Southeast Asian Mathematics Education Journal
2012, Vol. 2 No. 2, 61 - 73

Introducing Computer Adaptive Testing to a Cohort of Mathematics Teachers: The Case of Concerto

S. Kanageswari Suppiah Shanmugam
SEAMEO RECSAM
<kanageswari@recsam.edu.my>

Leong Chee Kin
SEAMEO RECSAM
<ckleong@recsam.edu.my>

Abstract

This article describes a study that explores on-line assessment, with the objectives to identify features that support or impede the usability of Concerto, an on-line adaptive testing software that was developed by the Psychometrics Centre of the University of Cambridge. We report on the analysis of data collected during a one-month in-service programme organised for secondary teachers and teacher educators from the Southeast Asian Minister of Education Organisation (SEAMEO) region. The study identifies the challenges the participants encountered during a one-day workshop and evaluates the difficulties of adopting Concerto to create a simple and an adaptive on-line mathematics test. While the small study limits the possibility of applicability for other samples, yet the findings of the study illustrate the complexity of using the Concerto's features and the commonly occurring difficulties, providing the basis for the development of some new workshop materials that will contribute to the improvement of introductory Concerto workshops that will be conducted in the future.

Keywords: Concerto; computer adaptive testing (CAT); on-line adaptive testing; assessment; testing

Introduction

Technology has long been incorporated in education as a tool that enhances teaching, learning and assessing strategies (Barker, 1974). Technology in testing has evolved from a simple gadget like a blackboard to the present sophisticated Smartphone, tablets and applets, while along the way witnessing gadgets like handhelds scientific calculators and the widely used computer. Computer based technology is possibly essential for sustainability in education. From computer assisted learning to the current computer adaptive testing (CAT), computers have continued to live up to its name as a powerful educational tool that is effectively used as a medium to assess students, apart from its essential role as a teaching and learning aid (Barker, 1974). With the Internet, CAT is slowly but steadily moving into the classrooms, making assessments efficient, accessible, economical and borderless.

Adaptive Testing

The first recorded adaptive test was created by Alfred Binet in 1905 which became the first intelligence test. He ordered the items according to their difficulty level. Test

administration began by presenting the students with an item that was believed to be a good estimate of the student's ability. If the student succeeded with that item, a moderately more difficult item was presented and if he failed, an easier item was presented. The process would continue until the student was unable to answer a few questions in a row (Linacre, 2000). Stanford-Binet Intelligence Test and Wechsler Intelligence tests are examples of current modern adaptive tests that follow the modest works of Alfred Binet (Partners in Learning, 2012). When computers came along, adaptive tests morphed into computer adaptive testing as computers overtook the role of test administrators and examiners. The computers selected items after making an estimation of the test taker's ability, administered the items and, provided objective and accurate scores without compromising test validity (Linacre, 2000).

Technology in Testing

Computer based technology has become a component in classrooms, especially in computer based testing (CBT) (Russell, 2006). Using computers in testing has proven to increase the effectiveness of test administration as it reduces costs of printing and shipping of test papers, scores accurately and reports the scores with great precision (Linacre, 2000). In addition, computers capitalise on ever-growing number of software to include multimedia and sound that adds colour and vibrancy, redefining the conventional, mundane static tests (Russell, 2006). The benefits of using computers in testing have transcended the objectives of measuring the product (students' work) to measuring the process which makes CBT an invaluable asset in testing. CBT has become mandatory in certain testing situations that require test takers to work in simulated environments such as assessing the skills required by air traffic controller. The use of simulations increased the authenticity of the items presented during the test administration and injected reality into the world of testing. CBT can be effectively used to cater for students with special needs, with the help of certain software like the speech-to-text software (Russell, 2006).

Over the years, CBT has grown from the infancy purpose of computer delivered examinations to computer adaptive testing that can provide accurate estimates of a person's ability by administering a small number of items that are closely matched (Linacre, 2000). With the introduction of the Internet, computer adaptive testing has undergone another uplift that opens a parallel world of on-line assessment.

