Journal of Computerized Adaptive Testing

Volume 3 Number 1

October 2015

Applications and Implementations of CAT

Implementing a CAT: The AMC Experience

John J. Barnard

DOI 10.7333/15100301001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 3, Number 1, October 2015
DOI 10.7333/15100301001
ISSN 2165-6592

Applications and Implementations of CAT

Implementing a CAT: The AMC Experience

John J. Barnard
EPEC Pty Ltd., Australia

This paper gives an overview of an assessment program with overseas trained medical graduates who wish to practice medicine in Australia. A non-technical discussion of how the program was migrated from a paper-and-pencil format through online administration to computerized adaptive testing (CAT) over approximately 15 years includes various practical aspects encountered and how they were dealt with. Although theoretical CAT aspects are briefly mentioned throughout, the focus is on some guidelines and experiences that might be useful to practitioners.

Keywords: *Assessment, online, computer, adaptive, testing, medical*

The Australian Medical Council (AMC) is an independent national standards body that, among other activities, assesses overseas trained doctors who wish to practice medicine in Australia. Multiple-choice questions (MCQs or items) have been used in the assessment of the medical knowledge component of the examinations since 1978. The MCQ exam is designed to cover knowledge over five disciplines, referred to as patient groups: Adult Health, Women's Health, Child Health (Pediatrics), Mental Health (Psychiatry), and Population Health. Adult Health is further subdivided into two main domains, namely, Medicine and Surgery; while Women's Health is subdivided into Obstetrics and Gynecology. To ensure that different types of items are included in each exam, items are also labeled in terms of clinical tasks, including Data Gathering, Data Interpretation, and Data Management. Through matching content requirements for each patient group and type of item, a two-dimensional matrix was developed to identify the items for each examination, termed *the blueprint*.

In addition to changes in the number and the type of MCQs used over the years (e.g., Haladyna, 2004; Barnard, 2012), 1998 marked a significant change in assessment processes for the AMC, particularly in terms of the analysis of the examination. Before 1998, both Type A (single correct answer) and Type J (multiple correct response pattern) were used in the MCQ exams. Since 1998, only Type A MCQs with five options and a single correct option have

been used. In the same year, Rasch measurement (e.g., Barnard, 2012; Bond & Fox, 2007; Andrich, 1988; Wright, 1977) replaced classical test theory (CTT; e.g., Crocker & Algina, 1986) as the underlying measurement framework. In addition to the well-documented shortcomings of CTT (e.g., Hambleton & Swaminathan, 1985), the Rasch model was considered for its measurement properties and for the relatively small number of candidates (on average approximately 300 candidates per exam) who take the exam for robust calibration. It has been suggested that when considering high-stakes situations such as certification and licensure examinations, attention should be given to utilizing psychometric methods such as the Rasch measurement framework rather than CTT in order to better inform item-level decisions (Petrillo, Cano, McLeod, & Coon, 2015).

Different equivalent forms of the exam were necessary because the exam was administered around the world in different time zones and there was an increasing demand for more frequent exams. A further change was made in 2005, namely, a shift from the paper-and-pencil format to computerized testing. This enabled scheduling convenience and an ability to accommodate more dates and times for the increased number of candidates. Other benefits, such as faster reporting of official results, were also enabled (e.g., Bodmann, 2004). Similar results were also seen in the transition from paper-and-pencil to computerized testing for the National Council Licensure Examination (NCLEX) in 1994 in the United States of America, and the examination for the Council on Certification of Nurse Anaesthetists (CCNA) in 1996 (Vrabel, 2004).

The randomized administration of items in fixed-length exams not only met the demands at the time for the AMC, but the electronic format of the items also yielded better quality images, detailed capturing of key strokes, and other advantages associated with online delivery (e.g., Mills, Potenza, Fremer, & Ward, 2002).

Using the Rasch-calibrated items linked to a common scale, information functions could be used to construct different equivalent forms of the test that met the blueprint demands and psychometric criteria (e.g., Wagner-Menghin & Masters, 2013; Kolen & Brennan, 2004; Samejima, 1977) for international administration. Studies to ensure that candidates in the same vicinity in the computer labs were not administered the same items at the same time were undertaken. It was concluded that this was not a factor of any significance due to the randomization of the items.

