Pugh *et al. Research and Practice in Technology Enhanced Learning* (2020) 15:12
https://doi.org/10.1186/s41039-020-00134-8

Research and Practice in
Technology Enhanced Learning

RESEARCH Open Access

# Can automated item generation be used to develop high quality MCQs that assess application of knowledge?

Debra Pugh[1,2,3*], André De Champlain[1,2], Mark Gierl[4], Hollis Lai[5] and Claire Touchie[1,2,3]

* Correspondence: dpugh@mcc.ca
[1] Medical Council of Canada, 1021 Thomas Spratt Place, Ottawa, Ontario, Canada
[2] Department of Medicine, The Ottawa Hospital, Ottawa, Canada
Full list of author information is available at the end of the article

## Abstract

The purpose of this study was to compare the quality of multiple choice questions (MCQs) developed using automated item generation (AIG) versus traditional methods, as judged by a panel of experts. The quality of MCQs developed using two methods (i.e., AIG or traditional) was evaluated by a panel of content experts in a blinded study. Participants rated a total of 102 MCQs using six quality metrics and made a judgment regarding whether or not each item tested recall or application of knowledge. A Wilcoxon two-sample test evaluated differences in each of the six quality metrics rating scales as well as an overall cognitive domain judgment. No significant differences were found in terms of item quality or cognitive domain assessed when comparing the two item development methods. The vast majority of items (> 90%) developed using both methods were deemed to be assessing higher-order skills. When compared to traditionally developed items, MCQs developed using AIG demonstrated comparable quality. Both modalities can produce items that assess higher-order cognitive skills.

**Keywords:** Automated item generation, Multiple-choice questions

## Introduction

In recent years, automated item generation (AIG) has been increasingly used to create multiple-choice questions (MCQs) for the assessment of health professionals (Gierl, Lai, & Turner, 2012; Lai, Gierl, Byrne, Spielman, & Waldschmidt, 2016). This move is in part due to changes to the assessment landscape which have led educators to seek ways to provide more frequent testing opportunities. For example, the introduction of competency-based education models, which require multiple data points to make meaningful decisions about competence, has increased the need for test items to support more frequent and tailored assessments (Lockyer et al., 2017). Similarly, progress testing, which is gaining in popularity, requires a large number of test items to allow for the creation of multiple test forms (Albanese & Case, 2016). Finally, the creation of new content is also needed to attenuate the impact of surreptitious sharing of test items between learners through social networks (Monteiro, Silva-Pereira, & Severo, 2018).

The aforementioned changes have serendipitously led to several advancements in item development, including MCQs (Pugh, De Champlain, & Touchie, 2019), of which one of the most promising has been AIG. In brief, AIG relies on the use of computer algorithms to generate a large number of MCQs by inputting and coding information derived from a cognitive model (Gierl et al., 2012). The cognitive model approach requires content experts to deconstruct and document their thought processes before developing the test item (Pugh, De Champlain, Gierl, Lai, & Touchie, 2016). While doing this, content experts are forced to articulate the factors that would lead them down a series of different paths to solve a clinical problem. For example, if a clinician is asked to articulate their approach to a patient presenting with hyponatremia, they will identify the factors that will allow them to diagnose and manage the patient. These factors may include historical features (e.g., recent fluid intake/losses or medication use), physical examination findings (e.g., volume status), and laboratory results (e.g., urinary sodium). Different diagnoses would be associated with a different set of presenting features (i.e., variables). In other words, the diagnosis and management will be very different in a patient who is taking a selective serotonin reuptake inhibitor, is clinically euvolemic, and has a high urinary sodium (i.e., syndrome of inappropriate antidiuretic hormone secretion) versus a patient who has a history of vomiting, is clinically hypovolemic, and has a very low urine sodium (i.e., dehydration). The resulting model accounts for these differences and can be translated into code to generate MCQs through linear optimization (Gierl & Lai, 2013).

One of the most apparent advantages of using AIG is that it allows for the production of a large number of test items to be developed in a relatively short period of time. In fact, one cognitive model, developed and coded over a 2–3-h period, can lead to the generation of dozens or even hundreds of MCQs (Gierl et al., 2012). This may be very appealing to educators and organizations who find that their need for content exceeds their ability to develop items using traditional methods, such as those introducing progress testing or competency-based models. In addition, because AIG produces items that look similar, but require different thought processes to arrive at different answers, the impact of sharing recalled items between test-takers may be attenuated.

