DIF in CAT had changed over the last 10 years. In the decade since the publication of her first chapter, no new DIF methods for CAT were introduced in Zwick's review. Moreover, only 10 new references were included in her updated manuscript (out of a total of 57 references in the 2010 chapter), of which five were published prior to the publication of Zwick's first chapter in 2000, five were published after 2000, and only two of the five references published after 2000 were found in referred journals (the other three citations appeared in technical reports). In short, relatively little research has been conducted on DIF in CAT since 2000 despite the explosion of research on and application of computer-based and computer adaptive testing over the same time period.

Gierl, Lai, and Li (2011) began to address this gap in the research literature by evaluating the performance of CATSIB in a multi-stage adaptive testing environment. CATSIB was selected because it was one of Zwick's (2000, 2010) three main CAT DIF methods. CATSIB (Nakadumar & Roussos, 2001, 2004), a modification of SIBTEST (Shealy & Stout, 1993) intended for CAT, is a statistical procedure used to first match reference and focal group examinees on a regression-corrected IRT-based ability estimate, and then compare the examinees on a weighted mean difference to determine the presence of DIF. To-date, CATSIB has received limited empirical evaluation in a small number of CAT environments using an item pretesting design (see Nandakumar & Roussos, 2001, 2004; Lei, Chen, &

Yu, 2006). In this design, pretest DIF items are administered in a non-adaptive manner. As a result, adaptation affects the non-DIF items used to match examinees, but not the pretest DIF items themselves. This design can be used when items on the matching subtest are either known or assumed to be free from DIF. Unfortunately, the CATSIB statistical results from the item pretest design may not generalize to a purely adaptive context, meaning when items in both the matching and studied subtests are administered adaptively. Therefore, the purpose of the Gierl et al. study was to evaluate the performance of CATSIB for detecting DIF when both the matching and studied subtest items are administered adaptively in the context of a realistic adaptive testing environment. Their adaptive test was simulated to model a realistic multi-stage adaptive test (MST), much like the MCC QE Part I (see also Roy, Gierl, Breithaupt, & Lai, 2011). Their MST model used a four-item testlet in a seven-panel administration, where the first panel contained a single four-item testlet that contained items of different difficulty levels for all examinees. The second to seventh panels each contained three modules with four items per module at three different difficulty levels. Each examinee wrote seven modules thereby completing 28 items. Three independent variables, expected to affect DIF detection rates, were manipulated: item difficulty (easy, medium, hard modules), sample size (small—100-175 examinees per group; moderate—200-300 examinees per group; large—300-450 examinees per group), and balanced/unbalanced design (same or different sample sizes in each group). Two types of dependent variables were used to evaluate the performance of CATSIB, Type I error and power. Gierl et al. reported that CATSIB met the acceptable criteria, meaning that the Type I error and power rates met 5% and 80%, respectively, for the large reference/moderate focal sample and the large reference/large focal sample conditions. Hence, their results revealed that CATSIB could be used to consistently and accurately detect DIF on an MST, but only with moderate to large samples. In other words, CATSIB identifies DIF items with adequate Type I error protection and power across a range of module difficulty levels when a minimum sample size of 475 examinees is used.

METHODS

In the present study a combination of statistical criteria were used to identify items that function differentially across five grouping variables that characterize examinees who wrote the MCC QE Part I. Hence, the focus of this study is on identifying DIF items statistically but not interpreting the DIF items substantively (i.e., conducting sensitivity reviews, which entails having panels of content specialists review items to ensure they meet basic standards consistent with fair and equitable assessment practices). The substantive analyses should be conducted as a next step in this research program using the results from the present study to understand why DIF occurs for some items on the MCC QE Part I. The items evaluated in this study constitute the operational item bank currently used by the MCC for the QE Part I. This bank consists of a sub-pool of 2806 active items from the MCC QE Part I. For our study, the items were categorized by their four levels of difficulty (ranging from Level 1-Easy to 4-Difficult) and by their six content areas (Internal Medicine; Surgery; Obstetrics and Gynecology; Pediatrics; Psychiatry; and Population Health, including considerations of the legal, ethical, and organizational aspects of the practice of medicine). Five demographics variables that could elicit DIF across the difficulty levels and content areas were evaluated: Gender (Male/Female); Country of Degree (Canada/Foreign); Birth Country of Examinee (Canada/Foreign); Citizenship (Canadian/Non-Canadian); and Language (English/French). These five demographic variables were selected because this information is routinely collected by the Medical Council for each examinee, they can be dichotomously coded, and they serve as variables that could elicit group differences at the item level on the MCC QE Part I.

DESCRIPTIVE STATISTICAL RESULTS FOR OVERALL TEST PERFORMANCE BY DEMOGRAPHIC GROUP

The ability estimates and their associated standard errors for the students in each demographic variable using data from the Spring and Fall 2010 administrations is shown in Table 1. These results

reveal that the grouping variable used for the DIF analysis yields different mean ability estimates[1]. The largest difference occurs for the Country of Degree variable, where examinees trained in Canada received higher MCC QE Part I test scores than foreign-trained examinees. The smallest difference occurs for the Gender variable, where females outperformed males on the exam, but the overall performance difference is small. Language-related performance differences were also small, but examinees who wrote the MCC QE Part I in French outperformed their counterparts who wrote the exam in English. A complete set of descriptive statistics using overall test performance for the demographic variables using in our DIF study as a function of difficulty level and content area is provided in the Appendix.

