Computerized Adaptive Testing as a means for Mathematics Assessment

Leung Chi Keung, Eddie
Department of Mathematics, Hong Kong Institute of Education

Introduction

Using information technology (IT) in education is a hot issue in Hong Kong recently. Many people connect using IT in education with computer-assisted instruction (CAI), computer-assisted learning (CAL) and seeking information from the internet. Indeed, using IT may also bring in other advantages such as a continuous and dynamic approach to testing.

Assessment is an essential part of local Primary and Secondary Mathematics teaching and learning. It provides useful information to a variety of users for a variety of purposes. The various purposes for which users require information from assessment can be summarized as formative, summative, evaluative, predictive, comparative and selective (Education Commission, 1990, p.64). There are many methods of assessments such as oral questioning, written work, portfolio, observation and use of tests. All these methods can be used formally or informally. Each method has its own advantages as well as limitations. One should choose the right form according to the specific objectives of the assessment.

In Hong Kong, the class sizes of primary and secondary schools are generally around 40. Assessment of student's mathematics knowledge based on observations on individual performance would be difficult and unreliable in such large classes. This makes written tests the most popular format of assessment used by the mathematics teachers for practical reasons such as validity and reliability. If a student succeeds on a written test, it would be fairly confident to believe that the student has possessed the skills and/or knowledge which the test aims to measure. The weakness of students who fail in the test would be reflected in their work and can be further identified by in-depth questioning. With the test results, the teacher can evaluate his or her teaching and devise planning for later stages.

What is CAT?

No one can deny that written tests are indispensable component of summative assessment, though this format cannot meet all the six Mathematics Assessment Standards proposed by the National Council for Teachers of Mathematics (1995). Yet most written tests meet the first and the most fundamental standard: "Assessment should reflect the mathematics that all students need to know and be able to do" (NCTM, 1995, 11). With the advances in computing technology and psychometrics, most paper and pencil (P&P) tests can be transformed into the format of computerized adaptive testing (CAT; see, e.g. Lord, 1980; Weiss, 1976; Wainer et al., 1990). One of the main advantages of CAT over P&P is that it enables more efficient and precise trait estimation (Owen, 1975; Weiss, 1982; Wainer, 1990). It must be emphasized that computerized adaptive testing is different from computerized administrated testing which usually refers to a mechanism that randomly select a test item or a subtest from a pool of items with regardless of the ability of the testee (e.g. see Beevers et al., 1995). In contrast, a CAT system *adaptively selects* an item according to the estimate of the ability of testee based on his or her responses to previous items. In other words, CAT is a dynamic system that can provide tailor-made tests for individuals. If one gets an item correct, then the next item would be more difficult. With the same token, the next item would be easier if one gets a wrong answer for the one right on the screen. Because of the adaptive nature of CAT, examinees always face items that closely match their own individually estimated ability. Consequently, individual test forms of CAT should be shorter as there are less inappropriate items for each individual. At the end of the test, no one is likely to get all answers wrong and scores zero mark; the less competent students would find some items that they could solve and hence retain their interest in the subject. Neither anyone is likely to get all answers correct and scores full mark; thus even the top students understand that there are rooms for improvement. It may happen that two testees get the same number of items correct, however they may have different scores that depend on the parameters like difficulty, discrimination and guessing of the items (Lord, 1980; Hambleton & Swaminathan, 1985). All the conversions of score on the same continuum are done by the statistical procedures of the system.

There are many technical issues on how to build, maintain and use a CAT system. Interested readers may refer to Wainer et al. (1990) to start with.

Is it the right time to establish CAT?

CAT systems have been successfully developed overseas in different areas such as French language proficiency (Burston, Harfouch & Monville-Burston, 1995), Japanese language proficiency (Brown & Iwashita, 1996), and ESL reading comprehension (Young, Shermis, Brutten & Perkins, 1996). Other large scale CAT systems such as Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT) and National Council Licensure Examination for Nurses are run in the United States (see Chang & Ying, 1997). The two new directions in the local education policies: target-oriented curriculum (TOC) and using information technology in education, have pushed the author to advocate the establishment of computerized adaptive testing for upper primary to junior secondary mathematics.