Computer Adaptive Testing

Computer adaptive testing has gained recognition as a test that is tailored to the test taker's ability. CAT has a reputation as a new wave in assessment, breaking away from the traditional paper-pencil test. The design of CAT allows pre-calibrated items to be administered from an item bank, hence rendering itself as a customised test that continuously re-estimates the true test taker's ability with improved precision of an individual measure. Since the selection of item difficulty is matched to a person's ability, the iterative process will converge to provide an accurate, reliable and valid measure of the person's ability (Linacre, 2000). As items are individually paced, there is no need for the test takers to be presented with irrelevant items that are easy or too difficult, hence reducing and eliminating 'unwanted' behaviour like boredom, fatigue, test anxiety and frustration.

It has been claimed that CAT provides immediate scoring and feedback to the test takers and cuts down testing time by half (Russell, 2006). In addition, administrating equivalently challenging items reduces over-exposure of items and thus increases test security. Cost of administrating tests is cut down as the need for hiring and training invigilators does not arise, which also reduces measurement error. Besides test takers, CAT also benefits test developers and test publishers as experimental items can be simultaneously validated with testing and, test revision is less strenuous since adding or removing items does not contribute to test reliability (Rudner, 1998).

The major setback of on-line CAT is the computer facilities and the quality of the Internet connection which are mandatory components and they must function well for without either one, CAT cannot be administered.

The open source software that was used in this study was Concerto which is an on-line adaptive testing software and will be briefly discussed in the next section.

Concerto: On-line Adaptive Testing Platform

Concerto is an open source, web based, adaptive testing platform that was developed by the Psychometrics Centre of University of Cambridge for creating and running rich dynamic tests. It combines the flexibility of HTML presentation with the computing power of the R language. A table is used to store the data in the form of items, their responses and their difficulty level or any other item parameters depending on the parameter model stipulated in the Item Response Theory (IRT) (Hambleton, Swamination & Rogers, 1991; Lord, 1980; Wainer, 1990). IRT is a statistical framework in which test takers can be described by a set of one or more ability scores that are predictive, through mathematical models, linking actual performance on test items, items statistics, and their abilities (Lilley & Barker, 2002). The

HTML template is used to create an introduction template, an item template and a feedback template and to present it to the test takers. The scores are calculated as the test takers advance into the test.

Description of the Study

The study described in this article was conducted with the intention of identifying the features of Concerto that support or impede the usability of the software as a platform to conduct on-line adaptive assessment. By doing so, the user-friendliness of Concerto as a tool to create on-line adaptive tests could be identified which could help to increase its use in classroom testing. In addition, the challenges that the participants encountered during the Concerto workshop would be identified and solutions sought that would help to alleviate these impediments. At the end of the study, the authors hoped to make some recommendations to improve the user-friendliness of the software itself so that CAT would be more widely used as a mode of assessment. Specifically, there are two main objectives in this study, namely:

1. To identify the features of Concerto that support or impede its usability as an on-line adaptive assessment
2. To provide a basis to revise and develop an improve version of Concerto workshop material for novice users of the software.

Context and Environment

The study reports on a cohort of 11 secondary teachers and one teacher trainer introduced to Concerto as part of their four-week in-service course where one of the authors was the facilitator for the one-day (6 hours) Concerto workshop. The main purpose of the course was to enhance meaningful secondary mathematics learning through interactive technologies. Concerto was one of the technologies introduced in this course. During the workshop, participants used the desktops with Internet connection provided in the workshop venue. This cohort consisted of two teachers each from Philippines and Malaysia, one each from Thailand, Brunei, Vietnam, Cambodia, Myanmar, Laos, Indonesia and Singapore.