Table 1.

Mean Ability Estimate and Associated Standard Error for Groups Used in the DIF Analysis as a Function of the Five Demographic Variables

Demographic	Group	Mean Theta	Standard Error	N	% N	Absolute Mean Difference
Gender	Male	-0.37	0.26	2679	50.0%	0.25
	Female	-0.12	0.27	2677	50.0%	
Degree	Canada	0.37	0.29	2504	46.8%	1.17
	Foreign	-0.80	0.24	2852	53.2%	
Birth	Canada	0.33	0.29	2273	42.4%	1.01
	Foreign	-0.68	0.25	3083	57.6%	
Citizenship	Canada	0.16	0.28	2845	53.1%	0.88
	Foreign	-0.71	0.25	2511	46.9%	
Language	English	-0.30	0.26	4590	85.7%	0.38
	French	0.08	0.28	766	14.3%	

OVERVIEW OF DIF DETECTION METHODS

Four different statistical criteria were used to identify DIF items. The benefit of using four criteria stems from the fact CATSIB may be differentially effective across study conditions leading to different conclusions about the presence of DIF because the sample sizes vary dramatically across these

[1] It may also be useful to note the score reported to candidates is a weighted combination of this ability score, and scores on other performance tasks. The total score reported to test takers is scaled to a reported scale ranging from 0 to 1000.

conditions (see, for example, %N column in the Appendix). Hence, we identify DIF items using four different decision-making criteria. The first criterion is a statistical significance test of the CATSIB effect size measure $\hat{\beta}_{UNI}$. CATSIB is used to test the statistical hypotheses H_o: $\hat{\beta}_{UNI}$ = 0 versus H_1: $\hat{\beta}_{UNI} \neq$ 0, where $\hat{\beta}_{UNI}$ is the parameter specifying the amount of DIF that occurs for an item when examinees in the reference and focal group are compared. $\hat{\beta}_{UNI}$ serves as a measure of the expected probabilistic difference of a correct response between examinees in the reference and focal group. That is, $\hat{\beta}_{UNI}$ = $ES_R(\theta^*) - ES_F(\theta^*)$, where $ES_R\ (\theta^*)$ and $ES_F\ (\theta^*)$ are the regression corrected expected scores in the reference and focal groups, respectively, conditional on the matching subtest. A regression correction procedure is used to produce adjusted scores that more reliably reflect examinees of equal ability levels across groups and, thus, are more meaningful for comparing group differences on the studied items. CATSIB then uses the weighted average difference of these adjusted scores (weighted by the proportion of examinees obtaining matching subtest score θ^*) to estimate the DIF index $\hat{\beta}_{UNI.}$ A test statistic for evaluating the null hypothesis is then used. Shealy and Stout (1993) demonstrated that SIB has a normal distribution with mean 0 and variance 1 under the null hypothesis. For the current study, a two-tailed, non-directional hypothesis is used to evaluate the presence of DIF at an alpha level of 0.05. An item that produces a statistical result that exceeds the null hypothesis level of 1.96 (i.e., an item that produces a statistical result that is below the alpha level of 0.05) is identified as a DIF item. This DIF detection criterion is considered most inclusive, meaning that the largest number of DIF items are expected to be identified because the smallest number of requirements are used to identify DIF items. This condition is also expected to have the highest Type I error rate (i.e., highest number of items falsely identified as DIF items) relative to the other study conditions.

The second criterion is a statistical test combined with the sample size recommendation provided by Gierl et al. of 475 examinees. An item that produces a statistically significant $\hat{\beta}_{UNI}$ at or below an alpha level of 0.05 and includes a total of 475 examinees or more in the analysis is identified as a DIF

item. This criterion is more constrained than that of the first condition because it includes both a statistical test and a sample size requirement. Hence, only items that were administered to at least 475 were analyzed. This condition is expected to have an inflated Type I error rate, but the rate should be an improvement compared to the prior condition because the sample size requirement should increase the power of the statistical significance test.

The third criterion focuses on the magnitude of the $\hat{\beta}_{UNI}$ effect size measure. $\hat{\beta}_{UNI}$ is the weighted expected score difference between the reference and focal group. The calculation of this difference score is affected by the sample size of the reference and focal groups, as larger samples yield more stable effect size measures. $\hat{\beta}_{UNI}$ also has the desirable characteristic of interpretability, meaning the importance of $\hat{\beta}_{UNI}$ can be evaluated. Nandakumar (1993) and Roussos (personal communication, October 28, 1999) claimed, for example, that if $\hat{\beta}_{UNI}$ is less than 0.05, then the expected score difference is less than 1/20 of a score point, which they considered to be small. Nandakumar (1993) and Roussos (personal communication, October 28, 1999) also claimed that if $\hat{\beta}_{UNI}$ is greater than that 0.10, then the expected score difference is greater than 1/10 of a score point, which they considered to be large. Thus, for the third criterion, an item that yields a $\hat{\beta}_{UNI}$ at or greater than 0.10 and includes a total of 475 examinees or more, as recommended by Gierl et al., is identified as a DIF item. This condition uses the same criteria used by Gierl et al. in their simulation study to identify DIF items. As a result, DIF detection in this condition is expected to be powerful (i.e., 80% accuracy, or higher) while maintaining adequate Type I error control (i.e., close to 5%).