As suggested by Clark et al. (1994, p. 11): "the TOC initiative would need to devise forms of assessment designed to measure students' learning against criteria embodied in standards, in order to measure what they were able to do and how well they could do it, and to highlight their strengths and weaknesses in order to inform future teaching and learning". Within the framework of TOC, attainment targets for individual topics of Mathematics at various stages will be laid down. More and more test items measuring students' achievement in these targets will be constructed with the effort of ED officials, teachers, publishers, tutorial centers, educational researchers and so on. Those items that satisfy the 3-parameter item response theory (IRT) models (Lord, 1980; Hambleton & Swaminathan, 1985) and passes the sensitivity test (Flaugher, 1990), can be calibrated and gathered to form a rich item bank that covers a wide range of abilities.

As the government is planning to equip schools with more computers and establish an intranet among schools, the CAT system can be developed and administered with the support and coordination by the government. If schools have sufficient resources and support, they may download the relevant item banks and establish their own CAT systems.

Is it worthwhile?

At least four parties: students, teachers, school administrators and officials of education department, will benefit from a well-developed CAT system coordinated by the government. Firstly, the pressure on teachers could be partially released. In the study of Leung, Man & Kong (1998), it is found that Mathematics teachers working in TOC schools have more pressure in setting test and examination papers. Teachers are

not sure whether the test set by them can cover all the attainment targets. If there is a CAT system containing the item bank for measuring the skills and concepts concerning a certain topic (for example, fractions in the Stage 2 of TOC), then the teachers may simply help the pupils to activate the CAT system and let the computer do the rest. This would save teachers a lot of time on the preparation of tests and marking of scripts when performing summative assessments. Teachers can then utilize their energy on planning and preparation of other kinds of assessments, purposeful and meaningful learning activities for their pupils. In addition, all students experience the same set of examination questions in a formal examination, some of them may try to cheat or look over the shoulders. These kinds of misbehaviour may arouse discipline problems that teachers have to tackle. But if the examination is in CAT form, any two neighbouring students are unlikely to face the same set of questions, thus reducing the number of student misbehaviours.

Secondly, the students can have objective assessments. The computers recognize neither the names nor the faces of individual students, so no marks will be added or deducted by impression. Besides, the CAT system can cater for individual differences by delivering tailor-made tests. A competent student will not face too many simple questions that may lead to an underestimate of his or her proficiency if careless mistakes are made. On the other hand, less able students will not face too many questions that are difficult to them. Thus, their confidence and interests in the subject would not be seriously hampered. In addition, students would spend less time on individual test as the test generated by a CAT system would generally be about half of the length of its P&P counterpart. Furthermore, a well-developed system would be able to immediately issue individual reports on the performance of the testees. Hence, the strengths and weaknesses of the students would be identified. If a student has unsatisfactory result in the test, he or she can re-take the test at the time that he or she feels confident after revision or remedial teaching. Once the students are familiar with the testing procedures and environment, teachers need not accompany the students in their second and subsequent trials. Students just need to book the computers and inform the teachers in advance. If the item bank is rich, the test items at various attempts are very unlikely the same. Since the students themselves can determine the dates for subsequent attempts after failure, their motivation of learning may be stronger when their sense of ownership of learning increases.

Thirdly, the school administrators can have a clearer picture on the achievements of their students and the teaching effectiveness of their staff since the teachers do not know in advance what test forms will be generated by a CAT system. It is not an unusual practice that teachers

give tips or similar quiz to their students once they know the test questions. There are several reasons for this kind of action: some teachers worry that the principals may invite them to explain if the performances of their classes are below average; some feel a higher sense of satisfaction if their classes apparently perform better than other classes; and some try to cover up the facts that they do not teach properly and so on. The information gathered from objective data would help the administrators to make better decisions and adjustments in school policies. Besides, schools will be aware that their achievements will be in comparison with other schools in an objective system. Then, they will develop clear and coherent educational goals and utilize their resources wisely to achieve the goals.