The teaching experience of these teachers ranged from three to twenty years with four teachers with an experience of 6 years and two teachers with 18 years of teaching experience. There were three male teachers while the rest were female. There was only one participant with a diploma and one with a Masters while the rest had a bachelor's degree. All the participants taught Mathematics and their academic qualifications were related in the Mathematics field except two who specialised in Economics and Education Management.

All these participants had no prior knowledge of Item Response Theory and R script. According to the participants, this workshop was the first of its kind that provided an exposure to IRT and none of them had ever heard of or applied R script knowledge in their university years, career or in any other course. In addition, they also admitted their lack of knowledge of CAT, even though they claimed that they have heard about it. These limitations were considered while conducting this workshop, and prevent the application of the findings to other samples or workshops.

Instrument Evaluation

The Concerto introductory workshop was evaluated using data derived from observations by the facilitator, conversations between the facilitator and the teachers, their created adaptive on-line test, and a 28-item questionnaire. Questionnaires have long been used to evaluate user interfaces (Root & Draper, 1983). The questionnaire used in this study was adapted from the 'USE Questionnaire' by Lund, as it had been taken through a complete psychometric instrument development process (Lund, 2001). It has four categories. The categories, the items concerned, and it purposes are as shown below.

- Usefulness (Item 1- Item 6): to evaluate the effectiveness of Concerto
- Ease of Use (Item 7- Item 17): to evaluate if it is easy to use Concerto
- Ease of Learning (Item 18 - Item 21): to evaluate if it is easy for the participants to learn the use of Concerto
- Satisfaction (Item 22 - Item 28): to evaluate if the participants are satisfied with Concerto

After the workshop, the participants were asked to indicate their level of agreement on this questionnaire based on a Likert scale ranging from 1 ('Strongly Disagree') to 5 ('Strongly Agree").

Design and Content of the Workshop

The workshop was divided into three sessions each of two hours The first session included knowledge of adaptive testing, computer adaptive testing (CAT), and an overview of Concerto and Item Response Theory (IRT). The various IRT models of one-parameter, two-parameter and three-parameter models were introduced for comparison purposes. However, only the one-parameter model was used to introduce and develop a simple adaptive on-line test in this workshop. The main idea here was to guide the participants to utilise Concerto to deliver a simple yet powerful tool for the development of an adaptive on-line test. Although, this platform relies on three main elements, that is, 1) a HTML presentation

layer, 2) an R Scripting logic, and 3) a SQL database backbone, care was exercised to avoid direct exposure to these elements. Knowledge of these elements was kept to a minimum to avoid information over-load that may cause adverse consequences of adding confusion to these participants who lacked the prior knowledge.

The second session was the guided hands-on activity to design a simple mathematics test consisting of a sample of four items. These items used the addition operation of differing item difficulty. In this session, they learned to create the three HTML templates. The first was the Introduction, where the user could input their name that would be used to customise other HTML templates. The second template consisted of the sample test items with dichotomous response options and the last template would provide the user's score. Then they created an item bank using a table where the teachers input the names of the columns, the test items, assigned values to the user's responses. They also created the corresponding response buttons. The R language and the LTM package were used to generate item parameters.

In the third session, the teachers were required to create an adaptive on-line mathematics test based on the input from the earlier two sessions. They were required to replace the four-item bank with a mock-up item bank consisting of 100 items with only one parameter (generated using the Rasch model) provided by the facilitator.

The workshop was conducted in the English language where some participants relied upon translation for comprehension.

Interpretation of Research Findings and Summary

In this section, some the results of the findings organised around the research questions are highlighted. In this study, the data derived from observations by the facilitator, conversations between the facilitator and the teachers, a 28-item questionnaire and their adaptive on-line test created were examined to evaluate the workshop.

