EQAO

LITERATURE REVIEW

Leveraging Multi-Stage Computer Adaptive Testing for Large-Scale Assessments

November 2020

Leveraging Multi-Stage Computer Adaptive Testing for Large-Scale Assessments

EQAO is transforming Ontario's approach to large-scale assessments. The agency has worked with measurement and technology experts to create an assessment design that will propel EQAO's modernization plan forward. The first step is a transition to a modernized digital interface. Digitalization opens new opportunities to engage Ontario's learners as they demonstrate their knowledge and skills.

In the 2020–2021 school year, EQAO will field test a digitalized and modernized platform for the Grade 9 Assessment of Mathematics, where students will complete a multi-stage computer adaptive test (msCAT). Leveraging the opportunities that come with technology, the assessment is designed to align to best practices, be relevant to students' learning experiences and include interactive and engaging items that can increase test engagement and lead to a positive test experience (Lifelong Achievement Group & Martin, 2015).

Researchers and jurisdictions that have implemented msCAT have demonstrated that this type of testing is well-suited for school-based accountability testing due to a range of benefits for students, educators and test developers (Bejar, 2014; Yamamoto et al., 2018a). Automated reports of student results will be available to students, parents and educators through an online system. For schools and boards, the move to msCAT affords longer test administration windows, reduced reporting time and precise measurement of proficiency levels for individual students (Hendrickson, 2007).

EQAO is modernizing at a time when schools and boards need flexibility and responsiveness. The items on the Grade 9 mathematics assessment will be machine scored, allowing schools and boards to access data and results quickly. Key features of msCAT allow for flexible test scheduling and the timely use of reliable data to inform improvement planning, program development and school and board decision making.

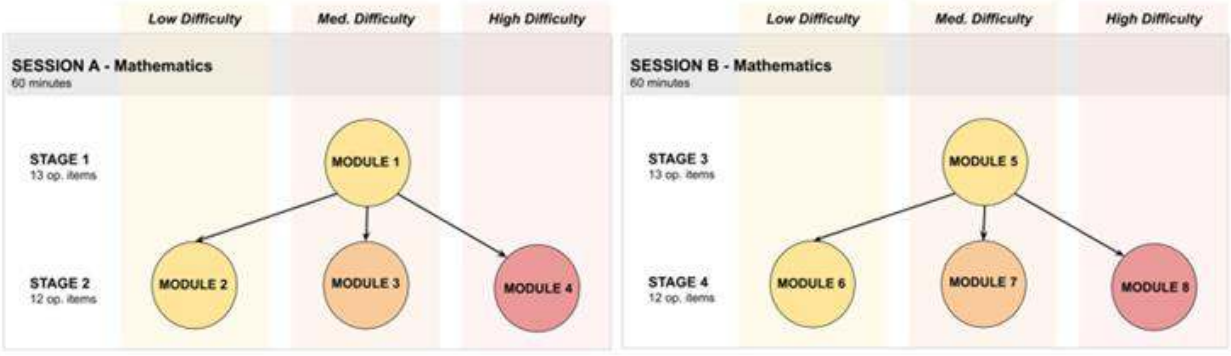
What is multi-stage computer adaptive testing (msCAT)?

A multi-stage computer adaptive test (msCAT) is a form of computer test that adapts to a student's level of proficiency according to the student's achievement on a set of items, called a *module* (Luecht & Sireci, 2011).[1, 2] An msCAT begins with a Stage 1 module, often consisting of items of medium difficulty overall, to gauge students' level of proficiency (Hendrickson, 2007). Depending on their achievement, students are next presented with one of three modules in Stage 2, which vary in overall difficulty (see Figure 1 below). This process is repeated depending on the number of stages of the test. As a result, students taking an msCAT receive different modules and complete different overall versions of the test. Although students may receive different items, modules are carefully constructed to meet specific statistical and content requirements (Luecht & Sireci, 2011). By employing statistical and psychometric models and analyses, student results are put on the same scale and can be compared (e.g., Levels 1 to 4) even though students may complete different modules.

[1] Multi-stage computer adaptive testing is also commonly referred to as multi-stage adaptive testing (MST) or computer adaptive multi-stage testing (ca-MST).

[2] Sets of items are also referred to as *testlets* in the literature.

Figure 1. *Diagram of an msCAT Model*



What are the differences between msCAT, computer adaptive tests (CATs) and linear tests?

