Building an Assessment Learning System on the Web

Reggie Kwan[1], Kenneth Wong[2], Chi-wing Yu[3], Philip Tsang[4], and Kat Leung[1]

[1] Caritas Francis Hsu College and Caritas Bianchi College of Careers
rkwan@cihe.edu.hk
[2] Hong Kong Institute of Higher Education
[3]Po Leung Kuk Tang Yuk Tien College
[4]The Open University of Hong Kong

Abstract. This paper presents the design and a preliminary assessment of an online assessment system for learning purposes. The prototype aims to enhance learning by helping teachers teach Key Stage 3 (KS3) Mathematics in Hong Kong The system was built on the concept domain model and the Rasch model to provide an adapting feature to students, such as, navigation support, optimal study path and direct guidance. This is based on the Response Theory (IRT) model to determine students' estimated ability level.. When an individual student enters the system, different problems depending on student's estimated ability level will be assigned to students. Once the system senses an individual student's responses are converging to a particular competence level, the system immediately calculates the student's estimated ability level, identifies his strengths and weaknesses, etc. An immediate score, solution of items, feedback recommended study path and direct guidance can be give instantaneously.

1 Introduction

Web-based learning is growing rapidly worldwide; more and more schools in Hong Kong are integrating learning materials on the Web as part of the standard pedagogy. Most believe that web-based education has the potential not only to provide savings in time and money, but, more importantly, it's flexibility and convenience may also revolutionize the way to live and play. Assessment for learning can be one of the ways of improving learning by gauging students' competency, e-assessment should be an integral part of any e-learning system [1]. As a matter of fact, the Curriculum Development Council published a report entitled "*Learning to learn: The Way Forward in Curriculum Development*" in 2001 to urge schools to put more emphasis on assessment for learning [2]. It is a process in which teachers seek to identify and diagnose student learning problems, and provide quality feedback for students on how to improve their work.

Most students, however, treat assessment as difficult or even a necessary evil. Fortunately, there are techniques that can turn assessment into something effective and efficient but not threatening. Unfortunately, much of the current assessment methods in Hong Kong schools are still the conventional paper-and-pencil test (PAPT), they are bounded by ***time*** and ***place***. It is no easy task to design an online assessment system that focuses on "Assessment for Learning", not to mention the migration process [3].

Mathematics is a challenging subject for students at all levels. Although, textbook publishers and software vendors have developed a lot of digital materials in learning of mathematics, not many of them put an emphasis on assessment for learning and using concept model to build the online learning kits. Most online assessment systems do not provide immediate feedback and recommend study path to make use of the assessed result to help students learn. Our prototype of the online assessment system, on the other hand, aims to enhance learning by helping teachers teach Mathematics from Secondary 1 to 3 (Key Stage 3) in Hong Kong. The system is built on the concept domain model and the Rasch model [4], [5], and [6]. The two integrated models are used for providing adapting features to students, such as, navigation support, optimal study path and direct guidance.

2 Rasch Model and Computer Adaptive Test

The Rasch model is a fairly simple Item Response Theory (IRT) model. IRT can overcome some of the problems and assumptions associated with Classical Test Theory (CTT) and to provide information for decision-making that is not available through CTT. The Rasch model is based on objective measurement [5]. It is based on the probability that an examinee with a given ability level will correctly answer a question representing a given difficulty [7]. Rasch models are also used for analysing data from assessments to measure things such as abilities, attitudes, and personality traits. The Rasch model can be used as an interval scale of scores for both the difficulty of items and the ability of the examinee tested. Interval scores are constant differences along the scale, add & subtract possible; e.g., ratings on students' ability, the difference between 3 and 2 is equal to the difference between 2 and 1, but still a student with 4 is not twice as good/bad as that of 2. These scores are reported in units called logits. Since Logits unit can do addition, subtraction, multiple, and division, it makes useful for educational gains, displays of strengths and weaknesses, and comparisons of different groups.

