THE DEVELOPMENT OF A COMPUTER ASSISTED DESIGN, ANALYSIS AND TESTING SYSTEM FOR ANALYSING STUDENTS' PERFORMANCE

Qingping He and Peter Tymms

The Development of A Computer Assisted Design, Analysis and Testing System for Analysing Students' Performance

Qingping He and Peter Tymms
CEM Centre, University of Durham,
Durham, DH1 3UZ, UK
Qingping.He@cem.dur.ac.u
P.B.Tymms@cem.dur.ac.uk

Abstract

Recent years have seen increased application of Computer Assisted Assessment (CAA) in education at various levels, and a variety of computer software systems have been developed for use in computer-based testing and analysis. However, many existing system are primarily designed to provide objective assessment of students and analysis of test items. The present study presents the development of a Computer Assisted Design, Analysis and Testing System (CADATS) that can be used by primary and secondary schools and other test organisations to undertake computer assisted assessment. The system incorporates an Item Response Theory (IRT) model – the Rasch model to facilitate the administration of IRT – based tests on computers and the analysis of test items and students' performance using modern test theories. Specifically, the system has been created to design and undertake computer-based tests (CBTs), including the Computer Adaptive Tests (CATs), and to undertake diagnostic analysis on students' performance at both individual and school levels in order to identify curriculum areas where students are under performing.

Keywords: Computer Assisted Assessment, Item Response Theory, Diagnostic Analysis, Students' Performance, Item Analysis.

Introduction

Recent years have seen increased application of Computer Assisted Assessment in education at various levels (e.g. Buchanan, 2000; Lin et al., 2001; Tymms, 2001; Gardner et al, 2002; Tsai and Chou, 2002; Tzuriel and Shamir, 2002; Ashton et al., 2003; Russell et al., 2003; Tymms et al., 2004; Wang et al., 2004). CAA offers a number of advantages over traditional paper-and-pencil based assessment, including the provision of timely information that can be used for diagnosing areas where a student has difficulties. A variety of computer software systems have been developed for use in computer-based test and analysis. However, existing systems are generally expensive and require skilled expertise to operate. In view of the limited IT

and other resources in primary and secondary education organisations, there is a need to develop assessment systems which are easy to use and can provide the necessary information that can help teachers to improve students' learning ability and performance.

The present study presents the development of a Computer Assisted Design, Analysis and Testing System (CADATS) that can be used by primary and secondary schools and other test organisations to undertake computer assisted assessment. The system could also be used by higher and further education organisations. The system will be easy to use. The system incorporates an Item Response Theory model – the Rasch model (see Rasch, 1960; Wright and Stone, 1979; Masters and Keeves, 1999 for detailed description of this model) in order to facilitate the administration of IRT – based tests on computers and the analysis of test items and students' performance using modern test theories (see, for example, Hambleton and Swaminathan, 1983; Masters and Keeves, 1999; Hambleton, 2000; Tymms, 2001; Wang and Kolen, 2001; Brown and Iwashita, 2002; Tonidande et al., 2002; Lilley and Barker 2003). The Rasch model has been used in the system due to its simplicity and unique features in analysing test data. This model measures the item difficulty and student ability on the same linear continuum, and can therefore provide objective measurement (see Wright and Stone, 1979).

CADATS has been specifically created to design and undertake computer-based tests, including the computer adaptive tests, and to undertake diagnostic analysis on students' performance at both individual and school levels in order to identify curriculum areas where students are under performing. The system involves the use of Extensible Markup Language (XML) and the Macromedia Flash technology, and this provides easy-to-understand and intuitive graphic representation of test results. Some preliminary results from a case study involving students from schools in Hong Kong who took the Year 7 Baseline Test in October 2003 provided by the Curriculum, Evaluation and Management (CEM) Centre at the University of Durham, UK, obtained using the system presented in this paper will be reported.