Linear tests, such as standard paper-and-pencil tests or computer-based fixed-form tests, present all students with the same items in the same order. In comparison, computer adaptive tests (CATs) adapt to a student's level of proficiency item by item. The response to each item informs the selection of the next item. The adaptation and computer functionalities of both msCATs and CATs provide detailed information about a student's knowledge and skills more efficiently than do linear tests. However, the advantage of msCATs (module-level adaptation) over CATs (item-level adaptation) is that they reduce computer processing demands while maintaining high measurement precision (Hendrickson, 2007; Yamamoto et al., 2018b). Table 1 summarizes the main differences between linear tests, msCATs and CATs. For more information about the comparison among linear tests, CATs and msCATs, see Hendrickson (2007) and Yamamoto et al. (2018b).

Table 1
Comparison of Linear Tests, CATs and msCATs

Linear tests	**Computer adaptive tests (CATs)**	**Multi-stage computer adaptive tests (msCATs)**
Fixed test, no adaptation • All students are presented with same items in same order • Test responses are scored after test submission • Students can revise all responses prior to test submission • Test yields less precise information about students who are struggling or excelling	Item-level adaptation • Students are presented with different items depending on their previous responses • Test responses are scored immediately after each completed item • Students cannot revise responses to past items after proceeding to a new item	Module-level adaptation • Students are presented with different modules depending on their previous responses • Test responses are scored immediately after each completed module • Students can revise responses within a module prior to proceeding to a new module

	• All student achievement levels measured with high precision • Demands on data management and computer processing to track performance on each item are high	• All student achievement levels measured with high precision • Data management and computer processing demands to track performance on each module are fewer

What are the benefits and drawbacks of msCATs, compared to linear tests and computerized adaptive tests (CATs)?

This section compares the three test formats and describes their relative advantages and disadvantages. Comprehensive reviews are available from Hendrickson (2007) and Yan, Lewis & von Davier (2014).

A main advantage of adaptive tests (CATs and msCATs) over linear tests is that adaptive tests are more efficient (i.e., more precise, often with fewer items) in measuring student proficiency across the range of the achievement scale (e.g., Levels 1 to 4). On linear tests, all students take the same test items. Items of low difficulty contribute little information to the measurement of student proficiency at the higher end of the achievement scale. Similarly, items of high difficulty contribute little information to the measurement of student proficiency at the lower end of the scale. Because of this, precision of measurement varies along the achievement scale with linear tests, and typically students of average proficiency are measured more precisely, while students at the two ends of the achievement scale (i.e., Levels 1 and 4 in Ontario) are measured less precisely.

On adaptive tests, items presented to students are tailored to their individual proficiency level, so equally precise measurement of all students across the achievement scale is possible, often with fewer items. Adaptive tests are well-suited for student achievement testing where the range of skills or abilities to be measured is very broad and high measurement precision is required.

Some studies suggest that matching items to skill levels can also help to improve test engagement among students. With msCAT designs where students are directed to modules better matched to their ability based on their previous answers, some students report higher self-efficacy and lower disengagement than with fixed online assessment designs (Lifelong Achievement Group & Martin, 2015; Martin & Lazendic, 2018). Other research into online testing has explored the potentially motivating properties of msCATs, suggesting that because “content is delivered to a more appropriate level of difficulty, students will be more motivated to persist and complete the test” (Lifelong Achievement Group & Martin, 2015, p. 8).

The following are other advantages of adaptive tests over linear tests: (1) answer copying can be reduced substantially (with msCATs), and even avoided (with CATs), because students sitting next to each other often take different modules of items or different items; (2) computerized testing (adaptive or not) provides more scheduling flexibility for testing and more efficient test result reporting (Hendrickson, 2007).

Adaptive tests also have some drawbacks for test developers compared to linear tests. For example, the amount of effort required to assemble, implement and administer adaptive tests is much greater than for linear tests. Adaptive tests are demanding for routing decision making, data management and computer processing. Developers of adaptive tests have less control than linear test developers in the administration of final tests. For example, test developers have less control over item ordering. With adaptive tests, there are also some unique item and test exposure control issues and security concerns (Hendrickson, 2007; Yan, et al., 2014).

Compared to CATs, msCATs offer additional benefits. First, msCAT developers have more control over item ordering and content balancing. Second, students writing CATs are not able to review or change their answers to previous items, but on msCATs, students can preview items, review their answers and make changes if needed within each module. Third, the work on routing or adaptation points, data management and computer processing for msCATs is less demanding than for CATs. A disadvantage of msCATs is that more items are needed to achieve the level of measurement precision that CATs offer (Hendrickson, 2007; Yan, et al., 2014).

After considering these relative advantages and disadvantages, some testing programs, such as Australia's National Assessment Program—Literacy and Numeracy (NAPLAN), which is administered to students in Grades 3, 5, 7 and 9, have chosen msCATs over CATs. Other testing programs, such as the Graduate Record Examination (GRE) revised general test, which is an admissions requirement for graduate schools in the United States and Canada, have moved from CATs to msCATs. The GRE administered CATs from 1993 to 2011 but changed to msCATs in 2011 (Robin, Steffen, & Liang, 2014).