Rasch model presents a simple relationship between the examinee and the difficulty of items. The mathematical formula of the Rasch Model is given bellow:

$$Log\left(\frac{P_i}{1-P_i}\right) = \boldsymbol{\theta_j - b_i} \qquad (1)$$

Where

P_i : probability for an examinee responding correctly.

θ_j : ability parameter of an examinee.

b_i : difficulty parameter of an item.

Figure 1 is the Rasch Model Test Characteristic Curve [8]. It shows the relationship between the probability P(i=1) and (θ_j - b_i) the difference between the examinee's ability level θ_j and the item difficulty b_i .

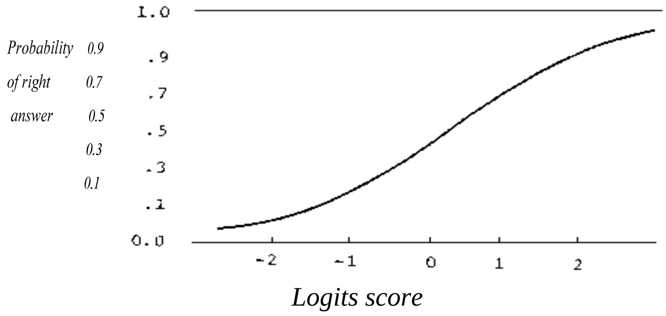

Fig. 1. The Rasch Model Test Characteristic Curve. [8].

If an examinee’s ability level exactly equals to the difficulty level b_i of the item, he/she will have a 50% chance of passing an item. Similarly, if the examinee’s ability level is greater than the difficulty level of the item, he/she will have over a 50% chance of responding correctly to this item. Conversely, if the examinee’s ability level is less than the difficulty level of the item, he/she will have less than a 50% chance of responding correctly to this item. The best design for the selection algorithm of items is that the difficulty level of an administrated item is close to the current ability level. The examinee ability level parameter and item difficulty parameter can be estimated iteratively through application of a process such as Conditional Maximum Likelihood estimation.

In analyzing Rasch data, there are two chi-square fit statistics should be concerned - Outfit and infit statistics [9]. Outfit statistics are more sensitive to extreme scores and infit statistics are more sensitive to unexpected patterns. Use of this two fit statistics information, the Rasch model helps the user identify any items that are not fitting the model, and any examinee whose scores do not appear to be consistent with the model.

Computer Adaptive Test (CAT) works like a good oral exam. Examinees receive the question in accordance with their ability. After the response is given, the result is calculated immediately. If an answer is correct, the next question generated will give a higher difficult item. If the answer is incorrect, the procedure will be reversed. The examinee's ability level can be estimated during the testing process [10]. Since the item selected next for obtaining ability estimates is based upon one's previous item performance, an algorithm must be chosen for sequencing the set of test items administered to the examinees. Therefore, using Rasch model to design such algorithm is very suitable.

3 Concept Domain Model

However, CAT may be optimal for determining an individual's overall ability level, it doesn't assure content balance and doesn't guarantee that one could obtain subtest scores. To overcome this concern, the algorithm should develop a set of construction rules to select the best questions. To optimize the online assessment system, the research team decided to construct a set of construction rules based on a concept domain model. The concept domain model consists of two parts: skillful tree and curriculum tree. The skillful tree can descript as a relationship between different skill interconnected together to form a network, shown as Figure 2 with appendix A shows the Mathematics Skill to be mastered in solving linear equations. The curriculum tree can described as a relationship between different knowledge elements interconnected together to form a network (with dependency), shown as Figure 3 with appendix B shows an example .