Another potential advantage of AIG is that it may be more likely to result in items that assess clinical reasoning or application of knowledge rather than factual recall, because of its reliance on cognitive models. Cognitive models, by design, force item writers to focus on problem conceptualization. This is important as educators strive to better understand and assess examinees' cognitive processes. Although once thought to be useful only in the assessment of lower-order skills (i.e., recall of facts), well-constructed MCQs have been shown to be beneficial in assessing clinical reasoning (Coderre, Harasym, Mandin, & Fick, 2004; Heist, Gonzalo, Durning, Torre, & Elnicki, 2014; Skakun, Maguire, & Cook, 1994). In fact, examinees have been shown to use both system I (automatic, non-analytic) and system II (analytic) cognitive processes when answering MCQs, which aligns with the processes that clinicians use in practice (Surry, Torre, & Durning, 2017). However, to date, there are no studies demonstrating that items developed using AIG do in fact target these higher-order skills.

Despite the many advantages of AIG, there is some concern that the items generated using this method may not be of the same quality level as those developed using traditional methods (in which each item undergoes rigorous committee review by a panel of content experts). Psychometrically, results from pretest items in a high-stakes exam suggest that items for health professionals developed using AIG display psychometric properties that are similar to those obtained using traditionally developed MCQs (Gierl et al., 2016). From a content expert perspective, a preliminary study was conducted and found that the quality of items generated using AIG was comparable to that of those developed using traditional methods (Gierl & Lai, 2013). In that study, researchers compared the quality of 15 MCQs developed using AIG to items developed using traditional methods, via eight pre-defined quality metrics. They found that items were comparable for seven of eight quality metrics. However, the quality of distractors (i.e., the incorrect options for MCQs) was significantly worse for items generated using AIG.

In response to the perceived concern on quality, much effort has been devoted to developing an approach to improve the quality of MCQ distractors generated using AIG. This approach has provided content experts with a framework for systematically developing a list of plausible distractors at the level of the cognitive model. In practice, this has led to the generation of high-quality distractors for MCQs, as evidenced by difficulty level and discrimination indices (Lai et al., 2016). However, although psychometrically sound, to date, there have been no follow-up studies that have examined the quality of these generated distractors from the perspective of content experts.

The Medical Council of Canada (MCC) develops and administers a written examination (MCC Qualifying Examination, Part I), that is one of the requirements for full licensure to practice medicine in Canada. Approximately, three quarters of this examination is comprised of MCQs. In the past few years, we have augmented our MCQ content development by introducing AIG (Gierl et al., 2012).

The purpose of this study was to evaluate the quality of the items generated as compared to those developed using traditional methods, as judged by a panel of experts. Specifically, this study (1) compared the constructs (i.e., knowledge versus application of knowledge) assessed by items developed using AIG versus traditional methods, and (2) compared the quality of items developed using AIG versus traditional methods. We hypothesized that the use of AIG would result in items of comparable quality to those developed using traditional methods but that AIG, because of its reliance on cognitive models, would result in items that would better assess higher-order skills.

## Research design and methods

The quality of MCQs developed using AIG and traditional methods was evaluated by content experts in a blinded review. A 5-h workshop (including training and scheduled breaks) was convened for the purposes of this quality assurance exercise. Participants were asked to rate MCQs developed by two methods (AIG vs traditional), using five indicators of quality, as well as provide a global rating of overall quality. They were then asked to make a judgment about whether or not the item tested recall or application of knowledge. This study was approved by the Ottawa Health Science Network Research Ethics Board (Protocol #20170332-01H).

### Participants

Using purposive sampling, we invited physicians with expertise in MCQ development ($n$ = 4) to participate in a workshop. These physicians all serve on test committees at

MCC and have received formal training on MCQ item development. However, none of these physicians were involved in developing the content used for this exercise. Two of the participants were family medicine physicians, one was a pediatrician, and one was a psychiatrist. There were two women and two men.

### Training

During the workshop, participants received a 1-h calibration session, led by one of the co-investigators (DP). As part of their training, participants were provided with examples of questions of high and poor quality for each of the quality metrics. Training was based on best practice for the development of MCQs (Haladyna, Downing, & Rodriguez, 2002) and reflected the instructions that are typically provided to all MCQ content developers at MCC.

They were also provided with guidance regarding what would constitute an MCQ that assesses simple knowledge (i.e., requires recall of facts to answer the question) versus application of knowledge (i.e., requires the use of knowledge to solve a clinical problem). Examples of each type were provided for training purposes, c.f. Table 1 for some examples from the workshop. For items that were deemed to assess the application of knowledge, participants were asked to make a decision about what was being assessed (i.e., diagnosis, management, communication, or professionalism). This step was added, as these are the physician activities specified in the newly developed MCC examination blueprint (Touchie & Streefkerk, 2014). Examples of each were provided as part of their training.

Finally, using frame-of-reference training (Newman et al., 2016), which helps raters develop a shared mental model, participants were calibrated through rating two practice MCQs along with a discussion as a group to ensure that all participants were rating consistently.