System Specifications

CADATS can be used to perform the following main tasks:

- Creating items and constructing item banks.
- Designing tests effectively by selecting items from an item bank. Both Classic and IRT – based tests (including Computer Adaptive Tests) can be designed using the system. Once there is a calibrated item bank (i.e. once properties such as item difficulty are know for each item), an IRT – based test can be designed which targets the specific ability level of the intended students.

- Conducting designed tests (both Classic and IRT – based tests, including CATs) on computers.

- Analysing test results. In addition to providing basic test statistics, the system will also be able to undertake detailed diagnostic analysis on student's performance at both individual and class/school levels. The system has the ability to generate information on the performance of students and test items that can be easily used by teachers to identify curriculum areas where students are under performing.

- Undertaking test item analysis using the Rasch model. This will enable the establishment of large calibrated item banks for a school in the long term. Such item banks will surely be very useful for monitoring students' performance progress by administrating tests composed of items selected from the item banks.

System Design and Implementation

CADATS has been designed to contain two major components: A, The Item and Bank Building, Test Designing and Results Analysis Component; and B, The Test Delivery Component. Both components are developed as a Microsoft Excel Visual Basic for Application (VBA) project embedded with the Macromedia Flash AxtiveX control.

- The Item and Bank Building, Test Designing and Results Analysis Excel workbook, embedded with the Flash AxtiveX control, provides the Graphical User Interfaces (GUIs) which interact with the worksheets in the workbook and the files on the computer's hard drive. This is used to write items and build item banks, design tests and undertake analysis of test results. Items are formulated as Flash movies or JPEG files stored on the computer's hard drive. Tests designed by the system are formulated as XML objects which can be loaded into the Test Delivery Component for conducting tests on computers.

- The Test Delivery Excel workbook is used to undertake computer-based tests and collect students' response data. This workbook is also embedded with the Flash AxtiveX control. The Flash ActiveX control acts as the GUI, which loads a test XML object and parse it in order to load individual questions for display during a test. The workbook can be either copied onto a shared network drive for multiple access by students to take a test or copied onto individual computers as a standalone testing system. Responses from students are exported as XML objects stored on the computer's hard drive which can be loaded into the Item and Bank Building, Test Designing and Results Analysis Component for analysis.

Analysis of Hong Kong Year 7 Baseline Maths Test Results Using CADATS – A Case Study

As one of the world's largest educational research institutions, the CEM Centre at the University of Durham has been working with both primary and secondary schools in the UK and abroad through the provision of a variety of monitoring projects to schools. The CEM Centre's value-added approach, which involves the analysis of data collected from baseline tests and other measures administered by the CEM Centre and school outcome measures, provides fair comparison between the progress made by pupils in a particular school and the large sample of others in a CEM Project. This provides an important basis for schools to undertake self-evaluation and management (see Fitz-Gibbon, 1997). The CEM Centre's Middle Years Information System (MidYIS) Project provides value-added analysis to schools with students in years 7, 8 and 9 (ages from 11 to 14). From 2002, over 20 schools in Hong Kong have been participating some of the CEM Centre's Projects. Some of these schools have been involved in the MidYIS Project and students from 5 of the schools have been taking the English version of the MidYIS Year 7 Baseline Test for Hong Kong.

The MidYIS Year 7 Baseline Test for Hong Kong consists of 9 separate sections: English Vocabulary; Maths; Perceptual Speed and Accuracy; Proof Reading; Cross-sections; Block-counting; Picture Sequences, Chinese Vocabulary and Chinese Reading Comprehension. The test is computer - based and delivered over the Internet. Results from the test for the schools are reported for five areas: Verbal (English Vocabulary), Maths, Non-Verbal (Cross-Sections, Block Counting and Perceptual reasoning), Skills (Proof-reading, and Perceptual Speed and Accuracy), and Chinese (Vocabulary and Reading Comprehension), and an overall test score. These results are standardised against all Hong Kong schools participating in this scheme to have a mean score of 100 and a standard deviation of 15 for each of the five score areas and the overall test score in order for easy comparison of test scores between the schools to be undertaken. The present investigation will focus on analysing the Maths test results for one of the 5 schools taking the baseline test using CADATS. The maths section of the Hong Kong MidYIS Year 7 Baseline Test contains questions from a variety of subject areas and requires substantial skills to answer. These include questions of numerical, algebra, geometry, problem solving, common knowledge (such as reading a watch), and others, in order to assess the maths proficiency of the students. There were 994 students in all from the 5 schools taking the baseline test in the academic year 2003/2004.