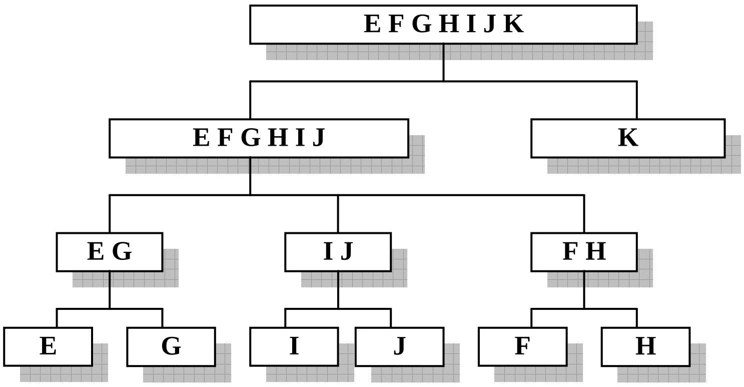

Fig. 2. A Skillful Tree

4 Building an item bank

To implement a CAT, an item bank containing all items is necessary. “An item bank is a large collection of test items organized and catalogued to take into account the content of each test item and also its measurement characteristics.” [6]. Actually, the item bank is a database of items. The size of the item bank should be big enough to cover the wide range of test content. The great advantage of an item bank is its flexibility. Tests can be long or short, easy or difficult depending on the aim of the test. Normally, the questions in CAT are drawn from an item bank. All individual items are carefully calibrated and ranked in difficulty. However, there are several disadvantages of building an item bank. No item bank is perfect. The items in an item bank must be continually re-calibrated. Therefore, all item bank has to continually maintain its standard. Such ongoing work requires a lot of resources.

It is expensive and time-consuming to establish an item bank especially on this CAT. Initially, there are about 100 existing items from the past few year tests which are suitable to the assessment system. However, those items are not calibrated and with difficulty levels. To assign difficulty levels of the items, research team members will first guess intelligently. As they have expertise in Computer Science, Education and Mathematics, they know what topic areas are harder than others for those at any stage of development and know which item is suitable for the assessment system. In addition, inspection of individual items gives indications of their relative difficulty. Initially, items are stratified into 10 different difficulty levels and related topic areas. Each research member gives his/her scale first, then an average will be derived accordingly. Consequently, a fairly stratifying of items by expert-perceived difficulty can often be accomplished.

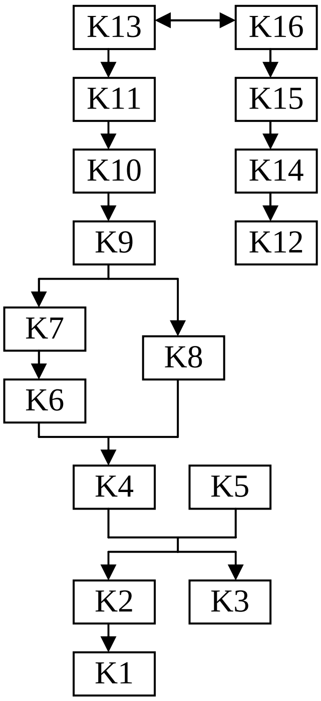

Fig. 3. Curriculum Tree

Nevertheless, this is only a preliminary state of estimated items' difficulty levels, all initial items must be examined and re-calibrated by a pilot conventional test. The pilot test will be administered later this year with students in this very course. The results of the test will be used to verify the item whether it is functioning as specified and to ascertain the item's precise difficulty. Then the item can be made part of the item bank.

5 System design and Architecture

There are tools that can do adaptive testing, and tools that can do tutoring, and tools that can do analysis of results from students doing on-line tests; but there is no single tool that incorporates all three of these functions. One of our goals is to develop an online assessment system incorporates all three of these functions.

