Implementation of an Improved Adaptive Testing Theory

Mansoor Al-A'ali
Department of Computer Science, College of Information Technology, University of Bahrain, Kingdom of Bahrain,
malaali@itc.uob.bh // mansoor.alaali@gmail.com

ABSTRACT

Computer adaptive testing is the study of scoring tests and questions based on assumptions concerning the mathematical relationship between examinees' ability and the examinees' responses. Adaptive student tests, which are based on item response theory (IRT), have many advantages over conventional tests. We use the least square method, a well-known statistical method, to reach an estimation of the IRT questions' parameters. Our major goal is to minimize the number of questions in the adaptive test in order to reach the final level of the students' ability by modifying the equation of estimation of the student ability level. This work is a follow-up on Al-A'ali (2007). We consider new factors, namely, initial student ability, subject difficulty, number of exercises covered by the teacher, and number of lessons covered by the teacher. We compared our conventional exam results with the calculated adaptive results and used them to determine IRT parameters. We developed the IRT formula of estimating student ability level and had positive results in minimizing the number of questions in the adaptive tests. Our method can be applied to any subject and to school and college levels alike.

Keywords

IRT, Item response theory, Testing methods, Adaptive testing, Student assessment

Introduction

Item response theory (IRT) is the study of scoring tests and questions based on assumptions concerning the mathematical relationship between the examinee's ability (or other hypothesized traits) and the questions' responses. Adaptive student tests, which are based on IRT, have many advantages over conventional tests. The first advantage is that IRT adaptive testing contributes to the reduction of the length of the test because the adaptive test gives the most informative questions when the student shows a mastery level in a certain field. Secondly, the test can be better tailored for individual students.

Adaptive assessment would undoubtedly present improved methods of assessment, especially with the availability of computers in all schools. As we know, evaluation and assessment are an integral part of learning. A good objective test at the end of each learning objective can reveal a great deal about the level of understanding of the learner. The possibility of differential prediction of college academic performance was discussed by researchers (Young, 1991). A good analysis of item response theory was presented in (Fraley, Waller, & Brennan, 2000) who applied the theory to self report measures of adult attachment. Error-free mental measurements resulting from applying qualitative item response theory to assessment and program validation, including a developmental theory of assessment, were discussed in Hashway (1998). The general applicability of item response models was extensively discussed by Stage (1997a, 1997b, 1997c, & 1997d). The use of the item response theory for the issue of gender bias in predicting college academic performance was discussed in Young (1991). Some researchers proposed implementing decision support systems for IRT-based test construction (Wu, 2000). The IRT algorithm aims to provide information about the functional relation between the estimate of the learner's proficiency in a concept and the likelihood that the learner will give the correct answer to a specific question (Gouli, Kornilakis, Papanikolaou, & Grigoriado, 2001). Figure 1 is a schematic representation of an adaptive test.

The IRT algorithm illustrated in Figure 1 aims to provide information about the functional relation between the estimate of the learner's proficiency on a concept and the likelihood that the learner will give the correct answer to a specific question (Gouli et al., 2001). In a conventional test, two matters are considered: the time and the length of the test. In an adaptive test, possible termination criteria are:

1. The number of questions posed exceeds the maximum number of questions allowed.
2. The accuracy of the estimation of the learner's proficiency reaches the desired value.
3. Time limitations: most popular adaptive tests have a time limit. Although time limitation is not necessary in adaptive testing, it can be beneficial. Students who spend too much time on tests may get tired, which can negatively affect the score.

4. No more relevant items in the item bank: when the item bank is small, or questions with a difficulty level suitable for the student do not exist, the test must be terminated.

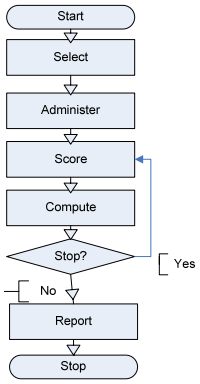

Figure 1. A schematic representation of an adaptive test

Conventional exams suffer from certain problems that must be considered carefully. First, assessments and assignments are normally given to test the different capabilities of students who are in the range of poor to excellent, and thus all students have to meet or resolve the same standard questions, which mostly do not match their own capabilities, and are either lower or higher. This means that students with high capabilities may waste their time in solving average assignments that do not excite, challenge, or interest them. Another aspect that we should keep in mind considerably is that a score of 75% on a test containing easy items has a different meaning from a score of 75% on a test containing difficult items (Segall, 2000). Second, testing for knowledge and understanding in the context of a specific course typically involves administering the same set of test items or questions to all the students enrolled in the course, usually at the same examination sitting.