The reliability of the maths test was estimated to be 0.82. An analysis using the Rasch model built into the system on the 65 maths items were undertaken, and difficulty values vary from –4.07 logits to 3.91 logits (see Wright and Stone, 1979, for the definition of logit). The model Infit values for all items are between 0.80 and 1.20, suggesting that the items meet the requirement of the Rasch model (see Bond and Fox, 2001). Figure 1 depicts the difficulty distribution of the maths items and the distribution of the estimated ability for the students from the 5 schools. For the ability

distribution, the shaded area represents all the students from the 5 schools. The average ability for the students from these schools is estimated to be 1.11 logits (equivalent to 100 on the standardised scale). This average ability is substantially higher than the average difficulty of 0 logits for the items, indicating that the maths test was relatively easy for the students. CADATS makes it easy to identify questions requiring specific attention, such as those which are extremely easy or extremely difficult. Once the items have been analysed using the Rasch model and their difficulty distribution is displayed graphically, clicking a question number on the difficulty distribution will display the content of the associated question. For example, the bottom diagram in Figure 1 illustrates the content of the most difficult question - Question Number 44 of the maths test.

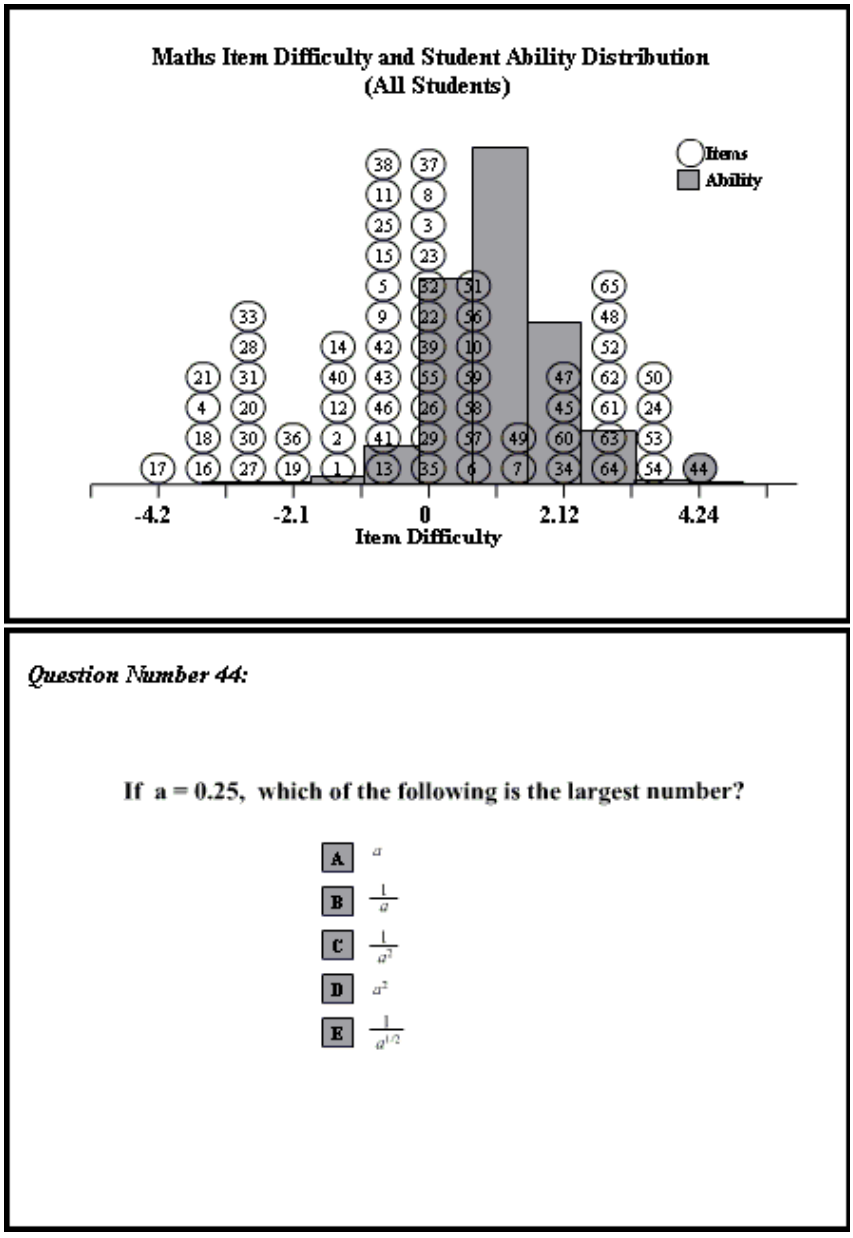

Figure 1: The item difficulty and student ability distributions (top) and the content of the most difficult question-Question Number 44 in the test (bottom).

As an example to demonstrate the use of CASATS to undertake diagnostic analysis on students' performance, one of the schools, referred to as School A in this investigation, was used in this case study. There were 198 students from the school taking the online baseline test. The average maths raw score for the school is 41.9 with a deviation of 5.8. The standardised score for this school is 99, with a standard deviation of 13.5 and an error estimated to be 1.1. Compared with a mean of 100 and deviation of 15 for all the 5 schools involved, the mean maths score for the school is slightly below the average of all schools, but not significantly different. Both the raw score and the estimated ability using the Rasch model have been used to present test results for each student. Figure 2 compares the distribution of student ability for the school with the distribution of item difficulty. The shaded area for the student ability distribution represents all the students from School A. Compared with Figure 1, the peak of the students' ability distribution for this school is more pronounced than that for all the students from the 5 schools. The standard deviation of the estimated ability for the school is significantly smaller than that of the estimated ability for all the 5 schools. The average student ability for the entire school is estimated to be 1.06 logits (equivalent to 99 on the standardised scale). This average ability is slightly lower than but not significantly different from the average student ability of 1.11 logits (100 on the standardised scale) estimated for all the 5 schools.