However, further increase in the number of candidates put more pressure on the development of a sufficient number of different parallel examination forms to cope with the increasing number of administrations. The initial two paper-and-pencil exams administered annually increased substantially over time; and the demand on item development capacity, linking, and equating and management infrastructure prompted a quest for a better solution.

Computerized adaptive testing (CAT) was considered at this time. In CAT, the testing experience is based on the principle that an item bank exists where items can be selected based on the answers from previous questions (Vrabel, 2004). An examinee is initially administered an item with a moderate level of difficulty from the item bank, where items are calibrated by their level of difficulty. If answered correctly, the next item administered would be more difficult; and if answered incorrectly, an easier item would be selected (Vrabel, 2004). This allows a very precise estimate of an examinee's ability. CAT allows for a more individualized and tailored examination for each examinee. Instead of administering different sets of items as separate examination forms, the bank of items can be used to compile unique combinations of exams to individual examinees (e.g., Wainer, 2000; Van der Linden & Glas, 2003; Reckase, 2003). This approach addressed not only security issues in the same venue and over time zones to improve security but also improved item usage (e.g., Chang & Ansley, 2003) and allowed for increased administration. Since February 2011, CAT was implemented for the medical knowledge component of the AMC examination process. To date thousands of examinees have been successfully evaluated in this mode of examination.

In high-stakes examinations administered over time, it is to be expected that some items might "leak" when examinees become familiarized with commonly used items. Administering different sets of items to individuals from an item bank, however, made the memorization of

some items a rather difficult exercise, as individuals received tailored examinations.

The intelligent item selection process, specific to each examinee, can yield precise estimates of their ability (e.g., Weiss, 2011). Perhaps most importantly, efficiency is improved through administration of fewer items to all candidates without compromising measurement precision (e.g., Weiss, 1982; Wang & Vispoel, 1998). This not only shortened the exam drastically for the AMC examinees but also reduced item exposure in order to mitigate leaking of items.

Pilot Evaluations

Trials

While conventional exams were administered online between 2005 and 2010, CAT trials were conducted with voluntary medical university students. In June 2008, a trial with 71 fifth-year students was primarily aimed at testing the CAT platform and comparing an integrated CAT (selecting items adaptively from the complete bank) with sequential CATs (selecting items in sets by discipline).

The students were divided into four groups (18 students in each of three groups and 17 students in the fourth group). All students were administered 175 items in three hours: first, a 30-item unconstrained CAT from the whole bank, consisting of 979 items, followed by a 22-item CAT from Medicine only; a 20-item CAT from Surgery only; four 15-item CATs each from the remaining four disciplines; and, finally, a 43-item randomized conventional test. Each group of students was administered a different version. The two common tests over the four versions served three main purposes, namely, to link the four versions; to explore the possibility of including pilot, or new, items for calibration under exam conditions; and to compare the ability estimates derived from the non-common tests within the four versions with ability estimates derived from the first "fully unconstrained" CAT. The first 30-item common test was thus administered in the same format in all four versions, but with items selected adaptively so that students were not administered the same items; in the second common 43-item test, the students were administered the set of items in random order.

In all four versions, items were selected through maximum Fisher information but differed in the starting item (fixed versus from within a range) and ability estimation method. In the first version, each of the six non-common CATs started with the same item in each discipline, and maximum likelihood estimation (MLE) was used to estimate ability. In the second version, the same fixed items as in the first version were used to start each test, but a Bayesian method (EAP) was used to estimate ability. The third version was similar to the first version, but the first item was randomly selected from within the range [–1;1] logits. The fourth version was similar to the second version, with the first item selected randomly from within the range [–1;1] logits. Mixed response strings were "forced" through administering a very easy and a very difficult item in the MLE versions. Seven (CAT) ability measures were thus derived for each student. The six discipline ability measures were weighed by the inverse of their standard errors and combined to yield a single measure. This measure was compared to the measure derived from the first common CAT, and promising results were found. (Due to the small samples and the voluntary character of the trial, correlations are not reported.) Although differences between fixed and random starting items were negligible, it appeared that the MLEs for these short tests were less biased. Ability estimates were within the range –2.5 to 2.5 logits (standard errors less than 0.4 logits), and mean differences between the overall abilities and the subtest derived mean abilities ranged between 0.081 and 0.102 logits for the four groups of students. These results suggested that the exam could potentially be an integrated CAT (assuming unidimensionality) or that a series of sequential CATs by patient group could be administered to yield more precise sub-measures in addition to an overall measure. It was found, within the limitations, that subset measures derived from the sequential CATs were more robust and precise than when derived from the integrated CAT.