5.1 System features

The distinctive features of our online assessment system are:

- students have choice to select Knowledge mode or Skilful mode to learn;
- students can select different knowledge elements to learning on the Knowledge mode;
- students can select different skills to practice on the skilful mode;
- students are given different problems during the test;
- selection of test question is based on a set of construction rules and Rasch model;
- the level of ability of the individual student is classified according to the their performance;
- students are automatically marked and their results were summarised in a report;
- feedback, recommended study path and direct guidance would be given to studnets when he complete a question.
- Teachers are able to access an individualized or group report of the students.
- all test questions and content materials are in form of hypermedia;
- the system have an explicit user-model which records some features of the individual student; and
- the system has a concept domain model, which is a set of relationships between knowledge elements in the information space;

System should be strongly adaptive, working in a well-structured information space; gathering data about the students' ability level and using this information to dynamically generate an optimal question to students. System should also be capable of altering the sequencing of question content, or appearance of the direct guidance and, an optimal study path on the basis of a dynamic understanding of the characteristics of the individual user.

5.2 Selection algorithm

An algorithm to choose the next best items is based on a set of construction rules and the Rasch model. To estimate examinee's ability, the algorithm is based on Rasch model. We use the algorithm described by [11] to estimate an examinee's ability. The next item generated will give an appropriate difficulty with the examinee's ability. This iterative process is part of the Rasch Model. Similarly, to estimate an appropriate content of next item, the algorithm uses a set of construction rules. We employs the concept domain model. If the user selects the Knowledge mode, the construction rules will focus on the content balance on different knowledge elements. If the user selects the Skilful mode, the construction rules will focus on the content balance on different skill type. The steps are shown on figure 4.

As examinees answer each question, the computer scores the question and uses that information together with the responses of the previous questions to determine which question is presented next. If an examinee gets a correct answer with a given item, the system will generate an item from the pool with slightly more difficult than the current one. Then the best next item will be selected with some constraints from this item pool based on a set of construction rules. If the examinee gets the item wrong, the process is similar, however the next item received will be easier. When the Stopping condition is obtained, the testing session ends. The stopping conditions are crucial factors for our system, they are when:

1. no more relevant questions are left in the item bank; or
2. a pre-determined time-limit was reached; or the examinee decided to quit the test.

The current estimated ability level was the final ability level of the examinee. Then examinees will be given an immediate feedback, study guides if any and score on their performance and examinees' performance can be tracked by using the computer to store performance data.

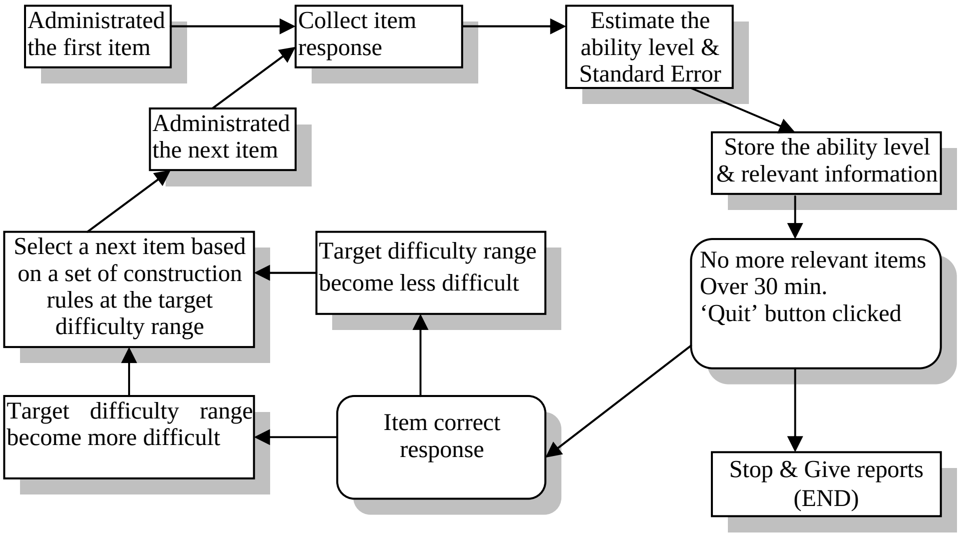
Administrated the first item
Collect item response
Estimate the ability level & Standard Error
Administrated the next item
Store the ability level & relevant information
Select a next item based on a set of construction rules at the target difficulty range
Target difficulty range become less difficult
No more relevant items
Over 30 min.
'Quit' button clicked
Target difficulty range become more difficult
Item correct response
Stop & Give reports (END)

Fig. 4. The Algorithm of the online assessment system

5.3 Design of the database

The system contains several main records: the item record, the item frequency record, the student record, the student response record, student test record, student ability record, knowledge domain record, and skill domain record. Details are elaborated as follows.