[Alternatively, we can also conceptualize stat sig is the baseline, Sample size being the guard for power, and Magnitude being the guard for Effect size, therefore, all three criteria should yield results of statistically significance, with high power due to its strict requirements in sample and magnitude]

The fourth criterion focuses on the magnitude of the combined outcome of a statistical test, a $\hat{\beta}_{UNI}$ effect size measure, and a sample size requirement. The condition is most exclusive, meaning that

it has the largest number of requirements (i.e., statistical test, effect size measure, sample size requirement) which will lead to the fewest number of identified DIF items. But this conservative condition is also expected to lack power, relative to criterion 3, because Type I errors will be minimized but some DIF items could also be missed due to the stringent Type I error control. A conjunctive rule (i.e., a combination of a statistical test and a DIF effect size measure) is often used to identify items that function differentially in operational testing programs[2]. For example, research at the Educational Testing Service has resulted in proposed values for interpreting the Mantel-Haenszel (MH) DIF detection approach. MH DIF values are classified as negligible, moderate, or large using the following conjunctive rule (Zieky, 1993, p. 342): Negligible or A-level DIF: Null hypothesis is retained or null hypothesis is rejected and $|\Delta_{MH}| < 1$; Moderate or B-level DIF: : Null hypothesis is rejected and $1 \leq |\Delta_{MH}| < 1.5$; Large or C-level DIF: Null hypothesis is rejected and $|\Delta_{MH}| \geq 1.5$, where Δ_{MH} is the effect size measure used with the MH DIF detection approach. Similarly, SIBTEST yields an overall statistical test and an effect size measure for each item. Roussos and Stout (1996, p. 220) proposed the following $\hat{\beta}_{UNI}$ values for classifying DIF as negligible, moderate, and large: Negligible or A-level DIF: Null hypothesis is rejected and $|\hat{\beta}_{UNI}| < 0.059$; Moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\hat{\beta}_{UNI}| < 0.088$; Large or C-level DIF: Null hypothesis is rejected and $|\hat{\beta}_{UNI}| \geq 0.088$. By adding the effect size magnitude used with criterion 3 to the combined decision outcome used with criterion 4, our final DIF detection condition for the present study can be specified. For the fourth criterion, an item that yields a statistically significant $\hat{\beta}_{UNI}$ at or below an alpha level of 0.05, has a $\hat{\beta}_{UNI}$ at or greater than that 0.10, and includes a total of 450 examinees or more in the analysis is identified as a DIF item. This condition is expected to yield the smallest number of items flagged with DIF.

[2] This outcome also reveals that operational testing programs are inclined to use a conservative procedure to identify DIF items which minimizes their Type I errors (i.e., identify items as DIF when they are not) but also results in the smallest number of identified DIF items due to lower power rates.

Next, the DIF detection results are presented. Three different outcomes are provided: overall, by difficulty level, and by content area for each of the five demographic variables. DIF is also evaluated using four different criteria. Criterion 1 is statistical significance test for the difference between the reference and focal groups; criterion 2 is a statistical significance test using a minimum sample size of 475 examinees across both groups; criterion 3 is the $\hat{\beta}_{UNI}$ effect size measure at or above 0.10 and a minimum sample size of 475 examinees across both groups; criterion 4 is a statistical significance test for the difference between the reference and focal groups, the $\hat{\beta}_{UNI}$ effect size measure at or above 0.10, and a minimum sample size of 475 examinees across both groups. These four standards are referred to as criterion 1 to 4, respectively.

GENDER DIF RESULTS

OVERALL DIF BY GENDER (MALE/FEMALE)

Overall, criterion 4 identified the fewest DIF items whereas criterion 1 identified the most. The proportion of DIF items ranged from a low of 2% to a high of 9%. Criterion 3, the outcome we consider to be most reliable and accurate, identified 4% of the items as gender DIF (see Table 2).

Table 2.

The total number and proportion of DIF items across the 4 flagging criteria for Gender

Flagging Criteria	Number of Flagged Items	% of Flagged Items
1	251	9%
2	57	2%
3	103	4%
4	45	2%

GENDER BY DIFFICULTY

The proportion of DIF items by gender across the four difficulty levels used to classify and administer items on the MCC QE Part I is shown in Table 3. Proportions rather than counts are reported because the total number of items included in the calculation using a specific DIF classification criterion varies across the study conditions. Hence, the proportions most clearly convey

information about the amount of DIF that is occurring within the MCC item bank across the demographic variables. For difficulty level 1 (easiest items), the number of DIF items was 0%; for difficulty level 2, the number of DIF items ranged from 0% to 2%; for difficulty level 3, the number of DIF items ranged from 1% to 8%; for difficulty level 4 (hardest items), the number of DIF items ranged from 1% to 35%. If we consider criterion 3 to be the most reliable and accurate DIF detection outcome, then gender DIF by difficulty level ranges from 0% to 1%.

Table 3.