In August 2008, a second trial was completed with 108 voluntary fifth-year medical students from another university. Whereas the June 2008 trial was designed to explore different

starting item options, ability estimation methods, and integrated versus sequential CATs, the August trial was designed to further explore the latter. Fisher (single point) information at the current estimate was used for item selection, and MLE was used for scoring in all the CATs. Two main versions of CATs were compiled for this trial. In the first main version, 54 students completed a 150-item unconstrained CAT. The second version, which was administered to the other 54 students, explored two ways of defining subset CATs, namely, by discipline and by clinical task. Twenty-seven students were randomly allocated to each of these two versions. In the discipline version, six sequential CATs comprising between 13 and 39 items were administered, with items selected proportionally and according to the blueprint from the six disciplines—irrespective of the three clinical tasks. In the clinical task version, three sequential CATs were compiled, i.e., data gathering, data interpretation, and data management. For the six discipline-based CATs, relatively precise measures were obtained, with standard errors ranging from 0.212 to 0.350 logits. Although the clinical-task CATs generally yielded smaller standard errors (ranging between 0.200 and 0.295 logits), and thus more precise measures due to longer tests, the discipline-based sequential CATs were preferred by AMC examiners from a content point of view.

A third trial was conducted in November 2008 with 188 medical candidates. After lengthy discussions, a decision was made that an integrated exam was preferred. This trial, therefore, was designed to compare a conventional online version of the exam with a fully integrated content-constrained CAT. This trial was conducted as a formal exam. The conventional part was a standardized AMC exam on which the cut score had been determined. A statistically significant high correlation of 0.816 (N =188) was found between the two versions, which were counterbalanced in order of administration. Using the conventional derived cut score in the CAT part, logistic regression analysis indicated that 86.7% of the passes were correctly predicted from the CAT as observed in the conventional exam. These results are in accordance with many studies that compared conventional exams with online exams and CATs, and were considered promising for the implementation of CATs in the AMC exams. For example, Kingsbury and Houser (1988) concluded that “scores from CAT and paper-and-pencil tests can be considered to be as interchangeable as scores from two paper-and-pencil tests would be”; whereas Olsen, Maynes, Slawson, and Ho (1989) reported that their analyses demonstrated high levels of comparability between paper-administered tests, computer-administered tests, and CATs, with correlations ranging between 0.81 and 0.91. These results are in accordance with many more recent studies (e.g., Puhan, Boughton, & Kim, 2007).

Simulations

After the decision was made that integrated content-constrained (by discipline) CATs were to be implemented in the AMC exams, further trials were conducted in 2009 and 2010 to obtain more detailed information. Post-hoc (real-data) and monte-carlo simulations were also used to further explore issues such as a starting rule, item usage and exposure, item review, test overlap, and piloting new items.

Information functions for items, pools (subsets of the item bank), or the entire bank can be used to inspect the measurement precision expected at each point on the ability continuum. The target information function is crucial for the implementation of CATs, as it will give an indication of the measurement precision, which is especially important at the cut score.

To minimize idiosyncrasies, different random seeding was used in 20 replications of the same requirements. The simulations were based on 120 items and 335 candidates. The 120 items were specified as the number of scored items intended in the CATs, and the 335 candidates was derived as an indicative average number of candidates per exam. Past AMC exams indicated that candidate abilities seldom fell outside the range –2 logits to 2 logits; thus, this range was used. In the simulations using CATSim (Weiss & Guyer, 2012), alpha and beta values were used to control the beta distribution to mimic the actual normal distribution as closely as possible. For the first 10 simulations, alpha and beta were both set at 5.0; and in the second 10 simulations, alpha and beta were both set at 1.0. The initial ability (θ) estimate

was set at a value close to the cut score of zero logits. θ was estimated by MLE as implemented in the preferred CAT algorithms, and subsequent items were selected by maximum Fisher information at the current θ estimate. A variable termination was set at a maximum standard error (SEM) of 0.20 in order to explore the number of items required for the specified precision. This level of precision was much higher than suggested by Wagner-Menghin & Masters (2013), who recommended 20 items per difficulty level to achieve a SEM of 0.39 if approximately 30 items are administered adaptively.