Item records stores all relevant item information, e.g. the item ID (*Qid*), content topic (*Qtype*), skill type(*Stype_A*, *Stype_B*,…) , date of upload (*Qdate*), question statements (*Stem*, *Alt_A*, …, *Alt_D*), keys (*Key_A*, …, *Key_D*) and difficulty level (*Rindex*).

The item frequency record contains the item ID (*Qid*), the number of right (*R_freq*) and wrong (*W_freq*) responses by students, total operation time (*Total_time*) by the students).

The student record stores all relevant student's information. The record consists of student ID (*SID*), first name (*F_Name*), last name (*L_Name*), class (*Class*).

The student response record stores student's information examine one item, e.g. student ID (*SID*), test indicator (*Tid*), item ID (*Qid*), student response on item(*R_time*), login date and time (*Date_test*).

The student test record stores student's information during the test,

e.g. student ID (*SID*), test indicator (*Tid*), executive time on item (*Ex_Time*), and executive time for the whole test (*Finish_Time*).

The student ability record stores student's ability information during the test, e.g. estimated ability level (*Ability_L*), standard error (*SE*), knowledge_domain(*K_1, K_2,K_3, …*), skill_domain(*S_1, S_2, …)*

Knowledge domain record and skill domain record store information of knowledge elements and skill type respectively.

5.4 Preparation of Item Bank

To implement a CAT, an item bank containing all items is necessary. "An item bank is a large collection of test items organized and catalogued to take into account the content of each test items and also its measurement characteristics." [9]. Actually, the item bank is a database of items. The size of the item bank should be big enough to cover the wide range of test content. The great advantage of an item bank is its flexibility. Tests can be long or short, easy or difficult depending on the aim of the test. Normally, the questions in CAT are drawn from an item bank. All individual items are carefully calibrated and ranked in difficulty. However, no item bank is perfect. The items in an item bank must be continually re-calibrated. Therefore, all item bank has to continually maintain its standard. Such ongoing work requires a lot of resources.

It is expensive and time-consuming to establish an item bank especially on this CAT. Initially, items presented in the item bank were chosen from the past examinations of Mathematics from 2006 to 2007. The questions were restricted to only one of the modules. It consisted of over 100 multiple-choice questions. The item bank was of smaller size compared with a real CAT for this pilot test. All items were categorised into different sub-sections based on the Concept Domain Model. All items were pre-calibrated by a software program called WINSTEP. WINSTEP gave a ranking of the difficulty level of all items. It was the only information used in this system. We employed the difficulty scale unit called logits [5].

5.5 System Architecture

The online assessment system architecture (Figure 5) consists of a Web Interface, a Main system, Database, Rasch Model and Concept Domain Model. The Web interface provides a communication channel between the system and examinees. It operates in conjunction with examinees and the online assessment system.