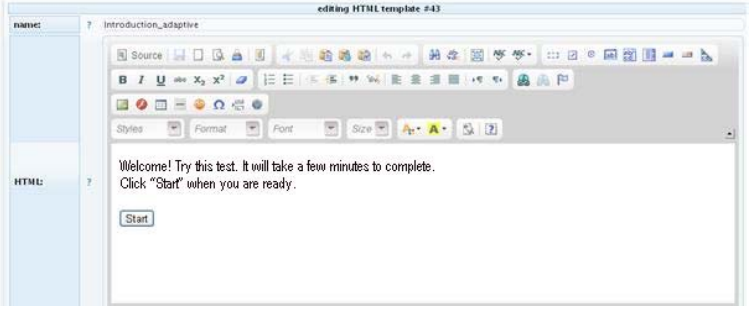

Figure 1. Introduction section of the on-line adaptive test

Simple Test and Adaptive On-Line Mathematics Test

It was observed that in the process of creating the simple test, the participants did not exhibit much problem in developing the HTML introduction, test items and feedback. However, they reported a significant level of frustration and confusion while setting the variable using the R script. Similar problems were encountered when using R code in creating the adaptive on-line mathematics test.

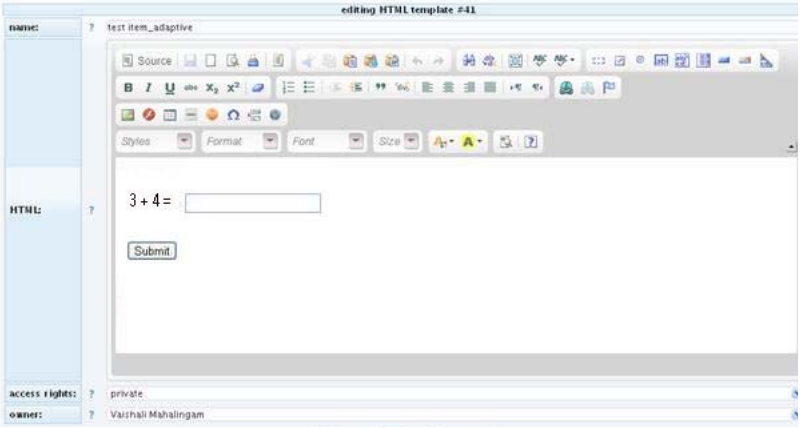

Figure 2. Test Item section of the on-line adaptive test

The teachers found the debugging feature of Concerto very useful and helpful. It was observed that they constantly used this feature before moving to the next section. It helped them to identify the problematic part of the section and eased troubleshooting. However, identifying the problems related to R code was a little difficult to rectify as it required some knowledge of the syntax and structure of R language. However, with further reading and researching on their own on R language as well as guidance from the facilitator, they managed to overcome their challenges.

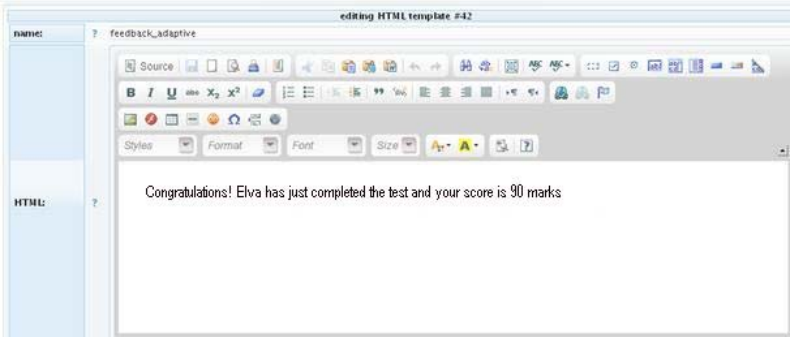

Figure 3. Feedback section of the on-line adaptive test

Out of 12 participants, 80% of them successfully constructed the simple test within the allocated time. This indicated the input session had provided sufficient and relevant basic surface understanding of Concerto as a tool to create an item bank and an adaptive test. This also indicated that the participants mostly had the necessary background knowledge to undertake the task to create an on-line adaptive test utilising a given mock-up item bank. The other 20% of the participants who were unsuccessful in developing the simple test on time managed to complete the task during the break with the help of their peers. Generally, it can be concluded that the features of Concerto did not pose acute difficulties that impeded its use in creating a simple test.