The Proportion of DIF Items As a Function of Gender Across the Four Difficulty Levels

Difficulty Level	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Easy	0%	0%	0%	0%
2	2%	1%	0%	0%
3	8%	3%	1%	1%
4—Difficult	35%	9%	1%	1%

GENDER BY CONTENT AREA

The proportion of DIF items by gender across the six content areas used to classify and administer items on the MCC QE Part I is shown in Table 4. For Internal Medicine, the number of DIF items ranged from 2% to 13%; for Surgery, the number of DIF items ranged from 0% to 10%; for Obstetrics and Gynecology, the number of DIF items ranged from 0% to 10%; for Pediatrics, the number of DIF items ranged from 0% to 10%; for Psychiatry, the number of DIF items ranged from0% to 9%; for Population Health, the number of DIF items ranged from 2% to 8%. Using criterion 3 as the most interpretable DIF detection outcome suggests that gender DIF by content area ranges from 0% to 3%.

Table 4.

The Proportion of DIF Items As a Function of Gender Across the Six Content Areas

Content Area	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Internal Medicine	13%	7%	3%	2%
2—Surgery	10%	2%	0%	0%
3—Obstetrics and Gynecology	10%	2%	0%	0%
4—Pediatrics	10%	3%	0%	0%
5—Psychiatry	9%	3%	0%	0%
6—Population Health	8%	5%	2%	2%

COUNTRY OF DEGREE DIF RESULTS

OVERALL DIF BY COUNTRY OF DEGREE (CANADA/FOREIGN)

Overall, criterion 4 identified the fewest DIF items whereas criterion 1 identified the most (see Table 5). The proportion of DIF items ranged a low of 1% for degree-related DIF to a high of 5%. The outcome for criterion 3 was 5%.

Table 5.

The total number and proportion of DIF items across the 4 flagging criteria for Country of Degree

Flagging Criteria	Number of Flagged Items	% of Flagged Items
1	131	5%
2	31	1%
3	131	5%
4	27	1%

COUNTRY OF DEGREE BY DIFFICULTY

The proportion of DIF items by country of degree across the four difficulty levels used to classify and administer items on the MCC QE Part I is shown in Table 6. For difficulty level 1 (easiest items), the number of DIF items ranged from 0% to 2%; for difficulty level 2, the number of DIF items ranged from 0% to 3%; for difficulty level 3, the number of DIF items ranged from 1% to 6%; for difficulty level 4 (hardest items), the number of DIF items ranged from 4% to 21%. Using criterion 3 as our point-of-reference, degree-related DIF by difficulty level ranges from 2% to 10%.

Table 6.

The Proportion of DIF Items As a Function of Country of Degree Across the Four Difficulty Levels

Difficulty Level	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Easy	0%	0%	2%	0%
2	0%	0%	3%	0%
3	2%	1%	6%	1%
4—Difficult	21%	5%	10%	4%

COUNTRY OF DEGREE BY CONTENT AREA

The proportion of DIF items by country of degree across the six content areas is shown in Table 7. For Internal Medicine, the number of DIF items ranged from 2% to 11%; for Surgery, the number of DIF items ranged from 0% to 5%; for Obstetrics and Gynecology, the number of DIF items ranged from 0% to 4%; for Pediatrics, the number of DIF items ranged from 1% to 4%; for Psychiatry, the number of DIF items ranged from 1% to 6%; for Population Health, the number of DIF items ranged from 2% to 8%. Using criterion 3 as a reliable and accurate DIF detection point-of-reference suggests that degree-related DIF by content area ranges from 2% to 11%.

Table 7.

The Proportion of DIF Items As a Function of Country of Degree Across the Six Content Areas

Content Area	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Internal Medicine	6%	2%	11%	2%
2—Surgery	5%	0%	2%	0%
3—Obstetrics and Gynecology	4%	0%	1%	0%
4—Pediatrics	4%	1%	3%	1%
5—Psychiatry	5%	1%	4%	1%
6—Population Health	4%	3%	8%	2%

BIRTH COUNTRY DIF RESULTS

OVERALL DIF BY BIRTH COUNTRY (CANADA/FOREIGN)

Overall, criterion 4 identified the fewest DIF items whereas criterion 1 identified the most (see Table 8). The proportion of DIF items ranged a low of 1% for birth country-related DIF to a high of 5%. The outcome for criterion 3 was 3%.

Table 8.

The total number and proportion of DIF items across the 4 flagging criteria for Birth Country

Flagging Criteria	Number of Flagged Items	% of Flagged Items
1	133	5%
2	22	1%
3	81	3%
4	18	1%

BIRTH COUNTRY BY DIFFICULTY

The proportion of DIF items by birth country across the four difficulty levels is shown in Table 9. For difficulty level 1 (easiest items), the number of DIF items ranged from 0% to 1%; for difficulty level 2, the number of DIF items ranged from 0% to 2%; for difficulty level 3, the number of DIF items ranged from 0% to 4%; for difficulty level 4 (hardest items), the number of DIF items ranged from 3% to 20%. Using criterion 3 as our point-of-reference, birth country-related DIF by difficulty level ranges from 1% to 5%.

Table 9.