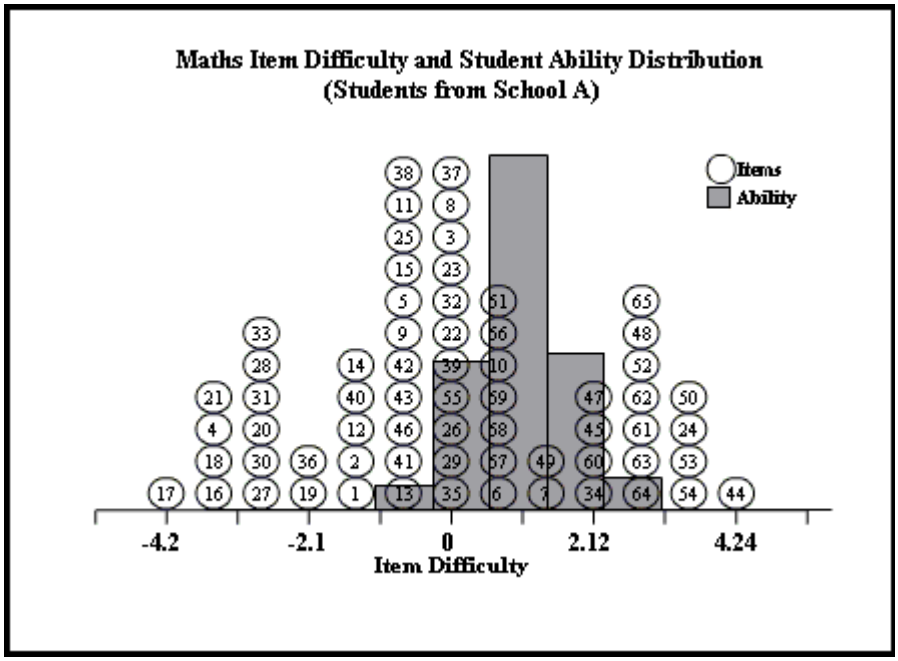

Figure 2: The item difficulty distribution and ability distribution for students from School A.

At the individual student level, detailed analysis can be undertaken using CADATS for a student at the item-by-item basis. Comparison with the performance of other students within the same school, or with a norm performance such as the national average performance or the average performance of a group of schools can also be carried out. CADATS records the response pattern for each student and presents the results graphically. Figure 3 shows the response pattern of Pupil 119 to the 65 maths items, providing a very clear picture of questions answered correctly (e.g. questions 17, 21, 4, 18, and etc), questions answered incorrectly (e.g. questions 27, 11, 35, 6, and etc), and questions skipped (e.g. questions 9, 8, 10 and 7). Figure 3 represents a typical response pattern: easier questions were generally answered correctly, while harder questions were answered incorrectly or skipped. However, it is noticed that some easy questions were also answered incorrectly, while some harder questions were answered correctly, reflecting the specific knowledge associated with each individual student. The easier question 27 was answered incorrectly by Pupil 119, while the medium difficulty question 9 was skipped. However, the more difficulty question 48 was answered correctly.

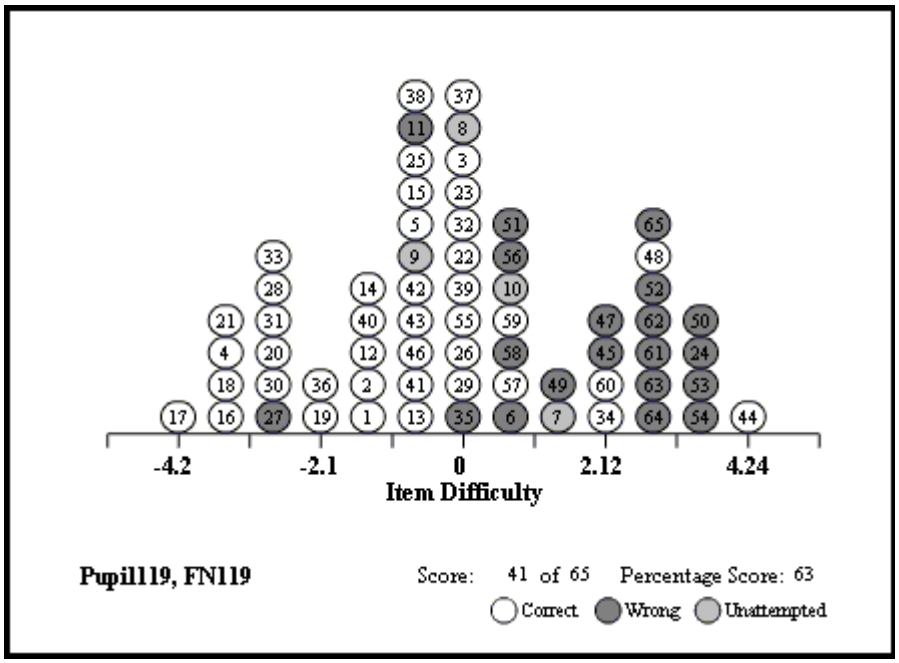

Figure 3: The response pattern of Pupil 119 to the maths items.