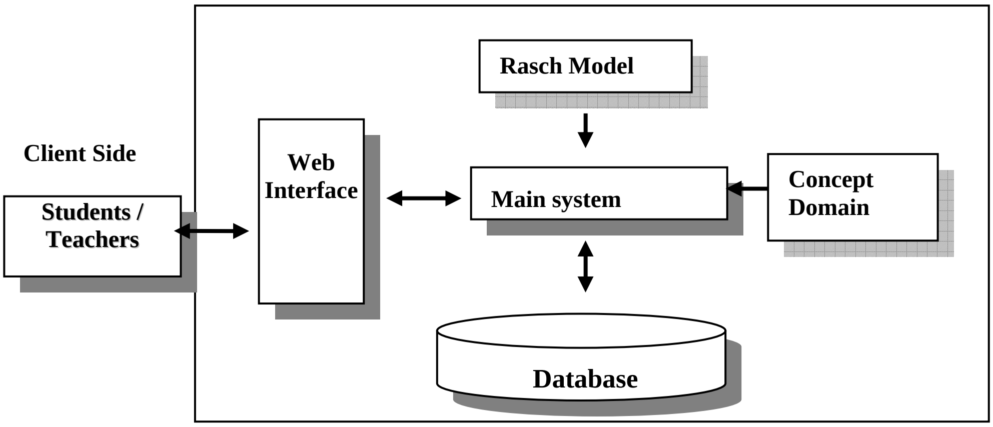

Fig. 5. Architecture of the online assessment system

The main system is the core module in the system, which provides the main tasks in the system: database connection, authentication, estimating the ability level and standard error, selecting the next item, determining the end of the test, giving feedback and generate summary reports. The Database contains item bank which all the pre-calibrated items of information, and examinee response record which contains all the relevant examinees information. The Concept Domain Model contains a set of relationships between different knowledge domains and different skill type domain.

5.6 Choice of Web tools and working platform for building the system

Although the Web framework has some limitations especially in its interface, it gives the system great benefit. The main advantage of the Web is that a user can access the system from anywhere on the Internet without any special interface programs. Two identical online assessment systems were developed, one on Linux and the other one on Window XP. The one using Linux provides a stable and economy working platform for students while the other one using Window XP was chosen for its popularity and the fact that XP is the most used operating systems in Hong Kong schools. In Linux system, an Apache server was used together with PHP and JavaScript to build the front pages and CGI, and uses mySQL as a database for handling information. In Window's platform, the system uses the Window 2003 server as a web server, PHP and JavaScript to build the front pages and CGI, and uses SQL server as a database for handling information.

6 Deployment & Further Development

Several schools are being selected to attend two pilot tests for the online assessment system. A single-group pretest-posttest evaluation design will be used to evaluate the effectiveness of the system. This design compares the same group of participants before and after the programme. The purpose of the single group pretest-posttest design is to determine if students improved after receiving such assessment system.

The research team will request the students to do both tests. Two tests are in traditional PAPT format that is multiple-choice questions; all items are selected from the item bank. Each participant has to answer all questions within the test period. They will not be allowed to leave the test until test session ends. Each participant will give feedback by filling in a questionnaire at the end. The result of the test will be analysed, items will be verified and the difficulty level of each item will be re-calibrated. Then, all re-calibrated items will be used as a part the regular item bank. A pretest pilot test will be deployed in December 2007. The protest pilot test will be scheduled on April 2008 after the students have completed the prototype of the online assessment learning. The process is similar to another project done by one of the authors [12].

Afterwards, the research team will analyse the result of the pilot tests, and the internal reliability and the content validity [13] and [14] will be measured. If the online assessment system fulfills all requirements from the research team, the system will be put in use to other schools. The current system only handles plain text multiple-choice questions. There are other areas that the research team intends to explore, such as a multi-media testing system, and to incorporate the system with electronic tutoring system in other learning programs.

7 Conclusion

Differentiating students’ abilities before they take a course was one of the concerns. An online assessment system has been built and focused on “Assessment for Learning”. The system can be an effective method of assessment in which the computer selects and presents test items to examinees according to the estimated examinees' ability levels. Using the Rasch Model, the system can estimate students' ability effectively and the estimated ability was a useful indicator for instructors’ reference. We learned that building such a system is an on-going process that requires a systematic and meticulous approach. The effectiveness of such a system remains to be seen and will be a great interest of the research team.

References

1. Kwan R. , Chan J. ,and Lui A. (2004) “Reaching an ITopia in Distance Learning—A Case Study”, AACE Journal, Vol 12, Issue 2, pp. 171 – 817
2. CDC (2001) “Learning to learn: The Way Forward in Curriculum Development”, the Curriculum Development Council, Hong Kong SAR, June 2001.