In one study, eight simulations were completed: two with positive θs and item difficulties (between 0 and 2 logits), two with negative θs and item difficulties (between –2 and 0 logits), two with positive θs (0 to 2 logits) and negative item difficulties (–2 to 0 logits), and two with negative θs (–2 to 0 logits) and positive item difficulties (0 to 2 logits) around the cut score. It was found that for 99.6% of the simulated candidates, an SEM of 0.20 or less was achieved within 120 items. Further simulations, removing either the SEM criterion or the maximum number of items criterion (120 items), yielded similar results. For example, in some cases SEMs less than 0.20 were achieved for all candidates with 71 or fewer items, while other simulations required slightly more than 120 items to achieve the set level of precision of an SEM of 0.2 logits or less. The range of item difficulties and/or candidate θs was modified to [–2.5; 2.5] and also to [–3; 3], resulting in slightly more items required for a small number of items to achieve an SEM of 0.20 or less. From these simulation results, it was concluded that high precision CATs can be expected if 120 items are adaptively administered from a bank of 1,800 items for conditions similar to those simulated. These results were also validated through post-hoc (real-data) CAT simulations in which estimates derived from the online exams administered between 2005 and 2010 were used and similar results were found.

The Blueprint and Exam

The AMC's item bank consists of thousands of calibrated items and is subdivided into banks of 1,800 items each. Only one bank is active at any time for security and other reasons. The items in each bank are located in a cell of a two-dimensional matrix with the six patient groups (disciplines) as columns and the three clinical tasks as rows. Each candidate is administered 35 Medicine, 25 Surgery, 15 Women's Health, 15 Child Health, 15 Mental Health, and 15 Population Health items that are scored. Randomization is used for balancing clinical tasks. In addition to the 120 scored items, 30 pilot items are included in the exam so that each candidate is administered 150 items. Once sufficient data has been collected on the pilot items, they are calibrated and linked to the constructed scale for use as future scored items. The trials showed that allowing 1.4 minutes per item was adequate so that the exam could be completed in 3.5 hours, as compared to 7 hours in the conventional exams administered previously.

The exam is thus a content-constrained fixed-length CAT. Variable-length CATs were initially considered but were not further considered because they cannot guarantee exact blueprint matching and lend themselves to criticism and possible accusations of unfairness if some candidates are administered more or fewer items than others, and therefore differential content coverage.

Dimensionality

No test can be perfectly unidimensional. An assumption has to be made that it is unidimensional "enough" if the items covary to an acceptable extent to measure something in common. This assumption was made for the AMC's MCQ exams where the construct is defined as medical knowledge and comprises six main patient groups. The unidimensionality assumption was tested extensively through factor analytic analyses using tetrachoric correlations, implementing Bejar's method (e.g., Liou, 1988), using *t* tests based on residual principal components, and evaluating fit indices (e.g., Hambleton & Swaminathan, 1985).

The relationship between the overall θ estimate (as a dependent variable) and the discipline/patient group θ estimates (as predictors) was explored. The results of the standard multiple regressions typically resulted in R^2s of 0.95, indicating that approximately 95% of the variance in the overall measure was explained by the model. This is a highly significant result ($p < 0.0005$). The standardized β coefficients showed that Medicine made the strongest unique contribution to explaining the overall θ, followed by Surgery. Principal components analyses were also conducted. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) values were without exception greater than 0.6 and the Barlett's Test of Sphericity value was significant ($p < 0.0005$). Using Kaiser's criterion, only one component had an eigenvalue of 1 or greater. The scree plots clearly showed a change in the shape of the plot, and therefore only the one component was extracted. The loadings of the variables indicated that all six patient groups loaded strongly (above 0.40) on the extracted component. From these results, it could be concluded that the unidimensionality assumption was valid to the extent that the different patient groups measured a common construct—medical knowledge.