Both the tasks of creating a simple test and an on-line adaptive test using Concerto required the participants to be logged in to the website at http://dev.myiqtest.org/concerto3_demo/cms/. One reason why some of the participants could not complete their task of creating a simple test was due to the slow and interrupted internet connection. The internet speed on that day was slow as indicated by the "Internet Traffic Report" at http://www.internettrafficreport.com/. The average packet loss was reported to be high, at about 80% in the Asia region. The probable reason for the high losses was due to the 8.9 magnitude earthquake that occurred a few days before the workshop which could have damaged the submarine cables. The slow and bad internet connection had pro-longed the time needed to save the test. At times, their data was lost as it could not be saved. Figures 1, 2, 3, and 4 show the sections of: the Introduction, Test Items, Feedback and Item Bank of a participant's on-line adaptive test respectively.

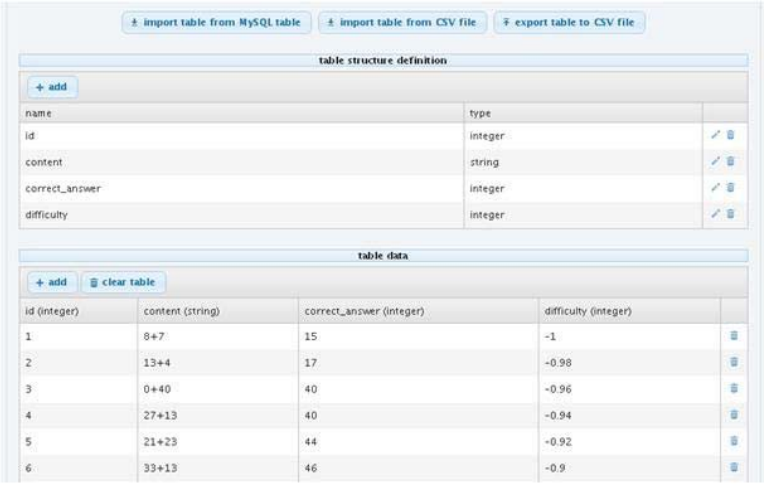

Figure 4. Item bank of the on-line adaptive test

When asked by the facilitator to recommend suggestions to improve the usability of Concerto, only three participants responded. Here are their verbatim responses:

"Interface (using buttons to just click & choose the 4 parts – Intro, feedback, items, test)"
"Guide and reading materials"
"It will be good if there were to be a 'Help' menu"

As to the question posed to them to find out if they needed further training in Concerto, 83% of the participants answered 'yes'. Below are some of their verbatim responses:

"I need more exposure to be more comfortable using it"
"It has new terms and computer languages I need to know"
"I want to know how to add pictures/ more advance keying in of formulas"
"I want to learn and improve the way of preparing test questions"
"I want to have an accurate test for students' ability"
"I don't understand deeply"

The above feedback indicated that generally the features of Concerto were user friendly although the participants needed more time to develop confidence and fully master the program. With assistance most of the participants could use the features of Concerto easily and efficiently to accomplish the tasks of creating a simple test and an on-line adaptive test. However, one participant indicated it would be more practical and easier to access if all the HTML templates were arranged using buttons or maybe tabs.

As R coding posed the greatest challenges to some participants, some instant guide located in the software itself such as a help tab or menu could be of help when they encountered problems. Maybe, a guide on the basic syntax of R language or some programming structure involving the R language could be provided to them.

Usability of Concerto as On-line Adaptive Testing Software

The 28 items that were administered were analysed. Simple analysis involving mean and correlation analysis were computed. Table 1 shows the results on item mean that were obtained.