The relative benefits and drawbacks of linear tests, CATs and msCATs are summarized in Table 2, adapting Yan, Lewis and von Davier (2014). Overall, msCAT is a "balanced compromise" (Hendrickson, 2007, p. 44) between linear tests and CATs, incorporating most of the advantages of linear tests and CATs. Bejar (2014) makes a strong argument that the most valuable use of msCATs may lie in school-based accountability testing, because of the wide range of skills and abilities that can be measured precisely as well as the many benefits that msCATs offer over linear tests and CATs.

Table 2
Benefits and Drawbacks of Linear Tests, CATs and msCATs

	Benefits	Drawbacks
Linear Tests	• Relatively straightforward to develop, assemble and administer • Gives test developers most control over the final test	• Offers the same items to all students • Measures inefficiently, especially at the two ends of the achievement scale
CATs	• Measures efficiently (i.e., facilitates more precise measurement across the achievement scale, with fewer items) • Facilitates efficient test result reporting • Discourages answer copying	• Requires most effort to implement • Gives test developers least control over final tests • Does not allow answer review or changes • Increases item exposure control difficulty

	• Facilitates flexible scheduling	• Costs more to administer • Entails demands on routing, data management and computer processing
msCATs	• Measures efficiently (i.e., facilitates more precise measurement across the achievement scale, with fewer items) • Facilitates efficient test result reporting • Requires moderate effort to assemble and implement • Gives test developers more control over the final tests • Allows item preview and answer review/change within a module • Reduces answer copying • Facilitates flexible scheduling	• Longer than CAT but shorter than linear tests • Increases item exposure concerns • Costs more to administer • Entails demands on routing, data management and computer processing (less so than CAT)

Source: Adapted from Yan, Lewis, and von Davier (2014)

It is possible for computer-based testing to be more accessible to students with special education needs, but there are also potential drawbacks regarding measurement. With the computer delivery mode, a wide variety of test accommodations, such as read-aloud, high contrast and zooming in and out, can be implemented at a low cost. However, along with these opportunities come challenges for measuring the abilities of students with special education needs (Bejar, 2014; Stone & Davey, 2011). For example, students with specific disabilities may have poor basic skills but good high-order skills. This distinctive knowledge pattern may not be accurately and stably measured by adaptive tests because such tests typically present students with more high-difficulty items after students get the low-difficulty items correct. To address the needs of students with different learning patterns and to reduce routing errors, it is important that the Stage 1 module or the routing module include items with a wide range of difficulty and cognitive complexity, and that the routing module be of sufficient length.

In the past several decades, many research studies on adaptive testing have focused on its psychometric and technical properties, but research on its psychological effects, for example, on student motivation and engagement, is limited. The available research findings are mixed. Some studies show that adaptive testing has no effect on student motivation and engagement (Ling, Attali, Finn, & Stone, 2017) or that it produces elevated anxiety (Martin & Lazendic, 2018). Others show higher motivation and engagement among female than among male students, as well as among Grade 9 students than among students in Grades, 3, 5 and 7 (Martin & Lazendic, 2018). Investigating the psychological effects of adaptive testing on students is an area for further research and will be a focus for EQAO.

Which jurisdictions have implemented msCAT designs, and what have they found?

Two international large-scale assessments have adopted an msCAT design: The Programme for the International Assessment of Adult Competencies (PIAAC), which is a standardized international survey that measures adults' (15- to 65-year-olds) proficiency in literacy, numeracy and problem solving (OECD,

2019a), and the Programme for International Student Assessment (PISA), which measures the skills of 15-year-old students in reading, mathematics and science (OECD, 2019b, p. 26).

A study conducted by Yamamoto and colleagues (2018a) that evaluated the design of PIAAC found that its msCAT design resulted in more efficient and more accurate measures of performance among both lower and higher performing respondents. Yamamoto et al. (2018b) conducted another study to investigate the technical qualities of PIAAC and PISA. Their study concluded that the msCAT design of PIAAC and PISA applies to large-scale assessments focused on individual-level scores and aimed at measuring complex constructs or different sub-scales of a construct. Thus, msCAT design is recommended when multiple items are associated with certain units or modules, and when items require context building. An msCAT design can also allow for the reporting of sub-scores by strand and skill.