**Table 1** Examples of MCQs assessing knowledge and application of knowledge

| |
|---|
| MCQ assessing knowledge |
| A 46-year-old man is brought to the Emergency Department after sustaining third degree burns to 30% of his body. Which one of the following intravenous solutions should be initially initiated? |
| 1. Normal saline at 500 cc/h |
| 2. Half normal saline at 250 cc/h |
| 3. Ringer Lactate at 150 cc/h |
| 4. Pentaspan at 300 cc/h |
| 5. 2/3 and 1/31 litre bolus |
| MCQ assessing application of knowledge |
| A 62-year-old woman with a history of confusion, back pain, and constipation presents for a follow-up visit. Laboratory investigations reveal a serum calcium of 2.9 mmol/L, a creatinine of 146 μmol/L, and a hemoglobin of 108 g/L. Which one of the following would help confirm the diagnosis? |
| 1. Parathyroid hormone level |
| 2. Bone marrow biopsy |
| 3. 25-OH vitamin D level |
| 4. Abdominal ultrasound |
| 24-h urinary calcium |

### Content development

The MCQs reviewed in this study were selected from an existing item bank at MCC. Items selected for the purposes of this study were created using either traditional methods or AIG. A detailed explanation of the process we used for AIG has been previously published (Gierl et al., 2012). In brief, the process involves (1) having subject matter experts create a cognitive model structure in which they document an approach to a clinical problem, (2) coding this information into an item model, and (3) generating items using the JAVA-based software called Item GeneratOR (or IGOR). For the traditional methods, MCQs were written by an individual content expert and then reviewed and revised by a committee of 8–10 content experts. Approval of one MCQ typically required between 10 and 20 min. For the AIG items, a cognitive model was developed by an individual content expert, and the model was reviewed and revised by a committee of 8–10 content experts. Following this, MCQs were generated from the model using AIG (typically between 80–100 items), and then these items underwent approval by one content expert. The process for generating and approving the pool of AIG items typically required 90–120 min.

### Content selection

Since multiple AIG items are developed from a single cognitive model, the item stems are very similar. For this reason, only one item from any given cognitive model was used for this exercise to help minimize the chance that participants would recognize similar items as being derived from cognitive models. For each cognitive model, one MCQ was randomly selected.

An equal number of items developed using traditional methods were then selected. To ensure that the traditionally developed and AIG items were testing similar content, the MCQs were matched by "Objective" (e.g., abdominal pain, and jaundice). This was done to further blind the participants as to the origin of the questions (i.e., they saw 2 questions related to abdominal pain and 2 questions related to jaundice).

The selected items represented each of the disciplines tested on the MCC Qualifying Examination, Part I (i.e., Pediatrics, Medicine, Surgery, Obstetrics and Gynecology, Psychiatry, and Population Health, and the Ethical, Legal, and Organizational aspects of medicine).

In total, there were 102 MCQs for review (51 developed using traditional methods and 51 generated using AIG). The items were presented in random order, and participants were blinded as to the method used to develop the items. The correct answer for each MCQ was indicated.

### Rating instruments

Participants were asked to anonymously rate the quality of each of the test items using a number of metrics, each rated along a 5-point Likert-like scale (ranging from "strongly disagree" to "strongly agree"). These metrics were chosen from a list of 31 item-writing guidelines (Haladyna et al., 2002). From this list of 31 items, six items were identified as being the most frequently cited in the literature on MCQ writing (> 80% of sources). One of these ("Use novel material to test higher-level learning") did not seem appropriate for this exercise, as all items would be novel to the examinee, and so it was excluded. Then, a final global rating of overall quality was added. Therefore, there was a total of six quality metrics for this exercise, c.f. Fig. 1. Of note, these were not the same metrics as used in the Gierl and Lai study (2013). See Table 2, for our rationale for either including or excluding these items.

In addition, participants were asked to make a judgment about the cognitive domain being assessed (i.e., knowledge recall or application of knowledge) by selecting from a list of five options. For items that were deemed to be assessing their application of knowledge, participants were asked to also identify an associated task (i.e., diagnosis, management, communication, or professionalism), c.f. Fig. 2.