When we consider classical test theory (CTT), we realize that CTT has a number of deficiencies. One of the problems with CTT is that it is test-oriented rather than item-oriented; that is, in the classical true-score model there is no regard for an examinee's response to a given item. As a result, CTT does not allow predictions to be made about how an individual or group of examinees will perform on a given item (Hildegarde & Jacobson, 1997).

To formulate a complete understanding of the assessment process in schools, we look at some drawbacks of pencil-and-paper tests. Some drawbacks of conventional pencil-and-paper tests are scoring and feedback. Instructors need a lot of time to correct test papers. This means that examinees cannot be informed about the result of the test immediately after its completion. We know that immediate feedback is important to the students for psychological reasons. It is motivational, helps them focus, and informs them if they have to work harder.

One of the solutions to these problems is to use computer adaptive tests (CATs). Advantages of CATs can include shorter and quicker tests, flexible testing schedules, increased test security, better control of item exposure, better balancing of test content areas for all ability levels, quicker test-item updating, quicker reporting, and a better test-taking experience for the test-taker. CATs are widely used these days and they give good results in many educational fields. CATs are used in many professional certification programs. Novell successfully introduced CATs into its certification program in 1991. The Educational Testing Service, the world's largest testing organization, published the Graduate Record Exam (GRE) as an adaptive test in 1993. TOEFL also uses a CAT. The Nursing Boards converted completely from paper-based testing to a computerized adaptive test in 1994.

Assessments can guide improvement—presuming they are valid and reliable—if they motivate adjustments to the educational system (Shute & Towle, 2003). "The question is no longer whether assessment must incorporate technology. It is how to do it responsibly, not only to preserve the validity, fairness, utility, and credibility of the measurement enterprise but, even more so, to enhance it" (Bennett & Persky, 2002).

Intelligent tutoring systems permit the modeling of an individual learner, and with that modeling comes the knowledge of how to perform an individualized assessment. Examinations can once again be tailored to meet the individual needs of a particular learner. In contrast with paper-and-pencil multiple-choice tests, new assessments for

complex cognitive skills involve embedding assessments directly within interactive, problem-solving, or open-ended tasks (Bennett & Persky, 2002).

One of our main objectives in this research was to study IRT in order to reduce the number of questions before reaching stability. It had been shown (Gouli et al., 2001; Eggen & Straetmans, 2000) that after 13–15 questions, the level of students' ability becomes stable. Therefore, the length of time required to complete the adaptive test is shorter than that required for the pencil-and-paper test. Our hypothesis is that the starting level of the adaptive test is arbitrary. Our starting point in the adaptive test was at the conventional level that we got from stage 1 of the method previously mentioned. Our target was to get a stable level after seven questions only, hence reducing the length of the test and the time required to complete the test. In order to test our hypothesis we had to build a system and evaluate and enhance IRT. We measured the effectiveness of IRT by comparing the students' achievement levels using the new IRT-based system with the results achieved by the students after taking an ordinary written test prepared by the teachers.

Testing, Adaptive Testing, and Item Response Theory

Linear tests are not adaptively administered tests and thus they are presented on the far-left side of Figure 2. Items on these tests are represented in sequence, that is, they are linear. The examinee is presented with the first item, then the second, then the third, and so on, in a predetermined fashion. Linear tests administered on the computer are also known as fixed-form tests.

In linear-on-the-fly testing (LOFT), unique, fixed-length tests are constructed for each examinee. The target of content and psychometric specifications should be met in constructing the test. However, the examinee's proficiency level is not a consideration when constructing form items; thus, these tests are not adaptive. A large pool of items is needed to develop this type of test because the test forms should be unique. The benefits of constructing LOFT tests are in presenting item exposure and rigorous content-ordering requirements. Therefore, the LOFT model has one major advantage over the linear model: improved security that comes from presenting different items across forms.

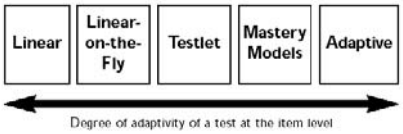

Figure 2. Types of Testing

Testlets are groups of items that are considered a unit and administered together. Usually, testlets are constructed based on previous knowledge of the difficulty of the items or based on content. More specifically, testlets are developed according to the order of items' difficulty or their ability to meet content specifications. Testlets are presented to examinees in units. Within testlets, examinees are given the opportunity to review, revise, and omit items. Items within a testlet may be assembled by similarities in level of difficulty, subject matter, or both. Hence, multistage testing will be possible.