Table 1
Mean for Participants' responses for each Item (N=12)

Item	Mean	Item	Mean
1	4.17	15	3.50
2	3.83	16	3.36
3	3.92	17	3.42
4	3.82	18	3.42
5	3.70	19	2.92
6	3.55	20	2.82
7	3.75	21	3.33
8	3.33	22	3.50
9	3.08	23	3.50
10	3.08	24	3.00
11	3.75	25	3.67
12	3.58	26	3.58
13	3.25	27	3.33
14	3.50	28	3.75

The results indicate Item 1 obtain the highest mean of 4.17 and Item 20 the lowest mean. With reference to Item 1, more than 83% agreed that Concerto was effective in assessing students. Item 20 refers to the item "It is easy to learn to use it". The low rating was most probably related to the problems they had encountered with the use of the R language and the frustration caused by the disruption to the internet connection. For the category "Usefulness" which consisted of Items 1 to Item 6, almost no participants selected "disagree". This indicated that almost all the participants were of the view that Concerto could be a powerful tool in assessing the students. They reported Concerto as possibly an effective method of assessing students, could also be more productive, take less time, allow better control, and a useful mechanism in assessing students as it could possibly accurately estimate students' true ability.

For the category "Ease of Learning" which is a grouping of Item 18 to Item 21, the mean range was from 2.82 to 3.42 which were among the items lowly rated by the participants. This can be conjectured that the participants faced some challenges to fully grasp the understanding of the introductory material well. As highlighted, the materials should be revised to overcome the difficulties faced with respect to the R language and to ensure that the internet connection is not disrupted and of high speed.

Correlation analyses conducted on the four categories, "Usefulness", "Ease of Use", "Ease of Learning" and "Satisfaction" revealed that there were no significant correlation among these categories. Correlation between "Ease of Use" and "Ease of Learning" is a low Spearman correlation coefficient of 0.235. This is in contradiction with many studies. According to Lund (2001), these two categories should be highly correlated. Obviously the small sample could explain this result, where the participants had rated those items in "Ease of Use" as favourable and items in "Ease of Learning" as unfavourable. Again, it is implied that the problems related to the R language need to be addressed.

Overall, based on the ratings for the items in the "Ease of Use" category and "Satisfaction" category, it can be conjectured that the workshop activities and materials were appropriate for the participants although some changes are needed to overcome the challenges they faced with the R language. They were satisfied with Concerto as an on-line adaptive test platform even though this was the first time they had used it. Specifically, this study indicated that the participants agreed that Concerto could:

1. Help them to be more effective in assessing their students;
2. Be more productive than paper-pencil test;
3. Give them more control in assessing their students;

4. Save time when assessing students;
5. Fulfill their expectations about the purpose of an assessment;
6. Give accurate details about their students' ability;
7. Provide a correct way to assess students as it was based on students' ability;

In addition, the participants also felt that they:

1. Preferred to use Concerto to conduct computer adaptive testing;
2. Could use it after the course;
3. Preferred to use it to design their test;
4. Were satisfied using it to design a simple test;
5. Were satisfied using it to design a computer adapted test.

Research Limitation

There were three main limitation of this study which involved the size of the sample and prior knowledge of the participants, the length of the intervention, and the instrument used.

Obviously the size of the sample prevents any conclusions being made that can be generalised to a larger population. As mentioned, the participants who attended this workshop even though had adequate computer literacy, they had no exposure to the IRT which provides the foundation in understanding CAT and, no knowledge of the R script language which is fundamental in designing the simple and adaptive tests. The participants had to learn to deal with syntax of R language as they were creating the two tests, a task that they had successfully accomplished with much feat.

There was also the limitation that a one day workshop is too short a time period for the participants to master the programme. At best the participants developed a surface understanding that would require further consolidation. The issue of retention of this knowledge could become the focus for future research.