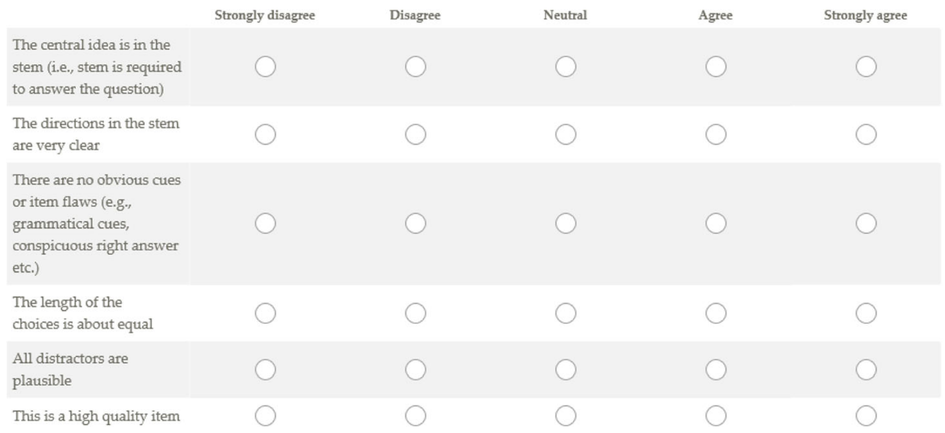
### Data collection

Ratings were anonymously collected using an online survey (Survey Monkey ®). Questions were numbered and corresponded to numbered items in the survey. Participants received the items in booklets of 10 MCQs at a time (with the exception of the final booklet which included 12 MCQs) to help prevent errors in data collection. Participants worked at their own pace and were asked to refrain from discussing the questions with each other. Two facilitators were present at all times to answer any questions.

### Analyses

Descriptive analysis was conducted to analyze ratings on the six quality metrics for items developed using either AIG or traditional methods. Overall percent agreement rates between judges were computed for the six quality metrics rating scales and the overall cognitive domain judgment scale (i.e., assessing knowledge versus application of knowledge).

A Wilcoxon two-sample test was run separately for each of the six quality metrics rating scales as well as the overall cognitive domain judgment scale. This statistic is a nonparametric analogue to a 2-sample or independent groups *t* test. For each analysis, our null hypothesis stated that there was no difference between AIG and traditionally developed item median rating for each given quality metric and for the overall cognitive judgment scale. A nominal type I error rate of 0.05 was retained for all analyses. A Holm-Bonferroni procedure was applied to control for the family-wise error rate



| | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| The central idea is in the stem (i.e., stem is required to answer the question) | ○ | ○ | ○ | ○ | ○ |
| The directions in the stem are very clear | ○ | ○ | ○ | ○ | ○ |
| There are no obvious cues or item flaws (e.g., grammatical cues, conspicuous right answer etc.) | ○ | ○ | ○ | ○ | ○ |
| The length of the choices is about equal | ○ | ○ | ○ | ○ | ○ |
| All distractors are plausible | ○ | ○ | ○ | ○ | ○ |
| This is a high quality item | ○ | ○ | ○ | ○ | ○ |

**Fig. 1** Quality metrics rating scales

**Table 2** Rationale for choice of quality metrics

| Items included in Gierl & Lai paper (2013) | Rationale for including or excluding |
|---|---|
| Every item should reflect specific content, as outlined in the test specifications | Excluded because all our items are mapped to objectives<br>Also, unreasonable to expect participants to know all our objectives |
| The question is based on important topics in the curriculum and is designed to measure key thinking and problem-solving skills | Excluded because it requires two judgments and we replaced with a question about problem-solving skills |
| The question is carefully edited, formatted, and presented using correct grammar, punctuation, capitalization, and spelling | Excluded because this is a high-stakes exam, and all items are professionally edited |
| The central idea is included in the stem, not the options | Included, but re-worded to respect the original language |
| The stem of the question is worded positively and avoids negatives such as NOT and EXCEPT | Excluded as there are no negatively worded items in our active test bank |
| Only one of the options is clearly correct | Excluded as this may be difficult for our panel to assess |
| The correct option is not cued by item writing errors | Included, but re-worded to respect the original language |
| All of the distractors are plausible | Included, but re-worded to respect the original language |

(Holm, 1979) due to the multiple dependent comparisons undertaken. This procedure controls for inflated type I error rate with a lower increase in type II error rate than that usually associated with a traditional Bonferroni correction.

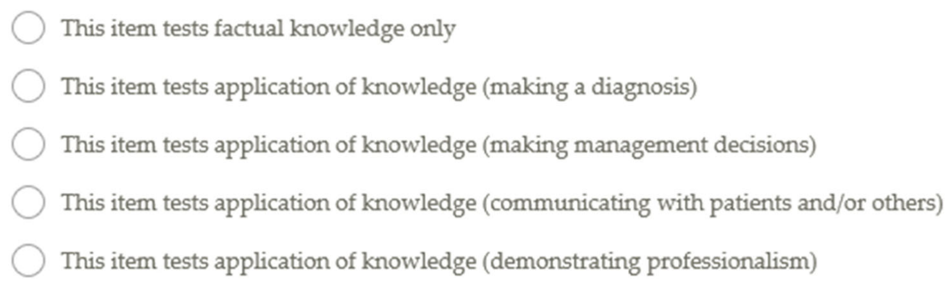
## Results

### Frequency: quality metrics rating scales

Table 3 provides the frequency distributions for the six quality metrics by item modality (AIG and traditionally developed). The distributions were highly (negatively) skewed, suggesting that categories "1" (*strongly disagree*) and "2" (*disagree*) were very sparsely selected by participating physician raters. Consequently, categories 1–3 were collapsed as category "1" (*disagree/neutral*) for the median tests, whereas categories "4" (*agree*) and "5" (*strongly agree*) remained as standalone response categories recoded respectively as categories "2" and "3".