A Calibrated Item Bank

Implementing and managing a CAT program requires careful considerations and some technical and psychometric demands. A first requirement is the availability of a calibrated item bank, since algorithms for item selection require that an estimate of the examinee's θ and item difficulty must be located on the same scale. Although it is possible to base CATs on CTT (Rudner, 2002), the sample dependent item statistics and especially the lack of relationship between item difficulty and person score makes this a less viable option. Rasch measurement was therefore implemented to achieve the desired measurement properties and accommodate the relatively small numbers of candidates.

In any sensible ongoing CAT program, the entire bank of items is not utilized as a whole. The AMC's bank is divided into banks of 1,800 items each, which are rotated once items have reached a certain usage frequency. This procedure not only improves security but also ensures acceptable exposure of items. The banks are linked so that performance and standards retain their meaning over banks. The equivalence of banks is ensured through matching the number of items in the cells of the blueprint and also by overlaying bank information functions.

Because the exam is primarily intended to make pass/fail decisions, most items have difficulties around the cut score of zero logits. Items for each bank were thus selected to have a "normal" bank information function in order to yield precise measures, especially for borderline performances. Approximately 2% of the items have difficulties less than −2 logits and approximately 2% of the items have difficulties greater than 2 logits. Figure 1 shows a typical bank information function.

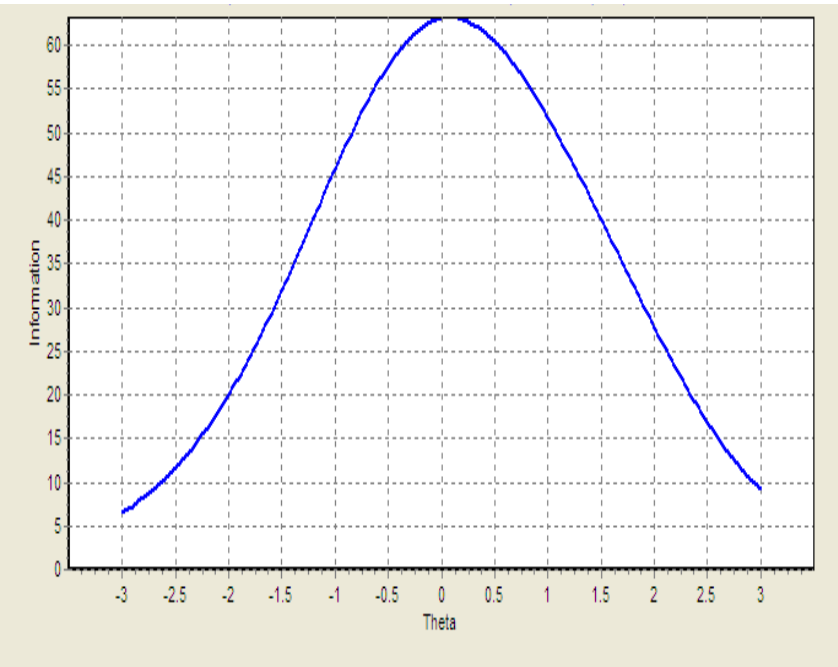
It should be noted that the number of items in a bank is not the only requirement for the administration of "good" CATs. The item difficulties should also span the difficulty continuum with more items where usage is expected to be high—usually around the cut score.

Any CAT program relies on the generation and addition of new items. Pilot items are seeded randomly in the exam, and responses are accumulated until enough data has been obtained for calibration and linking to the bank's scale. Difficulty estimates for the pilot items are obtained by using the θ estimates of the candidates as derived through the scored items.

Usage, Exposure, and Overlap

For many applications of CAT, item usage is not of prime importance. However, it becomes an issue if the CATs are not functioning well enough for the specific application and/or there is motivation for examinees to get access to the items to learn the answers. In a high-stakes exam such as the AMC's MCQ exam, the latter is likely to be an issue; and, therefore, monitoring item usage is important.

Figure 1. Typical Bank Information Function



Item usage is a joint function of bank size, bank information structure, and the distribution of candidate θ estimates. Note that usage has two meanings: Items should not be over-utilized; on the other hand, items should not be underutilized, as this would imply that the bank is not used effectively. In the AMC's program, item usage is closely monitored through plotting frequencies of items used over the difficulty continuum against the number of items available in the bank. It was found that if the administration algorithms remain unaltered, the patterns of item usage do not change significantly over exams and time.