The Proportion of DIF Items As a Function of Birth Country Across the Four Difficulty Levels

Difficulty Level	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Easy	0%	0%	1%	0%
2	0%	0%	2%	0%
3	3%	0%	4%	0%
4—Difficult	20%	4%	5%	3%

BIRTH COUNTRY BY CONTENT AREA

The proportion of DIF items by birth country across the six content areas is shown in Table 10. For Internal Medicine, the number of DIF items ranged from 1% to 7%; for Surgery, the number of DIF items ranged from 1% to 6%; for Obstetrics and Gynecology, the number of DIF items ranged from 0% to 5%; for Pediatrics, the number of DIF items ranged from 0% to 4%; for Psychiatry, the number of DIF items ranged from 1% to 5%; for Population Health, the number of DIF items ranged from 1% to 5%.

Using criterion 3 as a reliable and accurate DIF detection point-of-reference suggests that birth country-related DIF by content area ranges from 1% to 7%.

Table 10.

The Proportion of DIF Items As a Function of Birth Country Across the Six Content Areas

Content Area	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Internal Medicine	6%	1%	7%	1%
2—Surgery	6%	1%	1%	1%
3—Obstetrics and Gynecology	5%	0%	1%	0%
4—Pediatrics	4%	0%	2%	0%
5—Psychiatry	5%	1%	3%	1%
6—Population Health	2%	2%	5%	1%

CITIZENSHIP DIF RESULTS

OVERALL DIF BY CITIZENSHIP (CANADIAN/NON-CANADIAN)

Overall, criterion 4 identified the fewest DIF items whereas criterion 1 identified the most (see Table 11). The proportion of DIF items ranged a low of 1% for citizenship-related DIF to a high of 6%. The outcome for criterion 3 was 3%.

Table 11.

The total number and proportion of DIF items across the 4 flagging criteria for Citizenship

Flagging Criteria	Number of Flagged Items	% of Flagged Items
1	177	6%
2	37	1%
3	73	3%
4	20	1%

CITIZENSHIP BY DIFFICULTY

The proportion of DIF items by citizenship across the four difficulty levels is shown in Table 12. For difficulty level 1 (easiest items), the number of citizenship-related DIF items ranged from 0% to 1%; for difficulty level 2, the number of DIF items ranged from 0% to 1%; for difficulty level 3, the number of DIF items ranged from 1% to 5%; for difficulty level 4 (hardest items), the number of DIF items ranged

from 2% to 26%. Using criterion 3 as our point-of-reference, citizenship-related DIF by difficulty level ranges from 1% to 5%.

Table 12.

The Proportion of DIF Items As a Function of Citizenship Across the Four Difficulty Levels

Difficulty Level	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Easy	0%	0%	1%	0%
2	0%	0%	1%	0%
3	4%	1%	5%	1%
4—Difficult	25%	5%	4%	2%

CITIZENSHIP BY CONTENT AREA

The proportion of DIF items by citizenship across the six content areas is shown in Table 13. For Internal Medicine, the number of DIF items ranged from 2% to 7%; for Surgery, the number of DIF items ranged from 0% to 6%; for Obstetrics and Gynecology, the number of DIF items ranged from 0% to 8%; for Pediatrics, the number of DIF items ranged from 1% to 7%; for Psychiatry, the number of DIF items ranged from 1% to 6%; for Population Health, the number of DIF items ranged from 1% to 4%. Using criterion 3 as an interpretable point-of-reference suggests that citizenship-related DIF by content area ranges from 1% to 6%.

Table 13.

The Proportion of DIF Items As a Function of Citizenship Across the Six Content Areas

Content Area	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Internal Medicine	7%	3%	6%	2%
2—Surgery	6%	1%	1%	0%
3—Obstetrics and Gynecology	8%	1%	1%	0%
4—Pediatrics	7%	1%	1%	1%
5—Psychiatry	6%	1%	3%	1%
6—Population Health	4%	2%	4%	1%

LANGUAGE DIF RESULTS

OVERALL DIF BY LANGUAGE (ENGLISH/FRENCH)

Overall, criterion 4 identified the fewest DIF items whereas criterion 1 identified the most (see Table 14). The proportion of DIF items ranged a low of 2% for language DIF to a high of 9%. The outcome for criterion 3 was 4%.

Table 14.

The total number and proportion of DIF items across the 4 flagging criteria for Language.

Flagging Criteria	Number of Flagged Items	% of Flagged Items
1	251	9%
2	57	2%
3	103	4%
4	45	2%

LANGUAGE BY DIFFICULTY

The proportion of DIF items by language of examination across the four difficulty levels is shown in Table 15. For difficulty level 1 (easiest items), the number of DIF items ranged from 0% to 2%; for difficulty level 2, the number of DIF items ranged from 0% to 2%; for difficulty level 3, the number of DIF items ranged from 2% to 8%; for difficulty level 4 (hardest items), the number of DIF items ranged from 4% to 32%. Using criterion 3 as our point-of-reference, language-related DIF by difficulty level ranges from 2% to 6%.

Table 15.

The Proportion of DIF Items As a Function of Language Across the Four Difficulty Levels

Difficulty Level	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Easy	0%	0%	2%	0%
2	1%	0%	2%	0%
3	8%	2%	6%	2%
4—Difficult	32%	7%	6%	4%

LANGUAGE BY CONTENT AREA

The proportion of DIF items by language across the six content areas is shown in Table 15. For Internal Medicine, the number of DIF items ranged from 6% to 17%; for Surgery, the number of DIF items ranged from 0% to 7%; for Obstetrics and Gynecology, the number of DIF items ranged from 0% to 10%; for Pediatrics, the number of DIF items ranged from 0% to 8%; for Psychiatry, the number of DIF items ranged from 1% to 7%; for Population Health, the number of DIF items ranged from 3% to 7%. Using criterion 3 as a reliable and accurate DIF detection point-of-reference suggests that language-related DIF by content area ranges from 1% to 14%.