### Frequency: cognitive domain judgments

Similarly, the cognitive domain judgment item data matrix was very sparse for some categories. Consequently, the scale was recoded as either "1" (*this item tests factual knowledge only*) or "2" (*this item tests application of knowledge*), c.f. Fig. 3.



○ This item tests factual knowledge only

○ This item tests application of knowledge (making a diagnosis)

○ This item tests application of knowledge (making management decisions)

○ This item tests application of knowledge (communicating with patients and/or others)

○ This item tests application of knowledge (demonstrating professionalism)

**Fig. 2** Cognitive domain being assessed

**Table 3** Frequency distributions for item quality metrics by item modality

| | | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| **The central idea is in the stem** | AIG | 1 | 4 | 4 | 14 | 181 |
| | Traditional | 2 | 5 | 5 | 25 | 167 |
| **Directions in the stem are very clear** | AIG | 0 | 3 | 4 | 9 | 188 |
| | Traditional | 0 | 7 | 4 | 22 | 171 |
| **No obvious cues or item flaws** | AIG | 9 | 9 | 1 | 12 | 173 |
| | Traditional | 4 | 15 | 3 | 13 | 169 |
| **Length of choices about equal** | AIG | 4 | 10 | 2 | 19 | 169 |
| | Traditional | 3 | 10 | 3 | 28 | 160 |
| **All distractors are plausible** | AIG | 7 | 24 | 9 | 34 | 130 |
| | Traditional | 2 | 25 | 3 | 43 | 131 |
| **This is a high-quality item** | AIG | 7 | 33 | 22 | 50 | 92 |
| | Traditional | 8 | 38 | 20 | 56 | 82 |

## Descriptive statistics

Table 4 presents mean, standard deviation (SD), skewness, and kurtosis values for each of the six quality metrics and the overall cognitive domain judgment scale (i.e., assessing knowledge versus application of knowledge). The vast majority of items were deemed to be assessing higher-order skills (96.1% for AIG-generated items; 91.2% for traditionally developed items) rather than simple factual recall.

For items generated using AIG, mean judgment values (out of 3) ranged from 2.15 (*this is a high-quality item*) to 2.89 (*the directions in the stem are very clear*). For traditionally developed items, mean judgment values similarly varied from 2.08 (*this is a high-quality item*) to 2.78 (*the directions in the stem are very clear*). For items generated using AIG, judgments were least variable for quality metric 2 (SD = 0.41; *the directions in the stem are very clear*) and most variable for quality metric 6 (SD = 0.86; *this is a high-quality item*). Results for traditionally developed items were very similar. Judgments were the least variable for quality metric 2 (SD = 0.53; *the directions in the stem are very clear*) and most variable for quality metric 6 (SD = 0.85; *this is a high-quality item*).