Mastery Models tests are developed to provide accurate information about mastery/non-mastery. The main goals of mastery models are (1) covering the content domain and (2) making accurate mastery decisions. There are various models to implement mastery models, and they all share a major advantage, namely efficiency. Efficiency is observed in the ease of classifying all examinees based on simple rules. Eggen and Straetmans (2000) and Rudner (2002) provide good descriptions of classifying examinees into three categories.

Tests based on CAT delivery models present items depending on the performance of the examinee. The items that are presented have been pre-tested and item parameter estimates have been calculated. Using this information, examinees receive items that match their proficiency level at that time. Adaptive assessment systems can ask the most informative questions and determine when the student has displayed mastery of a particular concept. There is no need to pose further test items. The test may be better tailored for individual students.

Adaptive assessment has two main goals. First, the length of the test may be decreased because the adaptive test gives the most informative questions when the student shows a mastery level in a certain field. Second, the test may be better tailored for individual students. Adaptive assessment can provide accurate estimation of the learner's proficiency in an efficient way without forcing him/her to answer questions that are either too easy or too difficult for him/her (Gouli et al. 2001).

In the example shown in Figure 3, the examinee's level was about 50 out of 100, i.e., his capability for answering questions of varying difficulty is average. The first question the computer gave the examinee was of level 52, i.e., a slightly more difficult question than his initial estimated capability. The examinee correctly answered the first two questions and his level rose. When he answered the third question incorrectly, his level went down. The process continued until there was minimal error in estimating the examinee's level. The computer program becomes more and more certain that the examinee's ability level is close to 50 (Linacre, 2000).

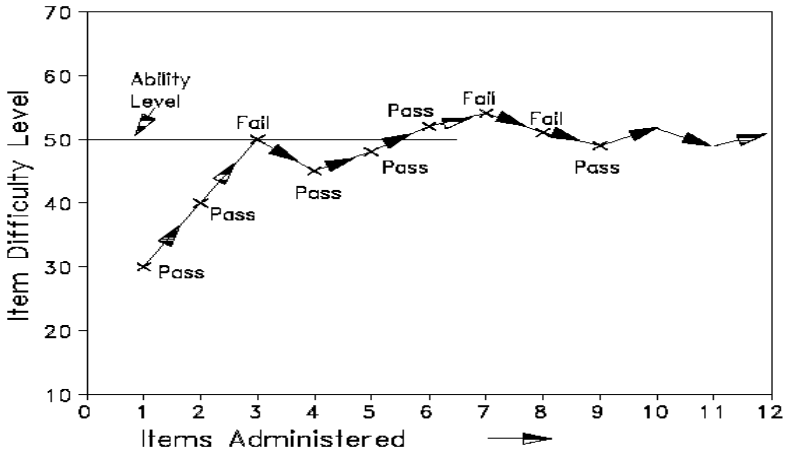

Figure 3. Example CAT Test Administration

In this research project, we use item response theory (IRT). IRT is a modern test theory designed to address the shortcomings inherent in classical test theory methods for designing, constructing, and evaluating educational and psychological tests (Hambleton, Swaminathan, & Rogers, 1991). One of the functions of IRT plots respondent behavior across the continuum of abilities underlying the concept of interest. In other words, IRT is adaptive.

IRT was used in the implementation of adaptive assessments because the proficiency estimate is perhaps independent of the particular set of questions selected for the assessment. Each learner gets a different set of questions with different difficulty levels while taking the adaptive assessment (Weiss, 1983). IRT-based item selection strategies (Weiss, 1983) are maximum information item selection strategy, and Bayesian item selection strategy.

Kingbury and Weiss used maximum information item selection strategy, in which the item pool is searched for an item that can give maximum information about the examinee. In Bayesian item selection strategy, used by McBride and Martin, the item selected from the item pool will maximally reduce the posterior variance of an individual's ability estimate. Bayesian item selection strategy uses prior information about the examinee more completely than maximum information item selection strategy (Weiss, 1983).

Eggen and Straetmans (2000) showed that optimum item selection is largely responsible for the major efficiency gains of CAT against a fixed, linear paper-and-pencil test. Efficiency gains means that fewer items are required to assess candidates with the same degree of accuracy, or that they can be assessed much more accurately with the same number of items. Table 1 shows that compared to pencil-and-paper tests, CAT achieves a higher percentage of correct answers when using maximum information, and fewer items are required to achieve a similar level of accuracy.