Item exposure is directly linked to item usage and also has at least two meanings, namely, in the context of test overlap and in terms of general item exposure. The former refers to the percentage of common items between exams administered to two or more candidates in the same session, whereas the latter refers to the rate at which an item is administered. Different formulae from discrete mathematics were used to explore test overlap between pairs of exams, and it was concluded that the ratio of 10:1 was sufficient to ensure that test overlap is not a factor. (The AMC implements a ratio of 15:1.)

The Sympson-Hetter method (Sympson & Hetter, 1985) is perhaps the most popular method implemented to control the exposure of items in CATs. This method and some of its modifications were initially considered for the AMC exams; but the trials and full implementation to thousands of candidates to date have shown that the exposure could be dealt with effectively through the CAT algorithm as implemented in the AMC's CATs.

Over-usage of items might result in item drift, i.e., differences in item parameter values over time. Due to implementation of stochastic processes, small variations are to be expected; but significant differences might be due to items becoming compromised through becoming known or to other factors that need to be closely monitored. Different methods can be used for this. A global item-oriented test for parameter drift using a Lagrange multiplier (λ) is based on evaluating a quadratic function of the partial derivatives of the log-likelihood function of the general model evaluated at the maximum likelihood estimates of a parameterized model (e.g., Bock, Muraki, & Pfeiffenberger, 1988; Suarez-Falcon & Glas, 2010). Alternatively, a method targeted at parameter drift due to item disclosure can be used. This latter method

addresses the one-sided hypothesis that the item is becoming easier and is losing its discrimination power. Whether λ statistics supporting the detection of specific model violations or the cumulative sum statistic is used, it is important to monitor parameter drift. To date, drift has not been detected as a problem in the AMC's CATs.

Starting Rule, Item Selection, and Scoring

Theoretically, a CAT can be started at any level of difficulty. To minimize exposure, a single item is usually not identified as the first item to be administered to all candidates. In the AMC program, random selection of the first item is limited to a specific difficulty range around the cut score. The purpose of starting around the cut score is to focus θ estimations in this area early so that it can quickly be determined whether the candidate passed or failed the exam. Using a Bayesian prior distribution, the initial item is randomly selected from the available items in the defined range to minimize the Bayesian posterior variance. The maximum posterior precision (MPP) item selection is continued until at least one correct and one incorrect response is obtained, after which the driver switches to a maximum likelihood scoring algorithm and a maximum information item selection algorithm.

Consecutive items are selected randomly from the content areas from which the item that yields the most (Fisher) information is selected using a randomesque factor of four (e.g., Kingsbury & Zara, 1989; Featherman, Subhiyah, & Hadadi, 1996). It was found that a factor of four sufficiently randomized the items and that loss in information was negligible even for the item providing the least information out of the four items. From a certain item position, the content area with the greatest divergence from the target is given priority in order to also meet the content constraint. This process ensures content balancing and minimizes overlap while ensuring that the exam remains integrated.

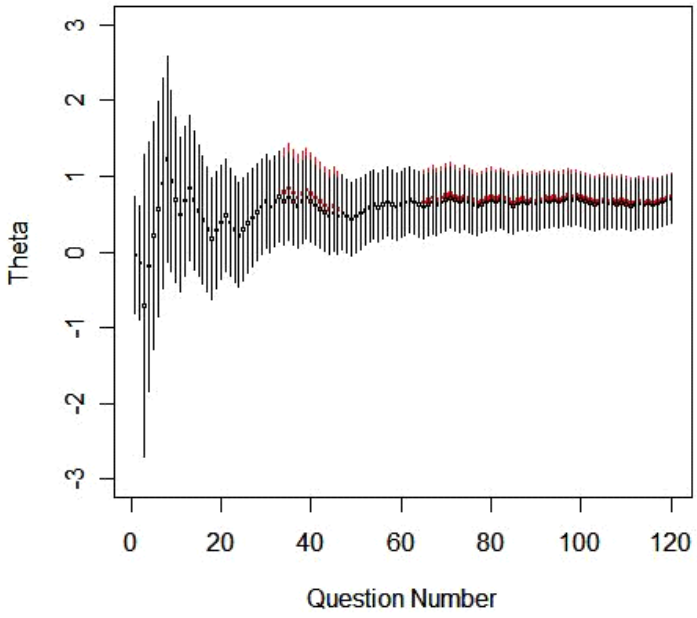
One of the main advantages of CATs over conventional tests is control of measurement precision. It has been widely reported that a CAT can yield at least the same precision of measures with half or fewer the number of items included in a conventional test (e.g., Babcock & Weiss, 2012). From a Rasch perspective, measurement precision is conditional on the trait being measured and is not a constant for a set of measurements that depends on the cohort, as is the case in CTT. Precision in the Rasch model depends on an individual interacting with a set of items and not on a cohort, thus allowing it to vary and to be evaluated at the candidate level.