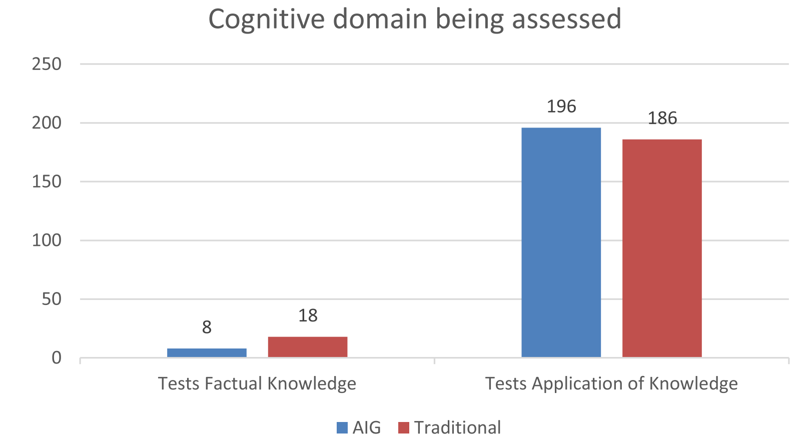


**Fig. 3** Frequency distributions for cognitive domain judgment by item modality

**Table 4** Mean, SD, skewness, and kurtosis values for each of the six quality metrics

| Item | Statistic | Development Modality | |
|---|---|---|---|
| | | AIG | Traditional |
| 1[a] | Mean (out of 3) | 2.84 | 2.76 |
| | SD | 0.47 | 0.55 |
| | Skewness | − 3.05 | − 2.22 |
| | Kurtosis | 8.43 | 3.87 |
| 2[b] | Mean (out of 3) | 2.89 | 2.78 |
| | SD | 0.41 | 0.53 |
| | Skewness | − 3.78 | − 2.42 |
| | Kurtosis | 13.63 | 4.83 |
| 3[c] | Mean (out of 3) | 2.75 | 2.72 |
| | SD | 0.61 | 0.65 |
| | Skewness | − 2.30 | − 2.08 |
| | Kurtosis | 3.69 | 2.66 |
| 4[d] | Mean (out of 3) | 2.75 | 2.71 |
| | SD | 0.59 | 0.61 |
| | Skewness | − 2.23 | − 1.91 |
| | Kurtosis | 3.60 | 2.39 |
| 5[e] | Mean (out of 3) | 2.44 | 2.50 |
| | SD | 0.80 | 0.74 |
| | Skewness | − 0.97 | − 1.01 |
| | Kurtosis | − 0.75 | − 0.31 |
| 6[f] | Mean (out of 3) | 2.15 | 2.08 |
| | SD | 0.86 | 0.85 |
| | Skewness | − 0.29 | 0.15 |
| | Kurtosis | − 1.59 | − 1.60 |

[a] The central idea is in the stem (i.e., stem is required to answer the item)
[b] The directions in the stem are very clear
[c] There are no obvious cues or item flaws (grammatical cues, conspicuous right answer, etc.)
[d] The length of the choices is about equal
[e] All distractors are plausible
[f] This is a high-quality item

With respect to the overall cognitive domain judgment item, mean values were 1.96 and 1.91 (out of 2), for AIG and traditionally developed items, respectively. SD values were equal to 0.19 and 0.28, respectively, for AIG and traditionally developed items.

Skewness and kurtosis are both useful indicators to assess the extent to which distributions differ from normality or a Gaussian curve. Values of zero are associated with a normal distribution. However, as a practical rule of thumb, values that range between [− 2.0, +2.0] generally are indicative of modest departures from normality (George & Mallery, 2010; Gravetter & Wallnau, 2012). These results show the majority of quality metric distributions that were highly skewed and leptokurtic, with the exception of scales “4” (*the length of the choices is about equal*) and “5” (*all distractors are plausible*). Item quality metric distributions are negatively skewed, suggesting that values are bunched up at the upper end of the 3-point scale (confirmed by the high mean judgment values). Finally, high positive kurtosis values for quality metric distributions for rating scales 1 to 4 suggest highly leptokurtic (or "peaked") curves, further confirmed by the low amount of variability in ratings (the low SD values).

Given that the distributions for four out of the six item quality metrics differed significantly from a normal distribution, we used non-parametric tests of significance, specifically tests that focus on the median, to compare AIG and traditionally developed quality metric judgments provided for each of the six rating scales and the overall cognitive domain judgment item.

### Wilcoxon two-sample tests

Table 5 provides the results of the Wilcoxon two-sample test for each of the six quality metrics and the overall cognitive judgment item. Note that these were sorted from lowest to highest adjusted empirical type I error rate. The first column provides the label for each indicator; the second column indicates the Wilcoxon median z-statistic; the third column outlines the (unadjusted) empirical type I error rate for each variable; and the final column shows the Holm-Bonferroni adjusted $p$ value. The Holm-Bonferroni method requires sorting the empirical $p$ values from lowest (i.e., most significant) to highest (i.e., least significant). The empirical $p$ value is compared to the adjusted nominal type I error rates of $\alpha/m$, $\alpha/m\text{-}1$, $\alpha/m\text{-}2$, etc., where $m$ is the number of tests undertaken to $\alpha$, in this case 0.05.

Using this correction, none of the empirical values are lower than the adjusted threshold. Therefore, none of the six-quality metric nor the overall cognitive judgment scale Wilcoxon two-sample tests were statistically significant. This suggests the distribution of ratings did not differ between traditionally developed items and their AIG-generated counterparts. Similarly, the distributions of items judged as either testing factual knowledge or application of knowledge that did not differ between AIG and traditionally written items. Another way to describe the findings is to state that the AIG items were not perceived as differing (better or worse) from traditionally developed items on any of the six-quality metrics or the overall cognitive domain judgment indicator.

**Table 5** Wilcoxon two-sample test results (Holms-Bonferroni adjusted) for each of the six quality metrics and the overall cognitive domain judgment item

| Item | Wilcoxon median z-statistic (out of 3) | Empirical type I error | Adjusted critical type I threshold |
|---|---|---|---|
| 2[b] | 2.55 | .01 | .007 |
| Overall cognitive domain judgment | 2.02 | .04 | .008 |
| 1[a] | 1.91 | .06 | .010 |
| 4[d] | 1.01 | .31 | .013 |
| 6[f] | 0.92 | .36 | .017 |
| 3[c] | 0.46 | .65 | .025 |
| 5[e] | − 0.43 | .67 | .050 |

Overall cognitive domain judgment is the item tests factual knowledge only/the item tests application of knowledge

[a] The central idea is in the stem (i.e., stem is required to answer the item)

[b] The directions in the stem are very clear

[c] There are no obvious cues or item flaws (grammatical cues, conspicuous right answer, etc.)