Since 2018, Australia's NAPLAN has been gradually implemented online, with 15% of schools administering NAPLAN online in 2018 and 50% in 2019. It is expected that by 2022 all schools in Australia will be completing the NAPLAN fully online. A study conducted in 2013 (ACARA, 2014) examined the measurement precision of the msCAT design compared to a fixed linear test design with Grades 3, 5, 7 and 9 students. The results show that the precision of measurement of student ability was higher for students taking the adaptive test than for those taking the linear test. Another study conducted in 2014 compared achievement, motivation and subjective experience of students taking the fixed linear design test and of those taking the msCAT (Lifelong Achievement Group & Martin, 2015). The results of the study found no differences in student achievement but higher accuracy of achievement estimates for students taking the msCAT. Moreover, the Grades 7 and 9 students taking the msCAT demonstrated higher self-efficacy and lower disengagement while taking the test and reported a more positive subjective experience (e.g., test ease, clarity, comprehension).

Concluding remarks

Overall, msCAT is well-suited to resolve many of the challenges of the previous paper-based EQAO testing program. In an advancing age of digital learning in which students are becoming more familiar and comfortable with technology, the computer aspect of msCAT design aligns with current student experiences, for example, an increasing emphasis on digital tools to enhance learning. The adaptive nature of msCAT meets students where they are, honours their strengths and weaknesses, and can lead to a more positive testing experience. The computer-based adaptive mode also eases the burden of test administration by eliminating lengthy and rigid test preparation procedures (e.g., printing, shipping) and the tight administration windows required for the security of paper-based tests. The msCAT format, through its provision of different modules to different students, increases security and flexibility in test scheduling. For the paper-based tests, reports are typically released months after a student completes the test because responses are manually scored. With msCAT, students, parents and educators will receive automated reports almost immediately following the completion of the test, which allows for more timely and detailed feedback.

In the course of field testing the Grade 9 Assessment of Mathematics, EQAO will be monitoring and evaluating student, school and board experience with the msCAT system, in order to track the changes associated with the agency's move to online adaptive assessments and the impact of those changes on the measurement of student learning and on user experience. Students, parents, educators and school and board leaders will be asked to share their experiences with test administration and reporting,

and psychometric and process data will be closely analyzed. Other success indicators, such as system performance and technical support data, will inform how msCAT is implemented. Drawing on insights from rapid-cycle evaluation, this monitoring and evaluation process is designed to allow EQAO to identify any areas in need of improvement and apply refinements to subsequent administrations. This approach is critical to ensuring that modernized EQAO assessments are delivering the benefits of msCAT design to students, educators and the education system as a whole.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2014). *Tailored Test Design Study 2013: Summary Research Report*.

Bejar, I. I. (2014). Past and future of multistage testing in educational reform. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3-20). Boca Raton, FL: CRC Press.

Campbell, C., Clinton, J., Fullan, M., Hargreaves, A., James, C., & Longboat, K. (2018). *Ontario: A Learning Province.* Toronto: Queen's Printer for Ontario.

EQAO. (2020). *Modernization and Digitization of EQAO Assessments.* Toronto: Queen's Printer for Ontario.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

Lifelong Achievement Group & Martin, A. J. (2015). *Online NAPLAN Testing and Student Motivation: Exploring Adaptive and Fixed Test Formats*.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than fixed-item test? *Applied Psychological Measurement, 41*(7), 495-511.

Luecht, R. M. & Sireci, S. G. (2011). *A review of models for computer-based testing.* Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27-45.

Organisation for Economic Co-operation and Development. (2018). *Education and Skills Online: Technical Documentation.* Retrieved from http://www.oecd.org/skills/ESonline-assessment/assessmentdesign/technicaldocumentation/ESOnline_Technical_Doc_Ch1-8_Oct2016.pdf

Organisation for Economic Co-operation and Development. (2019a). *Skills Matter: Additional Results from the Survey of Adult Skills.* Retrieved from https://doi.org/10.1787/1f029d8f-en.

Organisation for Economic Co-operation and Development. (2019b). *PISA 2018 Results (Volume I): What Students Know and Can Do*. Retrieved from https://doi.org/10.1787/5f07c754-en.

Robin, F., Steffen, M, & Liang, L. (2014). The multistage test implementation of the GRE revised General Test. In D. Yan, A. A., von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325-341). CPR Press: Boca Raton, FL.

Stone, E., & Davey, T. (2011). *Computer-Adaptive Testing for Students with Disabilities: A Review of the Literature* (Report No. RR-11-32).

Yamamoto, K., Khorramdel, L. & Shin, H-J. (2018a). Introducing Multistage Adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling, 60*(3), 347-368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018b). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 16-27.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3-20). Boca Raton, FL: CRC Press.

Education Quality and Accountability Office
2 Carlton Street, Suite 1200, Toronto ON M5B 2M9
Telephone: 1-888-327-7377 | Website: www.eqao.com

DmsCATlr_ne_1120