3. Kwan R., & Wong T.M. (2001) "Migrating to the On-line Environment: the experience of OUHK's School of Science & Technology", International Conference on Learning and Teaching On-line, Guangzhou, China.
4. Linacre J.M. (2000), Computer-Adaptive Testing: A Methodology Whose Time Has Come, Memo#69 Iom Research Memoranda, Institute for Objective Measurement, Inc. Available: http://rasch.org/memo69.pdf
5. Keeves J.P. and Alagumalai S. (1999), "New Approaches to Measurement", in Masters, G. N. and J. P. Keeves (Eds). Advances in Measurement in Educational Research and Assessment, Pergamon. (pp.23- 48)
6. Umar J. (1997), "Item Banking", in Educational Research, Methodology, and Measurement, John P. Keeves(editor), Pergamon Publishing.
7. Wright, B.D. and Stone, M.K. (1979). Best Test Design. Chicago: MESA Press.
8. Wright B. D. (1996) Reliability and Separation. Rasch Measurement Transactions 9:4 p. 472.
9. Linacre J.M., Wright B.D. (1994) Chi-Square Fit Statistics. Rasch Measurement Transactions 8:2 p. 350.
10. Rudner L.M. (19998), An On-line, Interactive, Computer Adaptive Testing Mini-Tutorial, ERIC Clearinghouse on Assessment and Evaluation, 1998.
11. Linacre J. M. (1999), "Individualized Testing in the Classroom", in Masters, G.N. and Keeves, J.P. (Eds). Advances in Measurement in Educational Research and Assessment. New York: Pergamon (pp. 186-194).
12. Wong, K. (2000). "Web-based Tool for a Computerised Adaptive Test System", Master Thesis, University of Victoria Wellington, New Zealand.
13. Wong K., R. Kwan R., and Chan J. (2002), "A Preliminary Evaluation of a Computerized Adaptive Test System on the Web", Kwan R., and etc. (Eds). Web-based Learning: Men & Machines, New Jersey: World Scientific Publishing, 2002, pp. 123-134.
14. Wong K. and Kwan R. (2002) "Building a Web-based Computerized Adaptive Testing System", International Conference on Information and Communication Technologies in Education. Badajoz, Spain, Vol I, 62-66.

Appendix A. Mathematics Skill to be mastered in solving linear equations

Code	***Skill***
A	Appropriate use of letters to represent numbers
B	Understand the language of Algebra
C	Formulation of Linear Equations in one unknown
D	Formulation of Linear Equations in 2 unknowns
E	Simplify +– on one side
F	Simplify ÷× on one side
G	Moving +– terms
H	Moving ÷× terms
I	Grouping
J	Removal of brackets
K	Simplification
L	Changing of subject of equation
M	Technique of substitution
N	Multiplication of Equations by a factor
O	Technique of Elimination

Appendix B. Example of Content Model

Code	***Concepts***
K1	Use Symbols or letters to represent numbers
K2	Understand the language of algebra including translating word phases into algebraic expressions or write descriptive statement for algebraic expressions
K3	Understand the concepts of equations
K4	Formulation of Linear Equations in one unknown
K5	Formulation of Linear Equations in two unknowns
K6	Solve simple equations involving one step in the solutions and check answers (involving whole numbers only)
K7	Solve problems by simple equations (involving only one step in the solution)
K8	Solve equations involving almost two steps in the solutions
K9	Solve problems by simple equations (involving at most two steps in the solutions)
K10	Solve linear equations in one unknown
K11	Solve literal linear equations
K12	Plot and explore the graphs of linear equations in 2 unknowns
K13	Solve simultaneous equations by algebraic method
K14	Solve simultaneous equations by graphical method
K15	Awareness of the approximate nature of the graphical method
K16	Explore simultaneous equations that are inconsistent or that have no unique solution