The AMC's 120-item CATs yield SEMs of about 0.185 logits on average. This can be interpreted as classical reliabilities well above 0.95 for each candidate (e.g., Thissen, 2000). When compared to the conventional exams comprising 240 scored items, the Spearman-Brown prophecy formula indicates that this precision would require more than 600 scored items administered conventionally, as compared to the 120 items administered in the CATs.

For each exam, the response patterns of candidates are closely inspected and pathway figures of their θ estimates are plotted. An example of such a pathway figure is shown in Figure 2. The θ estimate is shown as a circle, and the vertical lines show the decreasing standard error range for every provisional θ estimate. Note the convergence from around the 30th item, after which the θ estimate changed by less than 0.15 logits, while the SEM band was increasingly reduced.

In a CAT, a candidate's θ is re-estimated after each response, and the next item to be administered is determined by the most current θ estimate. In principle, this implies that previous responses cannot be changed in the process. However, since likelihood estimation is implemented, changes in responses can be accommodated and θ estimates can be derived from final responses. In the AMC CATs, candidates have to respond to each item until the end of the exam (item 150) has been reached, after which responses to any item(s) can be reviewed and changed in the available time. Then, using the item difficulties of the items administered and the final responses, final θ is estimated. Initial and final responses are recorded and Pearson correlations between initial and final abilities of approximately 0.993, on

Figure 2. Pathway Display of θ Estimates



average, were found over multiple exams. Thus, although there might be differences between final and initial θ estimates for some individual candidates, it can be concluded that, generally, it does not matter if candidates are allowed to revise and change some responses. Allowing candidates to change responses is thus more a policy decision than a psychometric issue.

Incomplete Exams

A candidate is allowed 3.5 hours to complete 150 items (120 scored and 30 pilot items), which gives an average of 1.4 minutes per item. The pilot items are administered randomly throughout the exam. If a candidate does not finish the exam in the available time, it means that the candidate had an advantage of more time per item, on average.

In order to maintain fairness, candidates who do not complete an exam are "penalized"; and the penalty should be a function of the number of scored items to which the candidate has not responded. In other words, the penalty should be more severe for a candidate who completed only (say) 102 scored items than for a candidate who completed (say) 114 scored items. A penalty procedure was derived from an equation for scoring items to which the candidates have not responded. This index is used together with the candidate's last θ estimate and SEM in the penalty. In addition to having results that match the blueprint, the main purpose of implementing this procedure is to discourage candidates from not completing the exam and thereby having a possible advantage over candidates who do complete the exam.

Repeat Candidates

In an ongoing program such as the AMC's examination, it is inevitable that there will be candidates who take the exam more than once after a failed attempt. Candidates who take the exam again after an unsuccessful first attempt to pass are administered a CAT in which previously administered items are masked (temporarily excluded from items available for administration). The exam is otherwise exactly the same as if it were a first attempt. Irrespective of where the candidates' exams start, they will converge to the same location unless the candidates actively increased their knowledge through studies, bridging courses, etc.

Feedback

In CATs all examinees theoretically answer about 50% of the items correctly. However, a candidate administered more difficult items will have a higher θ estimate than a candidate who was administered easier items. It is thus evident that performance cannot be reported in terms of number-correct scores but should be based on θ estimates. The θ estimates are commonly in the range of -2 to 2 logits and computed to at least three decimal places. For AMC reporting purposes, the θ estimates are converted to a scale with a mean of 250 and a standard deviation of 50. This conversion serves multiple purposes. First, the results are given as positive whole numbers; and, second, they cannot be confused with number-correct scores or percentages.