Table 1. CAT and a linear mathematics intake test

Method	Average number of items	% correct decisions
Paper-and-pencil	25	87.0
CAT(maximum information)	14.2	88.3
CAT (random selection)	20.2	85.2

Adaptive Assessment Algorithm

The IRT algorithm aims illustrated in Figure 1 provide information about the functional relation between the estimate of the learner's proficiency in a concept and the likelihood that the learner will give the correct answer to a specific question (Gouli et al., 2001).

Item Characteristic Curve in IRT

The item characteristic curve (ICC) is the basic building block of item response theory. There are two technical properties of an item characteristic curve. The first is the difficulty of the item. Under item response theory, the difficulty of an item describes where the item functions along the ability scale. The second technical property is discrimination, which describes how well an item can differentiate between examinees with abilities below the item location and those with abilities above the item location (Baker, 2001).

At each ability level, there will be a certain probability that an examinee with that ability will give a correct answer to the item. This probability will be denoted by P(θ). Its formula is given by

$$P(\theta)=\frac{1}{1+e^{-L}}=\frac{1}{1+e^{-a(\theta-b)}}$$

where:
b is the difficulty parameter
a is the discrimination parameter
$L = a(\theta - b)$ is the logistic deviate (logit) &
θ is an ability level.

The importance of item discrimination comes from the fact that it relates the strength of the relationship between a test item and the underlying (and unobservable) attribute being measured, for example, knowledge or learning.
In the case of a typical test item, this P(θ) will be small for examinees of low ability and large for examinees of high ability (Baker, 2001). Adding one more factor to the previous factor, the formula will be (Gouli et al., 2001):

$$P(\theta\ )=c+\frac{1-c}{1+e^{-2(\theta-b)}}$$

where *c* is unknown. Notice that $a = 2$ is given for the purpose of simplification.

Item Information Function

Item information function (IIF) is considered a very important value in IRT. IFF is used in estimating the value of the ability parameter for an examinee. Moreover, it is related to the standard deviation of the ability estimation. If the amount of information is large, it means that an examinee whose true ability is at that level can be estimated with precision; that is, all the estimates will be reasonably close to the true value. If the amount of information is small, it means that the ability cannot be estimated with precision, and the estimates will be widely scattered about the true ability (Baker, 2001).

In statistics, Sir R. A. Fisher defined information as the reciprocal of the precision with which a parameter could be estimated (Baker, 2001). Statistically, the precision with which a parameter is estimated is measured by the variability of the estimates around the value of the parameter. The amount of information is given by the formula:

$$I = \frac{1}{\sigma^{2}}$$

where σ^2 is a measure of precision of the variance of the estimators.

Estimating Item Parameters

I) Rasch model for estimating item parameters
Mathematical analysis shows that the Rasch model is statistically strong. Rasch model estimates are sufficient, consistent, efficient, and unbiased. It estimates the difficulty parameter and student ability (student level). There is an efficient method for approximating parameter estimates that could easily be calculated by hand. The drawback of this model is that there are no guessing factors and discrimination parameters. Besides, estimating items' parameters and students' abilities in a test with 20 or 30 items requires at least 100 examinees.

II) Chi-square goodness-of-fit
When the conventional test finishes, we observe a sample of *M* examinees' responses to *N* items in the exam. According to examinees' achievement results, we determine the examinees' ability levels. These ability levels will be distributed over the ability scale. According to Baker, "The agreement of the observed proportions of correct response and those yielded by the fitted item characteristic curve for an item is measured by the chi-square goodness-of-fit index" (2001).

III) Method of least square
This method works when the regression is approximately linear. The equation of a straight line is determined by the points that it passes through.

IV) Level estimation method
The student level estimator approach is a modification of the Newton-Raphson iterative method for solving equations method outlined by Lord (1980).

Item Types

I) Dichotomous items
With dichotomous items, there must be only one correct answer. This means that the examinee either answers the question correctly and gets a full mark or answers the question incorrectly and gets no mark. The common dichotomous item types are:

- Multiple choice: this kind of question gives a number of options to choose from (usually four), and the examinee has to choose the correct option.
- True–false: this kind of questions presents a statement that is either right or wrong, and the examinee has to decide whether it is correct or incorrect.
- Short answer: this kind of question provides a space that the student has to fill in with a short answer. If the answer is fixed and no other alternative answer is correct, then the question is dichotomous item.

II) Polytomous Items
Items in this category can be partly correct; the examinee might solve part of the question correctly and will receive part of the mark.