[d] The length of the choices is about equal

[e] All distractors are plausible

[f] This is a high-quality item

### Inter-judge agreement rates

Table 6 provides mean and SD inter-judge rates for each of the six-quality metrics and the cognitive domain judgment item.

For both AIG and traditionally developed items, the mean agreement rate (across all pairs of judges) was the lowest for the 6th quality metric (*this is a high-quality item*—0.27—for both AIG and traditionally developed items) and the highest for the 2nd quality metric (*the directions in the stem are very clear*—0.86 and 0.73—respectively, for AIG and traditionally written items). Not surprisingly, agreement was high on the cognitive domain judgment item (0.92 for AIG items; 0.88 for traditionally developed items) given the bi-category nature of this scale.

## Discussion

Educators' increasing need for test items has led to several advancements in the development of MCQs, including the use of AIG. AIG has led to improved efficiency of item development, as evidenced by high question output within a relatively short time frame compared to traditional methods (Gierl et al., 2012). Furthermore, an unanticipated consequence of AIG has been to improve the overall item development process because of the shift in focus to a macroscopic problem (i.e., developing a cognitive model) as opposed to the historically traditional microscopic activities (i.e., developing items on a one-by-one basis) (Pugh et al., 2016). However, regardless of the method used to develop an item, it is imperative to ensure that all items being produced are of high quality. This study demonstrated that medical educators' ratings of MCQs developed using the traditional method and those developed using our AIG method produced items that were indistinguishable, which is reassuring for a national, high stakes licensure examination.

We hypothesized that items generated using AIG would be more likely to be deemed as assessing higher-order skills than those developed using traditional methods, because of the former's reliance on cognitive models. However, the difference we found was non-significant (96.1% for AIG versus 91.2% for traditional). This may be related to the fact that all items were developed by highly experienced content experts who were

**Table 6** Mean agreement rate inter-judge agreement rates for each of the six quality etrics and the overall cognitive domain judgment item by modality

| Item | Mean (SD) | |
|---|---|---|
| | AIG | Traditional |
| 1[a] | 0.78 (0.17) | 0.68 (0.19) |
| 2[b] | 0.86 (0.10) | 0.73 (0.09) |
| 3[c] | 0.72 (0.18) | 0.71 (0.07) |
| 4[d] | 0.79 (0.05) | 0.71 (0.08) |
| 5[e] | 0.42 (0.18) | 0.49 (0.11) |
| 6[f] | 0.27 (0.16) | 0.27 (0.10) |
| Overall cognitive domain judgment | 0.92 (0.07) | 0.88 (0.04) |

Overall cognitive domain judgment is the item tests factual knowledge only/the item tests application of knowledge.

[a]The central idea is in the stem (i.e., stem is required to answer the item)

[b]The directions in the stem are very clear

[c]There are no obvious cues or item flaws (grammatical cues, conspicuous right answer, etc.)

[d]The length of the choices is about equal

[e]All distractors are plausible

[f]This is a high-quality item

instructed to avoid creating items that tested only factual knowledge. With less-experienced item writers, it is unclear if the use of cognitive models would lead to different skills being assessed. Some published studies have reported that about half of MCQs evaluated test factual knowledge only (Palmer & Devitt, 2007).

This study also sought to compare the quality of items generated using AIG to those developed using traditional methods. Our findings support those of Gierl and Lai (2013) which included a much smaller sample of questions. But in contrast to the Gierl & Lai study, the distractors produced using AIG in this study were indistinguishable from those created using traditional methods. This is likely due, in part, to the application of a systematic framework to develop plausible distractors (Lai, Gierl, Pugh, et al., 2016). Of note, the items developed using both methods were all written by content experts who had received extensive training on how to create high-quality items, which may very well account for the overall high ratings in all categories. Untrained writers have been shown to produce items that are of far lower quality (Jozefowicz et al., 2002). It is also possible that the raters erroneously judged items as assessing higher-order cognitive skills, when in fact, examinees may be using lower-order strategies to answer questions (Zaidi et al., 2018).

This study, however, does have some limitations, including the inherent subjectivity of the judgments made by the participants. However, this subjectivity was mitigated somewhat by using multiple raters, employing raters who had not been involved in the creation of the content being rated, and by providing frame-of-reference training for all participants. It is also possible that raters, although blinded, may have been able to guess which method had been used to develop a given item. This is unlikely, as previous work has demonstrated that raters' accuracy in guessing whether or not an item was generated using AIG ranges from 32–52% (Gierl & Lai, 2013). Also, content created using AIG was purposely matched to traditionally created MCQs to minimize the risk that raters would be able to ascertain the development method.