Table 16.

The Proportion of DIF Items As a Function of Language Across the Six Content Areas

Content Area	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1—Internal Medicine	17%	6%	14%	6%
2—Surgery	7%	1%	1%	0%
3—Obstetrics and Gynecology	10%	1%	1%	0%
4—Pediatrics	8%	1%	1%	0%
5—Psychiatry	7%	1%	2%	1%
6—Population Health	7%	4%	6%	3%

SUMMARY AND CONCLUSIONS

The purpose of the present study was to use a combination of statistical criteria to identify items that function differentially across five demographic grouping variables that characterize examinees who write the Medical Council of Canada Qualifying Examination Part I (MCC QE Part I). The MCC QE Part I is a licensure test written by all medical students seeking entry into supervised clinical practice in postgraduate training programs. It is a one-day, fixed-length, multi-stage, computer adaptive test used to assess the knowledge, clinical skills, and attitudes specified by the Medical Council as key objectives and competencies for medical training in Canada. Our study focused on DIF detection with the 2806 multiple-choice items previously administered in Section 1 of the computer-adaptive component of the test. A statistical DIF analysis involves administering the test, matching members of

the reference and focal group on a measure of ability derived from that test, and using statistical procedures to identify group differences on test items. An item exhibits DIF when examinees from the reference and focal groups differ in the probability of answering that item correctly, after controlling for the measure of ability derived from the test. The focus of the study was to identify DIF items using statistical analyses. There was no attempt to interpret the items substantively.

CATSIB (Nakadumar & Roussos, 2001, 2004), a modification of SIBTEST (Shealy & Stout, 1993) intended for CAT, was used to identify DIF items in the current study. CATSIB is a statistical procedure used to match reference and focal group examinees on regression-corrected IRT-based ability estimate, and then compare the examinees on a weighted mean difference to determine the presence of DIF. CATSIB was selected because it was one of Zwick's (2000, 2010) three main CAT DIF methods. CATSIB was also evaluated recently by Gierl et al. (2011) to assess its DIF detection performance when both the matching and studied subtest items are administered adaptively in the context of a realistic multi-stage adaptive test (MST), which is the same adaptive testing model used by the Medical Council of Canada for the MCC QE Part I. Gierl et al. reported that CATSIB met the acceptable criteria, meaning that the Type I error and power rates met 5% and 80%, respectively across a range of module difficulty levels when a minimum sample size of 475 examinees is used.

In the current study, four different criteria were used to identify DIF items. The benefit of using four criteria stems from the fact CATSIB may be differentially effective across study conditions leading to different conclusions about the presence of DIF because the sample sizes vary dramatically across the demographic variables (see %N column in the Appendix). Hence, we identify DIF items using four different decision-making criteria thereby producing a range of DIF detection outcomes. Criterion 1 is statistical significance test for the difference between the reference and focal groups; criterion 2 is a statistical significance test using a minimum sample size of 475 examinees across both groups; criterion 3 is the $\hat{\beta}_{UNI}$ effect size measure at or above 0.10 and a minimum sample size of 475

examinees across both groups; criterion 4 is a statistical significance test for the difference between the reference and focal groups, the $\hat{\beta}_{UNI}$ effect size measure at or above 0.10, and a minimum sample size of 475 examinees across both groups. We provide preferential treatment to the interpreted outcomes for criterion 3, as this was the same criterion used Gierl et al. to evaluate CATSIB in a multi-stage testing environment similar to that used with the MCC QE Part I. Hence, criterion 3 was expected to yield the most reliable and accurate DIF detection results, in light of the research currently available on CATSIB.

Overall, there was very little DIF in the item bank used for the MCC QE Part I. The proportion of gender DIF items ranged a low of 2% to a high of 9%. Criterion 3 identified 4% of the items as gender DIF. The proportion of DIF items for the Country of Degree variable ranged from a low of 1% to a high of 5%, with criterion 3 flagging 5% of the items. For Birth Country, the proportion of DIF items ranged a low of 1% to a high of 5%, with criterion 3 flagging 3% of the items. The proportion of DIF items for Citizenship ranged from a low of 1% to a high of 6%, with criterion 3 flagging 3% of the items. Finally, for language, the proportion of DIF items ranged a low of 2% to a high of 9%, with criterion 3 identifying 4% of the items.

When we evaluate DIF by difficulty level, again, very few items were identified. However, one trend does exist. The number of DIF items increases as difficulty level increased. In other words, there are more DIF items as the difficulty levels moves from easy (i.e., Level 1) to hard (i.e., Level 4). If we use the results from criterion 3, as an example, across all five demographic variable conditions then the average proportion of DIF items from Levels 1 to 4 is 1%, 2%, 4%, and 5%, respectively. This result indicates that the MCC QE Part I bank contains item that elicit more DIF at the higher difficulty level, and hence, higher ability levels.