Currently, only the overall performance is used to determine whether a candidate passed or failed the exam. Because each exam is content constrained and includes at least 15 items from each patient group, diagnostic feedback is also provided in terms of descriptors of performance for each patient group.

Discussion and Conclusions

Migrating a paper-and-pencil exam to an online exam and CAT requires trials, simulations, and psychometric considerations. Before the program is implemented, bank size, estimation algorithms, constraints, and other important aspects of an examination program can be investigated through the trials and simulations. Once implemented, CATs have many advantages over conventional testing, including security, measurement precision, and efficiency.

References

Andrich, D. (1988*). Rasch models for measurement*. Newbury Park, CA: Sage.

Babcock B., & Weiss, D.J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1-18. CrossRef

Barnard, J.J. (2012). *A primer on measurement theory*. Melbourne, Australia: Excel Psychological and Educational Consultancy.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275-285. CrossRef

Bodmann, S.M. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31*(1), 51-60. CrossRef

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). London, England: Lawrence Erlbaum.

Chang S.-W., & Ansley, T.N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40* (1), 71-103. CrossRef

Crocker, L.M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston Inc.

Featherman, C.M., Subhiyah, R.G., & Hadadi, A. (1996, April). *Effects of randomesque item selection on CAT item exposure rates and proficiency estimation under 1- and 2-PL models*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). London, England: Lawrence Erlbaum.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff. CrossRef

Liou, M. (1988). *Unidimensionality versus statistical accuracy: A note on Bejar's method for*

detecting dimensionality of achievement tests. Retrieved from the University of Minnesota Digital Conservancy, http://purl.umn.edu/104313. CrossRef

Kingsbury, G.G., & Houser, R.L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. CrossRef

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking. Methods and practices* (2nd ed.). New York, NY: Springer. CrossRef

Mills, C.N., Potenza, M., Fremer, J.J., & Ward, W.C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. London, England: Lawrence Erlbaum.

Olsen, J.B., Maynes, D.D., Slawson, D., & Ho, K. (1989), Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. *Journal of Educational Computing Research, 5*(3), 311-326. CrossRef

Petrillo, J., Cano, S.J., McLeod, L.D., & Coon, C.D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Journal of the International Society for Pharmacoeconomics and Outcomes Research, 18*(1), 25-34. CrossRef

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment 6(*3), 4-20.

Reckase, M.D. (2003, April). *Item pool design for computerized adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Rudner, L.M. (2002, April). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(2), 233-247. CrossRef

Suarez-Falcon, J.C., & Glas, C.A.W. (2010). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*. *56*(1), 127-143. CrossRef

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Personnel Research and Development Center.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed), *Computerized Adaptive Testing: A primer* (2nd ed., 159-183). Mahwah, NJ: Lawrence Erlbaum.

Van der Linden, W.J., & Glas, C.A.W. (Eds.). (2003). *Computerized adaptive testing: Theory and practice.* Dordrecht, Netherlands: Kluwer.

Vrabel, M. (2004). Computerized versus paper-and-pencil testing methods for a nursing certification examination: A review of the literature. *CIN Computers, Informatics, Nursing 22*(2), 94-98. CrossRef

Wagner-Menghin, M.M., & Masters, G.N. (2013). Adaptive testing for psychological assessment: How many items are enough to run an adaptive testing algorithm? *Journal of Applied Measurement, 14*(2), 1-12.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). London, England: Lawrence Erlbaum.

Wang, T., & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35* (2), 109-135. CrossRef

Weiss, D.J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6* (4), 473-492. CrossRef

Weiss, D. J., & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing*. St. Paul, MN: Assessment Systems Corporation.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-166. *CrossRef*

Author Addresses

John J. Barnard, EPEC Pty Ltd., P. O. Box 3147, Doncaster East, VIC, 3109, Australia; Medical School, University of Sydney, Edward Ford Building A27, Sydney, NSW, 2006, Australia. Website: www.epecat.com; Email: John@EPECat.com

,