Figure 4 further shows the response pattern from Pupil 119 for the 65 maths items in relation to the average item percentage scores for all the students from the 5 schools. Again this diagram makes it easy to identify items which a student did not perform well in relation to the average performance of all the students taking the test.

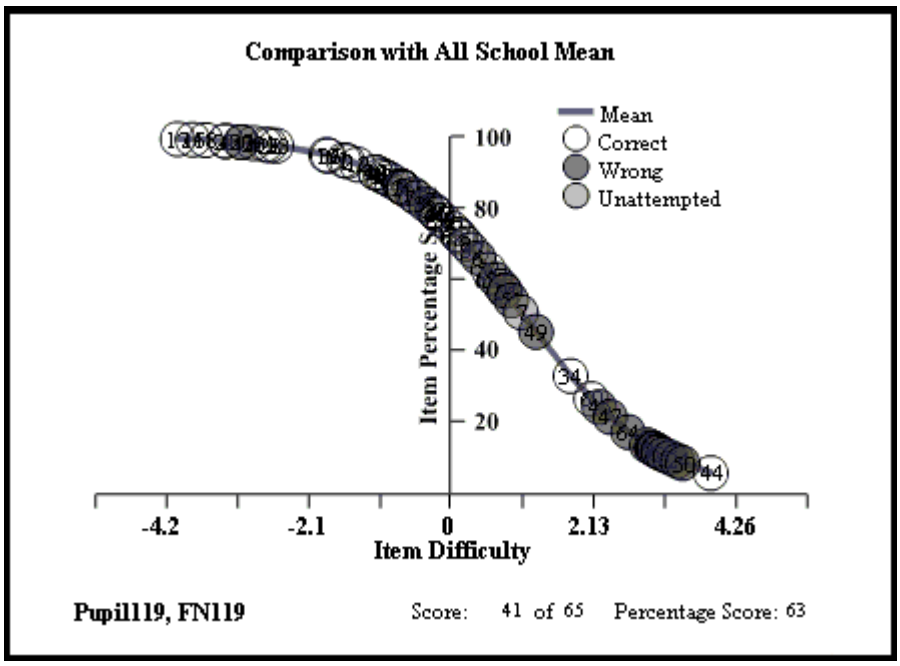

Figure 4: The response pattern of Pupil 119 in relation to the item percentage scores for students from all 5 schools.

At school level, CADATS can be used to compare students' performance within a school or with the average performance of students from a group of schools. For example, Figure 5 compares the item percentage scores for students from School A and those for all the students from the 5 schools which is represented by the solid line. This diagram gives a clear indication of the performance on the maths test for the entire school in relation to the average performance of all the schools. If an item percentage score for the overall school is on the line, then the performance of the school as a whole on that item can be assumed to be close to the average performance of all the schools. If the item percentage score is above the line, then the performance of the school on that item can be assumed to be above the average performance of all the schools. If, however, the item percentage score is below the line, the performance of the school on that item can be assumed to be below the average performance of all the schools. It is clear from Figure 5 that for the majority of the maths items, the item percentage scores are on or close to the line, indicating that School A performed as the average of the 5 schools on those items. There are however a few items (e.g. Question Numbers 39, 5, 49 and 34), the item percentage scores for the school are significantly below the line, indicating that for those items the school as a whole performed less well as the average performance of all the schools. The bottom diagram of Figure 5 shows the content of item number 34, on which the school under performed. This item is a number series question which calculates the value for the next term as the sum of its preceding two terms. A school should pay particular attention to and enhance teaching in the subject areas associated with the items which it under-performed.