Results

Before estimating question parameters and starting the adaptive test, students were given conventional tests. Students answered about 100 questions in three stages. In each stage, they solved an exam of 34 questions. Forty-five students of the intermediate class were examined in these conventional tests. These questions were of a first-intermediate level. The questions related to three math topics of first-intermediate level.

The least square method was used to estimate the difficulty level and discrimination parameters of each question. We used this method because it decreases the load of using computer processors, and requires no statistical tables such as chi-square.

It is agreed that the difficulty level ranges from +3, which is very difficult level, to –3, which is a very easy level (Baker, 2001). Table 1 shows that the level of difficulty of the questions ranges from +2.9 to –2.9. This means that the questions cover all possible ability levels of students. In table 2, discrimination ranges from –0.3 to –1.22. Positive discrimination values means that the question is not valid and should be revised. There are questions with discrimination greater than –0.3. However, we do not include them in the process of selection of items because they will not be good items because they fall outside the scope of almost all students being tested and would not contribute anything of value to the results.

Table 2. Analysis Of question difficulty

Max difficulty	Min difficulty	Average difficulty
2.9	–2.9	0.51

Table 3. Analysis of discrimination

Max discrimination	Min discrimination	Average discrimination
–0.3	–3.15	–1.22

After the conventional test we tested the adaptive assessment on 24 students. Two of the students did not complete the exam. According to the hypothesis, we should have modified the formula to shorten the number of questions to reach the stable level. We started the adaptive test with the conventional level of the students. Each student started with his previous level and with the consideration that we were certain of 85%, that is, $\sigma = 0.15$. The graph in Figure 1 shows the conformance of the adaptive test level and the conventional test level.

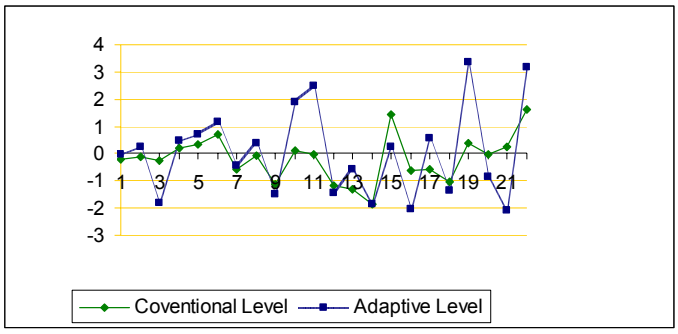

Figure 4. Conformance between conventional and adaptive levels

In Figure 4, about 20% of the points of the conventional levels are not close enough to their corresponding points in the adaptive level. While the average student level in the conventional tests is –0.17, the average in the adaptive tests is 0.031. Transforming these values to the 100 scale, we find that these values are 47.14 and 50.51, respectively. We could say that the averages are almost near each other. The difference between the average of the conventional level and the adaptive level is –0.20. In the 100 scale it is –3.37.

As an example, the fourth point shows that the conventional level is close to the adaptive level: the conventional level is about 0.47 and the adaptive level is about 0.2. This means that the difference is 0.27. In the 100 scale, this difference is about 4 marks. The nineteenth point shows that the conventional level is 0.13 and the adaptive level is 1.9; the difference between these points is 1.8. In the 100 scale, the difference is about 29 marks.

The correlation between the conventional level and the adaptive level is 0.63, which is considered to be high-moderate relation. The correlation is considered high if it is 0.70 and above.

It is not strange to have some deviation in the results of the two kinds of exams due to human error. After elimination of 20% of the points that are very far from their corresponding points, we get Figure 5, which is more accurate than

Figure 4. The average difference of the conventional level and the adaptive level in this case 2 is 0.028. In the 100 scale, it is 0.46.

After the elimination of 20% of the points, the correlation increases and becomes 0.74. This value is considered to be high.