Another limitation was that the instrument that was used could not be validated due to the small sample of 12 participants. Although Lund (2001) provided evidence about the reliability of the instrument and posited that the items that contributed to each scale were of approximately equally weighted and exhibited high Cronbach's Alpha (Lund, 2001), it cannot be applied here due again to the size of the sample.

Conclusion

CAT although relatively new in the field of assessment in the SEAMEO region, has the potential to be a useful tool in testing students' performance and could become one of the basic testing procedures. There are numerous potentials and advantages with efficiency as the most important (Georgiadou, Triantafillou, & Economides, 2006). While access to technology is becoming increasingly widespread in schools and homes, nevertheless, technology is still only marginally integrated into educational assessment at all levels (Erstad, 2008). According to Laborde (2001) and Lagrange, Artigue, Laborde, and Trouche (2003), the successful integration of technology is a rather complex and tedious process. It is argued that high quality in-service programmes for teachers are essential for successful technology integration. This study presented the evaluation of an in-service programme using Concerto, an open source, on-line adaptive testing platform software. While acknowledging the limitations of a small sample, the results of this study identified difficulties and challenges that the teachers faced while participating in a one-day introductory workshop for learning the use of the new software to create an on-line adaptive test. These findings suggested how this technology in-service programme for teachers could be improved. The study immediately resulted in the development of an improved and enhanced design of several new handouts to ease difficulties of novice users and, devised plans to extend the duration of future workshop for effective and smooth transition of knowledge. In addition, the evaluation tools were improved in our quest to strive for improving in-service programme of teachers as well as developing appropriate materials.

References

Barker, P. (Ed.).(1974). Designing multi-media workstations. In P. Barker (Ed.), *Multi-media Computer Assisted Learning* (pp.53-77). New York, NY: Nicholas Publisher.

Erstad, O. (2008). Changing Assessment Practices and the Role of IT. *International Handbook of Information Technology in Primary and Secondary Education , 20*(2), 181-194.

Georgiadou, E., Triantafillou, E.,& Economides, A. (2006). Evaluation parameters for computer adaptive. *British Journal of Educational Technology , 37*(2), 261-278.

Hambleton, R. K., Swamination, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications Inc.

Laborde, C. (2001). Integration of technology in the design of geometry tasks with Cabri-Geometry. *International Journal of Computers for Mathematical Learning, 6*, 283-317.

Lagrange, J. B., Artigue, M., Laborde, C., & Trouche, L. (2003). Technology and mathematics education: a multidimensional study of the evolution of research and innovation. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, &F. K. S. Leung

(Eds), *Second International Handbook of Mathematics Education*, (pp. 237-269). Dordrecht: Kluwer Academic Publishers

Lilley, M. & Barker, T. (2002). The development and evaluation of a computer-adaptive testing application for English language. In *Proceedings of the 6th Computer-assisted Assessment Conference* (pp.). Loughborough University, UK.

Linacre, J.M. (2000). *Computer-Adaptive Testing: A Methodology Whose Time Has Come*. Retrieved from http://www.rasch.org/memo69.pdf

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

Lund A. M. (2001). Measuring Usability with the USE Questionaire. *Society for Technical Communication Newsletter, 8*(2).

Partners in Learning (2012).*History of CAT.* Retrieved from http://performancepyramid.muohio.edu/pyramid/shared-best-practices/Technology/Computer-Adaptive-Testing/History-of-CAT.html

Root, R. W., & Draper, S. (1983). Questionnaires as a Software Evaluation Tool Interface Design 4 Analyses of User Inputs *Proceedings of ACM CHI'83 Conference on Human Factors in Computing Systems* . New York:ACM.

Rudner, L. (1998). *An on-line, interactive, computer adaptive testing mini tutorial*. ERIC Clearinghouse on Assessment and evaluation.

Russell, M. (2006). *Technology and assessment: the tale of two interpretations*. United Sates of America: Information Age Publishing Inc.

Wainer, H. (1990). *Computerized adaptive testing*. NJ: Lawrence Erlbaum Associates.