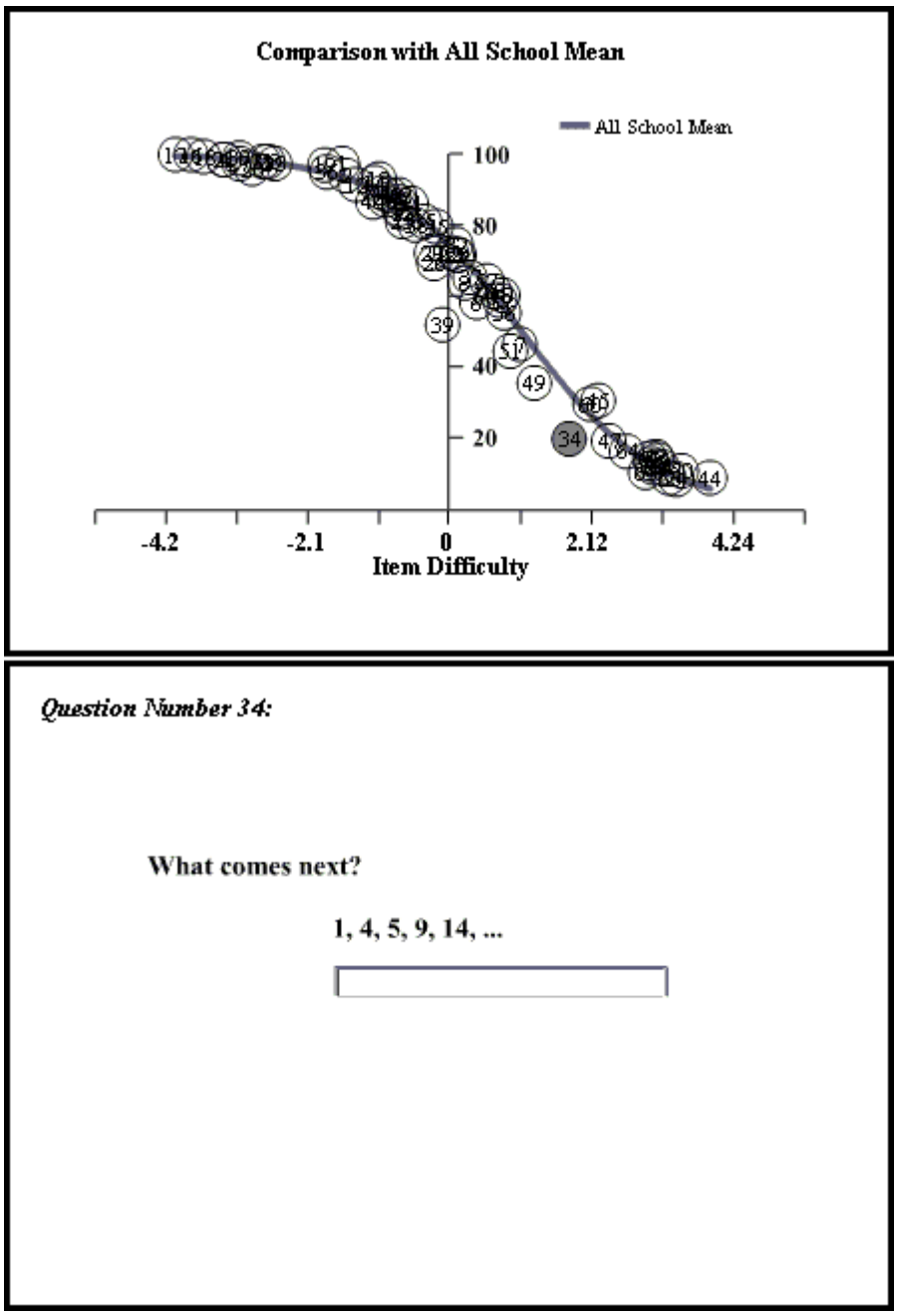

Figure 5: The percentage item score distribution for students from School A in relation to that for students from all the 5 schools (top) and the content of item 34 (bottom).