Last but not least, the Education Department can obtain more objective data by replacing some of the Hong Kong Attainment Tests in Mathematics with CAT systems. The delivery of the test and the marking can be done by the computers directly. On one hand, it saves teachers lots of time on marking and on the other hand, schools and teachers have less improper ways to boasting up their students' achievement. The officials can use the information to monitor the general standards of students' achievement and assess the effectiveness of new educational initiatives. This would lead to a better decision on resources allocation and future directions.

Conclusion

Computerized adaptive testing is one of the many methods of mathematics assessment. It may not be able to measure all kinds of intellectual ability of students such as communicative skills. However, it can help answer questions commonly asked by various parties such as "How good is my mathematics compared with the same age group?", "How good is my child at mathematics?" and "Are there any differences between students' mathematics achievement this year and the last year?". It provides objective measurement on students' knowledge in mathematics concepts and skills. Its implementation will certainly reduce the workload of mathematics teachers who may then spend more time on the planning and preparation of purposeful and meaningful learning activities for their students. The information gathered from objective data may lead to better decisions and adjustments on teaching-learning cycles, setting of educational goals and targets, resources allocation and professional support.

With the two recent directions in education policy: the implementation of TOC and using IT in education, it is the right time to start establishing a CAT system. CAT can replace many P&P tests of

Mathematics. There are many technical issues involved in developing a CAT system. So it may take several years to put the first CAT in mathematics for the public use even if we start planning it now.

References

Beevers, C.E., McGuire, G.R., Stirling, G. & wild, D.G. (1995). Mathematical Ability Assessed by Computer. *Computers & Education, 25*(3), 123-132.

Brown, A. & Iwashita, N. (1996). Language Background and Item Difficulty: The Development of a Computer-adaptive Test of Japanese. *System, 24*(2), 199-206.

Bruston, J, Harfouch, J. & Monville-Burston, M. (1995). The French CAT: An Assessment of its Empirical Validity. *Australian Review of Applied Linguistics, 18*(1), 52-68.

Chang, H., & Ying, Z. (1997, June). *Multi-stage CAT with stratified design.* Paper presented at the Annual Meeting of Psychometric Society, Goltlinsberg, NT.

Clark, J., Scarino A. & Brownell J. (1994). *Improving the Quality of Learning.* Hong Kong: Hong Kong Bank Language Development Fund/Institute of Language in Education.

Education Commission (1990). *Report No. 4: The curriculum and behavioural problems in School.* Hong Kong: Hong Kong Government.

Flaugher, R. (1990). Item Pools. In H. Wainer et al. (Ed.), *Computerized Adaptive Testing: A Primer.* New Jersey: Lawrence Erlabum Associates, Publishers.

Hambleton, R.& Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer Nijhoff.

Leung, C.K., Man, Y.K. & Kong, S.C. (1998, Nov.). *The Impact of the Implementation of Target-Oriented Curriculum (TOC) on Primary Mathematics Education: Teachers' Perspective.* Paper presented at the 15th Annual Conference of Hong Kong Educational Research Association. Hong Kong Baptist University.

Lord, M.F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

National Council of Teachers of Mathematics (1995). *Assessment Standards for School Mathematics.* Reston, Va.: National Council of Teachers of Mathematics.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Wainer, H. et al. (Ed., 1990), *Computerized Adaptive Testing: A Primer*. Hillsdale, New Jersey: Lawrence Erlabum Associates, Publishers.

Weiss, D.J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C.L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing*, 24-35. Washington, DC: United States Civil Service Commission.

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

Young, R., Shermis, M.D., Brutten, S.R. & Perkins, K. (1996). From Conventional to Computer-adaptive Testing of ESL Reading Comprehension. *System, 24*(1), 23-40.