## Conclusion

MCQs developed using AIG has shown to be of high quality and to measure higher-order skills. More importantly, they were indistinguishable from traditionally developed items from the ratings collected. However, it is important to highlight the fact that the quality of items is a direct result of the quality of the cognitive model used. Training of content experts in both the principles of MCQ development and in the approach to developing cognitive models is key to ensuring that the items generated are able to assess the constructs of interest.

### Abbreviations

AIG: Automated item generation; MCQ: Multiple-choice question; MCC: Medical Council of Canada; SD: Standard deviation

### Acknowledgements

Not applicable.

### Authors' contributions

All authors contributed to the conception and design of this study. Andre De Champlain conducted all analyses. Debra Pugh wrote the first draft of the manuscript. The final manuscript was reviewed and revised with input from all authors. All authors read and approve the final manuscript.

### Funding

This study was funded by "in kind" contributions from the Medical Council of Canada.

### Availability of data and materials

A file containing all data was uploaded as a separate document along with the manuscript.

### Ethics approval and consent to participate

This study was approved by the Ottawa Health Science Network Research Ethics Board (Protocol #20170332-01H).

### Consent for publication

The authors consent to the publication of this manuscript and the accompanying dataset.

### Competing interests

Debra Pugh, Andre De Champlain, and Claire Touchie are employees of the Medical Council of Canada. At the time of the study, Debra Pugh, Mark Gierl, and Hollis Lai were paid consultants for the Medical Council of Canada.

### Author details

[1]Medical Council of Canada, 1021 Thomas Spratt Place, Ottawa, Ontario, Canada. [2]Department of Medicine, The Ottawa Hospital, Ottawa, Canada. [3]Faculty of Medicine, University of Ottawa, Ottawa, Canada. [4]Faculty of Education, University of Alberta, Edmonton, Alberta, Canada. [5]Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada.

Received: 15 January 2020 Accepted: 18 May 2020
Published online: 05 June 2020

### References

Albanese, M., & Case, S. (2016). Progress testing: critical analysis and suggested practices. *Advances in Health Science Education, 21*(1), 221–234.

Coderre SP, Harasym P, Mandin H, Fick G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Medical Education, Nov 5, 4, 23.

George D, Mallery P. (2010). SPSS for windows step by step: a simple guide and reference 17.0 update. 10th Edition, Pearson, Boston.

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726–733.

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Evaluating the psychometric properties of generated test items. *Applied Measurement in Education, 29*(3), 196–210.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765.

Gravetter, F. J., & Wallnau, L. B. (2012). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth/Cengage Learning.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309–333.

Heist, B. S., Gonzalo, J. D., Durning, S., Torre, D., & Elnicki, D. M. (2014). Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: a mixed methods study. *Journal of Graduate Medical Education, 6*(4), 709–714.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156–161.

Lai, H., Gierl, M. J., Byrne, B. E., Spielman, A., & Waldschmidt, D. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education, 80*, 339–347.

Lai, H., Gierl, M. J., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine, 28*(2), 166–173.

Lockyer, J., Carraccio, C., Chan, M.-K., Hart, D., Smee, S., Touchie, C., Holmboe, E. S., & Frank JR on behalf of the ICBME Collaborators. (2017). Core principles of assessment in competency-based medical education. *Medical Teacher, 39*(6), 609–616.

Monteiro J, Silva-Pereira F, Severo M. (2018). Investigating the existence of social networks in cheating behaviors in medical students. BMC Medical Education, Aug 9;18(1), 193.

Newman, L. R., Brodsky, D., Jones, R. N., Schwartzstein, R. M., Atkins, K. M., & Roberts, D. H. (2016). Frame-of-reference training: establishing reliable assessment of teaching effectiveness. *Journal of Continuring Education in the Health Professions, 36*(3), 206–210.

Palmer EJ, Devitt PG. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Medical Education, Nov 28;7,49.

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical Teacher, 38*(8), 838–843.

Pugh, D., De Champlain, A., & Touchie, C. (2019). Plus ça change, plus c'est pareil: Making a continued case for the use of MCQs in medical education. *Medical Teacher, 41*(5), 569–577.

Skakun EN, Maguire TO, Cook DA. (1994). Strategy choices in multiple-choice items. Academic Medicine Oct; 69 (10 Suppl),S7-S9.

Surry, L. T., Torre, D., & Durning, S. J. (2017). Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Medical Education, 51*(10), 1075–1085.

Touchie C, Streefkerk C. for the Blueprint Project Team. (2014). Blueprint project – qualifying examinations blueprint and content specifications. Ottawa, Ontario. Accessed 10 Jan 2020 at:https://mcc.ca/media/Blueprint-Report.pdf.

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work? *Academic Medicine, 93*(6), 856–859.

## Publisher's Note