One reason why the level 4 items may elicit more DIF is that these items are selected for review by expert panels less often than items in the other sections. A summary of the average difficulty (Table

17a) and discrimination values (Table 17b) for all items in the MCC bank, as a function of difficulty level and content area, suggests this interpretation may be accurate. The results from these tables clearly reveal two outcomes. Outcome one: The average item difficulty value increases from level 1 to 4 for all six content areas, as one would expect, however the standard deviation for level 4 items is noticeably larger than the standard deviation for the items in the other three levels (see Table 17a). This results indicates that the items in level 4 are more heterogeneous and variable than the items in the other levels. Outcome two: The average item discrimination value as well as the standard deviation for the discrimination value is the lowest for level 4 items across all six content areas. That is, the items in level 4 consistently yield the lowest discrimination power. Taken together, these outcomes indicate that the items in level 4 have more diverse difficulty values but lower overall discrimination values compared to the other items on the test. This diversity in item difficulty combined with lower overall discrimination power may lead to more spurious response patterns among examinees or subgroups of examinees resulting in higher DIF detection rates.

Table 17a.

The Average IRT Difficulty (b-parameter) Value As a Function of Difficulty Level and Content Area

Average Difficulty Parameters		Content Area					
Difficulty Level		1	2	3	4	5	6
1	Mean	-3.61	-3.50	-3.61	-3.77	-3.61	-3.49
	SD	1.00	0.80	0.81	0.83	0.79	0.81
2	Mean	-1.78	-1.91	-2.01	-2.05	-2.03	-1.94
	SD	0.38	0.35	0.37	0.37	0.37	0.36
3	Mean	-0.16	-0.55	-0.50	-0.60	-0.70	-0.58
	SD	0.58	0.45	0.52	0.53	0.42	0.43
4	Mean	2.40	1.65	1.91	1.63	1.58	1.43
	SD	1.14	1.10	1.17	1.04	1.26	1.03

Table 17b.

The Average IRT Discrimination (a-parameter) Value As a Function of Difficulty Level and Content Area

Average Discrimination Value		Content Area					
Difficulty Level		1	2	3	4	5	6
1	Mean	0.36	0.37	0.37	0.37	0.35	0.42
	SD	0.13	0.12	0.19	0.14	0.13	0.16
2	Mean	0.45	0.40	0.36	0.38	0.39	0.46
	SD	0.21	0.16	0.16	0.18	0.16	0.18
3	Mean	0.39	0.35	0.29	0.34	0.33	0.40
	SD	0.16	0.14	0.10	0.15	0.13	0.17
4	Mean	0.26	0.28	0.22	0.26	0.26	0.26
	SD	0.10	0.12	0.07	0.10	0.09	0.10

When we evaluate DIF by content area, again, few items were identified. However, the number of DIF items does vary by content area. Using the results from criterion 3 aggregated across all five demographic variables, Internal Medicine produced the largest number of DIF items at 8%; Population Health items elicit the second largest number of DIF items at 5%; Surgery, Obstetrics and Gynecology, Pediatrics, and Psychiatry all have relatively low levels of DIF at 1%, 1%, 2%, and 3%, respectively. Hence, future studies designed to investigate the potential causes of DIF might begin by focusing on examinee performance in the content areas of Internal Medicine and Population Health, as the items in these areas seem to produce the most group differences. However, the more important finding from this study, in our point-of-view, is that very few items in the MCC QE Part I bank elicit group differences by Gender (Male/Female); Country of Degree (Canada/Foreign); Birth Country of Examinee (Canada/Foreign); Citizenship (Canadian/Non-Canadian);or Language (English/French). In other words, there are very few item-level group differences on this important Canadian licensure test.

REFERENCES

Gierl, M. J., Lai, H., & Li, J. (2011). *Evaluating the performance of CATSIB in a multi-stage adaptive testing environment.* Manuscript submitted for publication.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43,* 245-264.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16,* 159-176.

Nandakumar, R., & Roussos, L. (2001, July). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests.* Law School Admission Council Computerized Testing Report 97-11.

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177-199.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Roy, M., Gierl, M. J., Breithaupt, K., & Lai, H. (2011, August). *Analytic methods to evaluate item and test fairness: A case study of the Medical Council of Canada Qualifying Examination Part I (MCCQEI).* Paper presented at the annual meeting of the Association for Medical Education in Europe, Vienna, Austria.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer.

van der Linden, W., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 221-244). Dordrecht, The Netherlands: Kluwer.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 331-352). New York: Springer.

APPENDIX

A complete set of ability estimate descriptive statistics for the demographic variables using in our DIF study, as a function of item difficulty level (ranging from 1-Easy to 4-Difficult) and content area (Internal Medicine; Surgery; Obstetrics and Gynecology; Pediatrics; Psychiatry; and Population Health, including considerations of the legal, ethical, and organizational aspects of the practice of medicine).