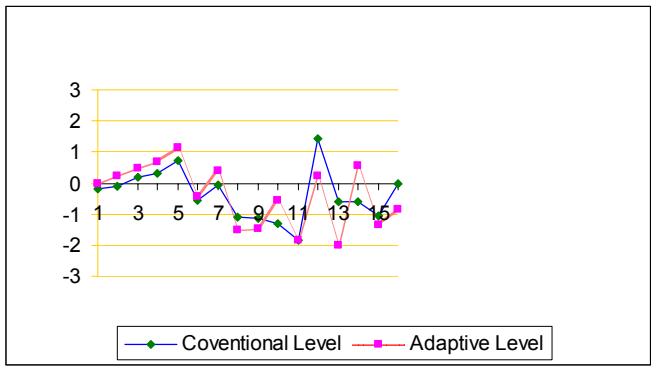

Figure 5. Conformance between conventional and adaptive levels

In Figure 6, we started with the subject difficulty as a starting point of the adaptive test. The results here were done by simulation of real students' results. Notice that the previous chart had a student's initial conventional level as a starting point. In this graph, we could say that the conformance of the two lines is acceptable. The students' conventional levels were the same as they were in the previous chart. While the average of these students' levels in conventional tests is –0.17, the average in adaptive tests is –0.0036. Transforming these values to the 100 scale, we find that these values are 47.14 and 49.9, respectively. Both charts' results are almost the same average. The average difference of the conventional levels and adaptive levels is –0.12. In the 100 scale it is –2.8. The correlation between the conventional and adaptive tests levels is 0.81. The average number of questions to reach the stable or final level is 12 questions.

A question arises after finding that both starting points, the students' conventional level or the difficulty of the subject, reached final levels after an average of 12 questions. Why do the levels of students change? This could be interpreted by the fact that students' levels changed based upon their readiness for the exam, motivation, time of exam, health and other conditions, and the fact that some students wanted to get better results. However, those changes were within an acceptable range in general.

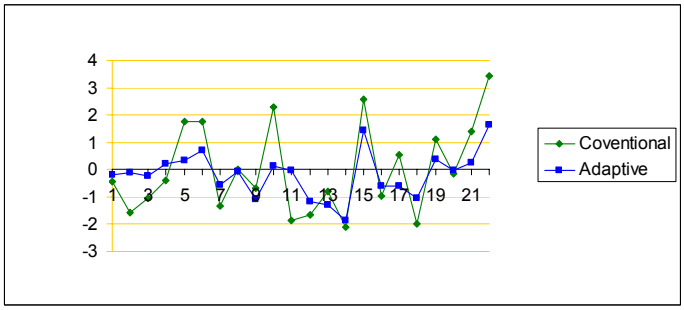

Figure 6. Conformance between conventional and adaptive levels

Our improved IRT formula and model

One of the objectives was to find a new formula that contributes new factors to the IRT model. These factors are initial student ability, subject's difficulty, number of exercises covered by the teacher, and number of lessons covered by the teacher. As shown in Figure 7, the number of lessons covered and the number of exercises covered directly affect the student's ability level.

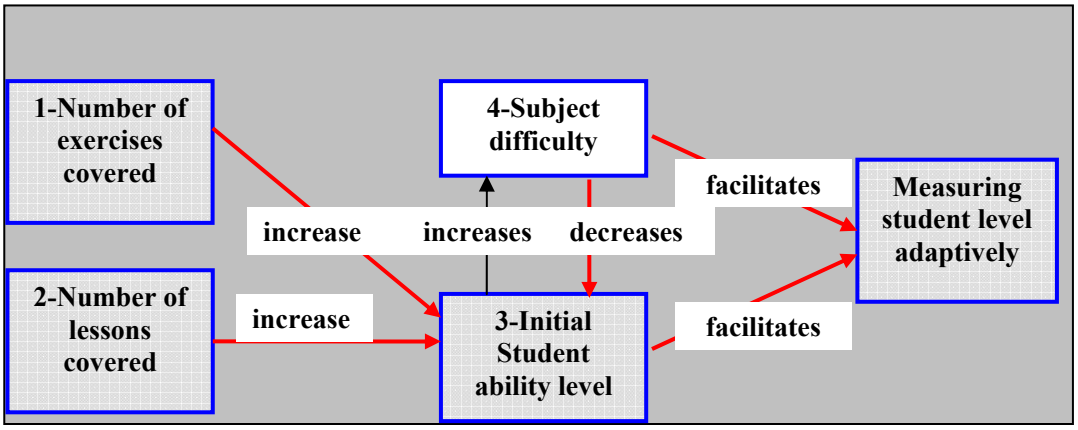

Figure 7. Our newly added factors in IRT model

Moreover, the increase in students' ability levels will in turn increase ease of subject and vice versa. As a matter of fact, the number of exercises and the number of lessons covered do not work by themselves in measuring students' levels because the difficulty of the subject affects these two factors directly. For instance, if the subject is very difficult, the increase in number of exercises and lessons might not help in producing good student results. Therefore, there should be further research that measures these factors in different subject difficulties.