Conclusions

With the rapid development of computer hardware and software and the investment in Information and Communication Technologies (ICT), the use of computer based testing and computer assisted assessment in education organisations has increased substantially in recent years. Computer assisted assessment have many advantages over conventional paper-and-pencil based testing and assessment. In addition to making decisions on students

based on test scores, CAA provides other useful information that paper-and-pencil based testing and assessment lack. Many existing systems are expensive and in many cases can only provide objective assessment of students and analysis of test items. CADATS presented in this paper provides a variety of functions that can be used to generate diagnostic information on the performance of students and test items, which can be easily used by teachers to identify curriculum areas where students are under performing either individually or in the school as a whole. Such information provides a basis for a school to help individual students according to their specific needs and to target areas where the whole school is under performing. Further work will involve refining the system to extend its capability in analysing test items. A full evaluation of the system by school teachers will also be undertaken and suggestions from the teachers will be taken into account when refining the system.

Acknowledgements:

The authors would like to thank Chris Wheadon, Brian Henderson, Robert Coe and Robert Clark for their critical comments and suggestions on the system reported in this paper and SM Tsui for organising the online test in Hong Kong.

References

Ashton, H.S., D.K. Scholfiled and S.C. Woodger (2003) Piloting summative Web assessment in secondary education. *2003 CAA Conference Proceedings*: 19-29. University of Loughborough, UK.

Bond, T. and C. Fox (2001) *Applying The Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah NJ: Lawrence Erlbaum Assoc.

Brown A. and N. Iwashita (2002) Language background and item difficulty: the development of a computer-adaptive test of Japanese. *System* 24: 199-206.

Buchanan, T. (2000) The efficacy of a world-wide web mediated formative assessment. *Journal of Computer Assisted Learning* 16: 193-200.

Fitz-Gibbon, C. T. (1997) The Value Added National Project: Final Report *Feasibility studies for a national system of Value Added indicator*. School Curriculum and Assessment Authority, UK

Gardner, L., D. Sheridan and D. White (2002) A Web-based learning and assessment system to support flexible education. *Journal of Computer Assisted Learning* 18: 125-136

Hambleton R. (2000) Emergence of item response modelling in instrument development and data Analysis. *Med Care* 38 (Suppl II): II60-II69.

Hambleton R. and H. Swaminathan (1983) *Item response theory: Principles and applications*. The Netherlands: Kluwer-Nijoff.

Lilley, M. and T. Barker (2003) An evaluation of a Computer Adaptive Test in a UK university context. *2003 CAA Conference Proceedings*: 171-182. University of Loughborough, UK.

Lin, S.S.J., E.Z.F. Liu and S.M. Yuan (2001) Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning* 17: 420-432.

Masters G. and J. Keeves (1999) *Advances in measurement in educational research and assessment*. The Netherlands: Elsevier Science.

Rasch G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.

Russell, M., A. Goldberg and K. O'Connor (2003) Computer-based testing and validity: a look back into the future. *Assessment in Education: Principles, Policy and Practice* 10: 279-293.

Tonidandel, S., M.A. Quiñones and A.A. Adams (2002) Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *The Journal of Applied Psychology* 87: 320-332.

Tsai, C.C. and C. Chou (2002) Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning* 18: 157-165.

Tymms, P.B. (2001) The development of a computer-adaptive assessment in the early years. *Educational and Child Psychology* 18: 20-30.

Tymms, P.B., C. Merrell, and P. Jones (2004) Using baseline assessment data to make international comparisons. *British Educational Research Journal* (in press).

Tzuriel, D. and A. Shamir (2002) The effects of mediation in computer assisted dynamic assessment. *Journal of Computer Assisted Learning* 18: 21-32.

Wang T. and M.J. Kolen (2001) Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement* 38: 19-49.

Wang, T.H., H. Wang, W.L. Wang, S.C. Huang and S.Y. Chen (2004) Web-based Assessment and Test Analyses (WATA) system: development and evaluation. *Journal of Computer Assisted Learning* 20: 59-71.

Wright, B. D. and M.H. Stone (1979) *Best test design.* Chicago, IL: MESA Press.