DIFFICULTY LEVEL

Difficulty Level = 1 (Easiest Items)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.56	0.25	197	3.7%
	Female	-0.36	0.26	194	3.6%
Degree	Canada	0.20	0.28	175	3.3%
	Foreign	-0.87	0.24	215	4.0%
Birth	Canada	0.13	0.28	160	3.0%
	Foreign	-0.79	0.25	231	4.3%
Citizenship	Canada	-0.09	0.27	202	3.8%
	Foreign	-0.82	0.24	188	3.5%
Language	English	-0.51	0.26	336	6.3%
	French	-0.17	0.27	55	1.0%

Difficulty Level = 2

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.48	0.26	198	3.7%
	Female	-0.27	0.26	198	3.7%
Degree	Canada	0.24	0.28	185	3.5%
	Foreign	-0.82	0.24	211	3.9%
Birth	Canada	0.18	0.28	168	3.1%
	Foreign	-0.72	0.25	228	4.3%
Citizenship	Canada	-0.01	0.27	210	3.9%
	Foreign	-0.75	0.25	186	3.5%
Language	English	-0.42	0.26	340	6.3%
	French	-0.09	0.27	56	1.1%

Difficulty Level = 3

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.45	0.26	189	3.5%
	Female	-0.24	0.27	190	3.6%
Degree	Canada	0.26	0.28	182	3.4%
	Foreign	-0.80	0.24	198	3.7%
Birth	Canada	0.20	0.28	164	3.1%
	Foreign	-0.69	0.25	215	4.0%
Citizenship	Canada	0.02	0.27	204	3.8%
	Foreign	-0.73	0.25	175	3.3%
Language	English	-0.39	0.26	324	6.1%
	French	-0.07	0.27	55	1.0%

Difficulty Level = 4 (Most Difficulty Items)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.45	0.26	216	4.0%
	Female	-0.25	0.26	216	4.0%
Degree	Canada	0.25	0.28	204	3.8%
	Foreign	-0.81	0.24	228	4.3%
Birth	Canada	0.20	0.28	185	3.5%
	Foreign	-0.70	0.25	247	4.6%
Citizenship	Canada	0.01	0.27	231	4.3%
	Foreign	-0.74	0.25	201	3.8%
Language	English	-0.40	0.26	370	6.9%
	French	-0.07	0.27	62	1.2%

CONTENT AREA

Content Area = 1 (Internal Medicine)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.44	0.26	240	4.5%
	Female	-0.23	0.27	241	4.5%
Degree	Canada	0.28	0.28	225	4.2%
	Foreign	-0.77	0.25	256	4.8%
Birth	Canada	0.21	0.28	204	3.8%
	Foreign	-0.66	0.25	277	5.2%
Citizenship	Canada	0.01	0.27	256	4.8%
	Foreign	-0.69	0.25	225	4.2%
Language	English	-0.38	0.26	412	7.7%
	French	-0.04	0.27	69	1.3%

Content Area = 2 (Obstetrics and Gynecology)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.50	0.26	169	3.2%
	Female	-0.27	0.26	169	3.2%
Degree	Canada	0.23	0.28	158	3.0%
	Foreign	-0.83	0.24	180	3.4%
Birth	Canada	0.17	0.28	144	2.7%
	Foreign	-0.73	0.25	195	3.6%
Citizenship	Canada	-0.03	0.27	180	3.4%
	Foreign	-0.76	0.25	158	3.0%
Language	English	-0.43	0.26	290	5.4%
	French	-0.12	0.27	48	0.9%

Content Area = 3 (Pediatrics)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.46	0.26	185	3.5%
	Female	-0.24	0.27	185	3.5%
Degree	Canada	0.27	0.28	173	3.2%
	Foreign	-0.83	0.24	197	3.7%
Birth	Canada	0.22	0.28	157	2.9%
	Foreign	-0.72	0.25	213	4.0%
Citizenship	Canada	0.04	0.28	197	3.7%
	Foreign	-0.76	0.25	174	3.2%
Language	English	-0.40	0.26	317	5.9%
	French	-0.04	0.27	53	1.0%

Content Area = 4 (Psychiatry)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.47	0.26	199	3.7%
	Female	-0.28	0.26	199	3.7%
Degree	Canada	0.25	0.28	187	3.5%
	Foreign	-0.83	0.24	212	4.0%
Birth	Canada	0.19	0.28	169	3.2%
	Foreign	-0.73	0.25	229	4.3%
Citizenship	Canada	0.00	0.27	212	4.0%
	Foreign	-0.77	0.25	187	3.5%
Language	English	-0.42	0.26	342	6.4%
	French	-0.08	0.27	57	1.1%

Content Area = 5 (Surgery)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.50	0.26	202	3.8%
	Female	-0.32	0.26	202	3.8%
Degree	Canada	0.20	0.28	189	3.5%
	Foreign	-0.85	0.24	216	4.0%
Birth	Canada	0.13	0.28	172	3.2%
	Foreign	-0.75	0.25	233	4.3%
Citizenship	Canada	-0.05	0.27	215	4.0%
	Foreign	-0.78	0.25	190	3.5%
Language	English	-0.45	0.26	347	6.5%
	French	-0.14	0.27	58	1.1%

Content Area = 6 (Population Health)

Criterion	Category	Mean Theta	Mean SE Theta	N	% N
Gender	Male	-0.50	0.26	202	3.8%
	Female	-0.32	0.26	202	3.8%
Degree	Canada	0.23	0.28	190	3.5%
	Foreign	-0.82	0.24	215	4.0%
Birth	Canada	0.16	0.28	172	3.2%
	Foreign	-0.72	0.25	232	4.3%
Citizenship	Canada	-0.07	0.27	215	4.0%
	Foreign	-0.75	0.25	189	3.5%
Language	English	-0.45	0.26	347	6.5%
	French	-0.14	0.27	58	1.1%