For now, we considered these two factors: subject difficulty and initial student ability level. The new factors will give the new shape of the formula as:

$$\theta_{n+1} = \theta_n + \frac{\sum_{n=1}^{n} s_i(\theta_n)}{(I_{initial}) + \sum_{n=1}^{n} I_i(\theta_n)}$$

where (θ_0)

$$\theta_0 = \begin{cases} \theta_{init} = \text{the initail student level from database} \\ \theta_{diffLevel} = \text{the difficulty level of subject if } \theta_0 \text{ does not exist} \end{cases}$$

and $I_{initial}$ is

$$I_{initial} = \begin{cases} 5 & \text{if } \theta_{init} \text{ exists} \\ 1 & \text{if } \theta_{init} \text{ does not exist} \end{cases}$$

5 means that the standard error, σ, equals 0.45, because we know that θ_0 gives some certainty of the level. In fact, we used this value because it is near the middle of the student ability level and will not negatively affect the process of estimation. 1 means that the standard error = 1, which means that we are not sure and do not know anything about the student level.

Figure 8 shows the performance of a typical student. It is one of the ideal results that we found. This student started in a level of 46 and was almost stable for the first 7 questions. Then his level increased slowly until he became stable at 50. The difference is only 4 marks. This change is expected because he might be a little better or a little worse than his conventional level.

Now, consider Figure 9. About 57% of the students follow the best pattern, in which their levels changed little up or down. 18 percent of the students were fluctuating in their levels. 15 percent of the students decreased in their levels constantly. 10 percent of students increased in their levels constantly. Most students were actually in their correct levels from the beginning of the exam. That shows that students actually start almost from the right level position. Students whose level changed constantly, increasing or decreasing, were the excellent or weak-level students. Students who fluctuated in their level were students who hesitated or revised one or two of the three exam topics.

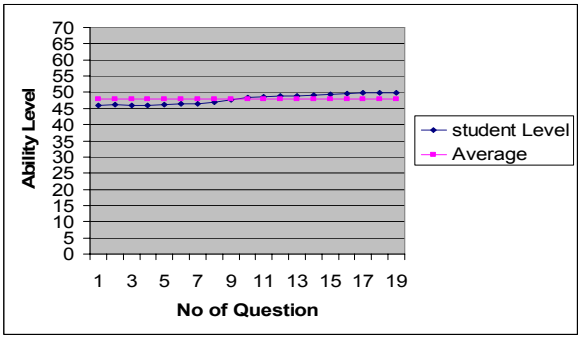

Figure 8. Number of questions vs. level of student

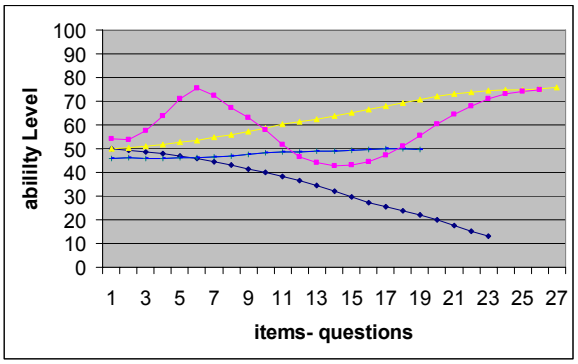

Figure 9. Number of questions vs. level of a student

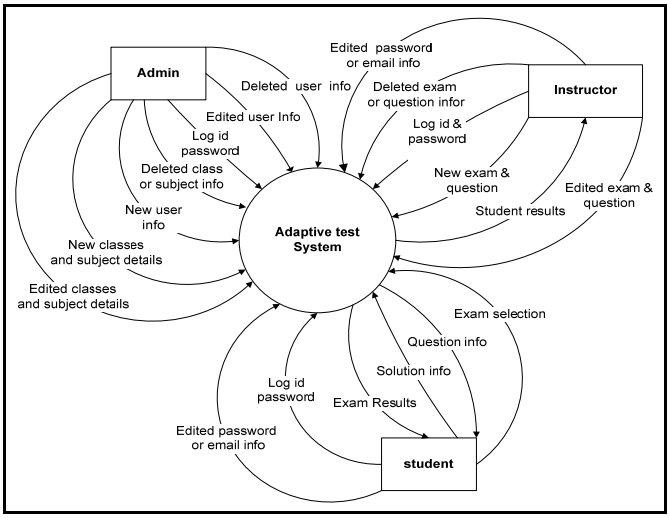

Figure 10. System context diagram

The improved IRT System Implementation

The model shown in Figure 10 was implemented using C++ as the programming platform and Access as the database. The system provides a number of functionalities through a number of processes as illustrated in Figures 11, 12, 13, and 14. The system facilitates a number of functional requirements such as: a login process; a student registration process; add new exams and their questions; create, delete, and edit exam questions; calculate and save

question factors according to IRT; and a wide variety of reports for students' results. Most importantly, the system is the implementation of the adaptive testing process for each student and for a group of students in a given class.

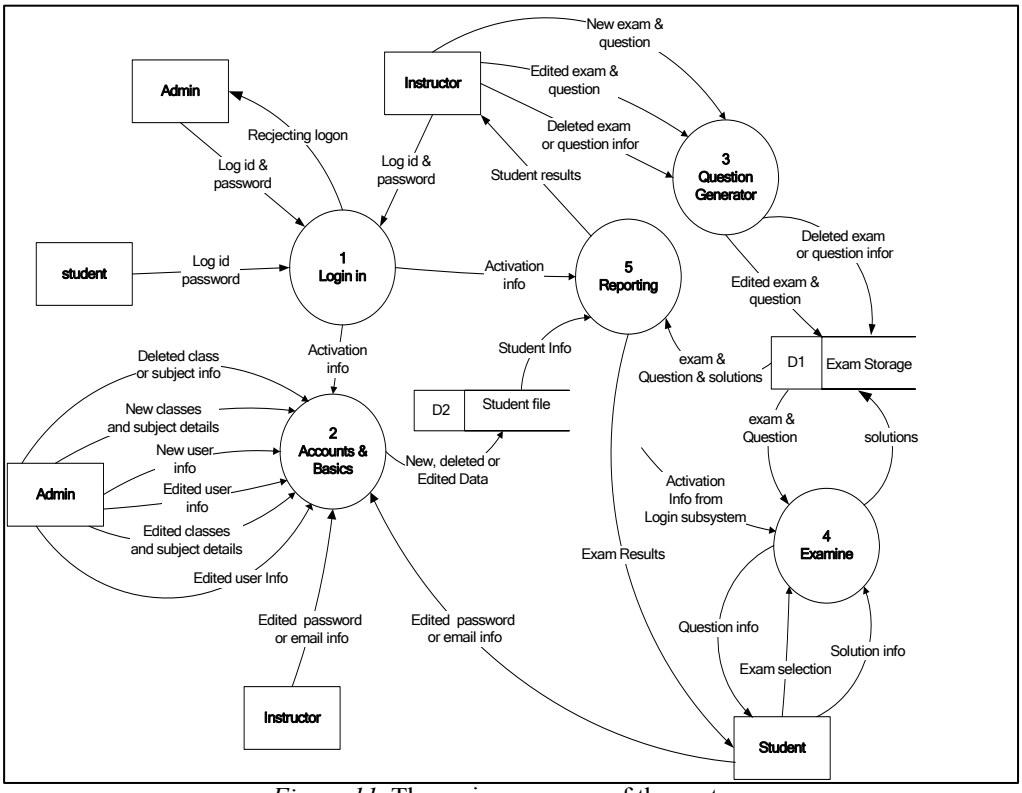

Figure 11. The main processes of the system

The system provides a search based on some criteria, such as exam creation date and the subject of an appropriate exam. The student can select the exam he/she wants to write and answer its questions. The system makes sure that the factors of IRT exist for each question of that exam. The system selects an appropriate question from the database depending on the student's level and presents the question to the student. Depending on student's answer, the system calculates and saves the level of the student as he/she answers each questions using IRT. If the stopping criteria are met, the system finishes the exam.

In order to achieve all the functional requirements through the system processes, the system deals with a database consisting of a number of tables or data stores, including: student table, instructor table, admin table, exam table, question table, question choice table, class table, stage table, teach table, subject table, question solution table, and student-level table.

The context diagram shown in Figure 10 gives the empirical overview of the system. There are three main players in the system, namely, the instructor, the student, and the system administrator.

Figure 11 shows the main processes of the system. Process 1 is a main process responsible for auditing the login activities of all types of users. Process 2 is responsible for all the user accounts, including new classes, subject details, and instructor information. Process 2 is further detailed in Figure 12. Process 3 is the question-generator process that stores all questions in the exam storage data store. Process 3 is further detailed in Figure 13. This process receives from the instructor the new exam questions, the edited questions, and the deleted questions. Process 4 is the process for conducting the exams. Process 4 deals with the student and allows exam selection, shows question information, and provides question solutions. Process 4 is further detailed in Figure 14.

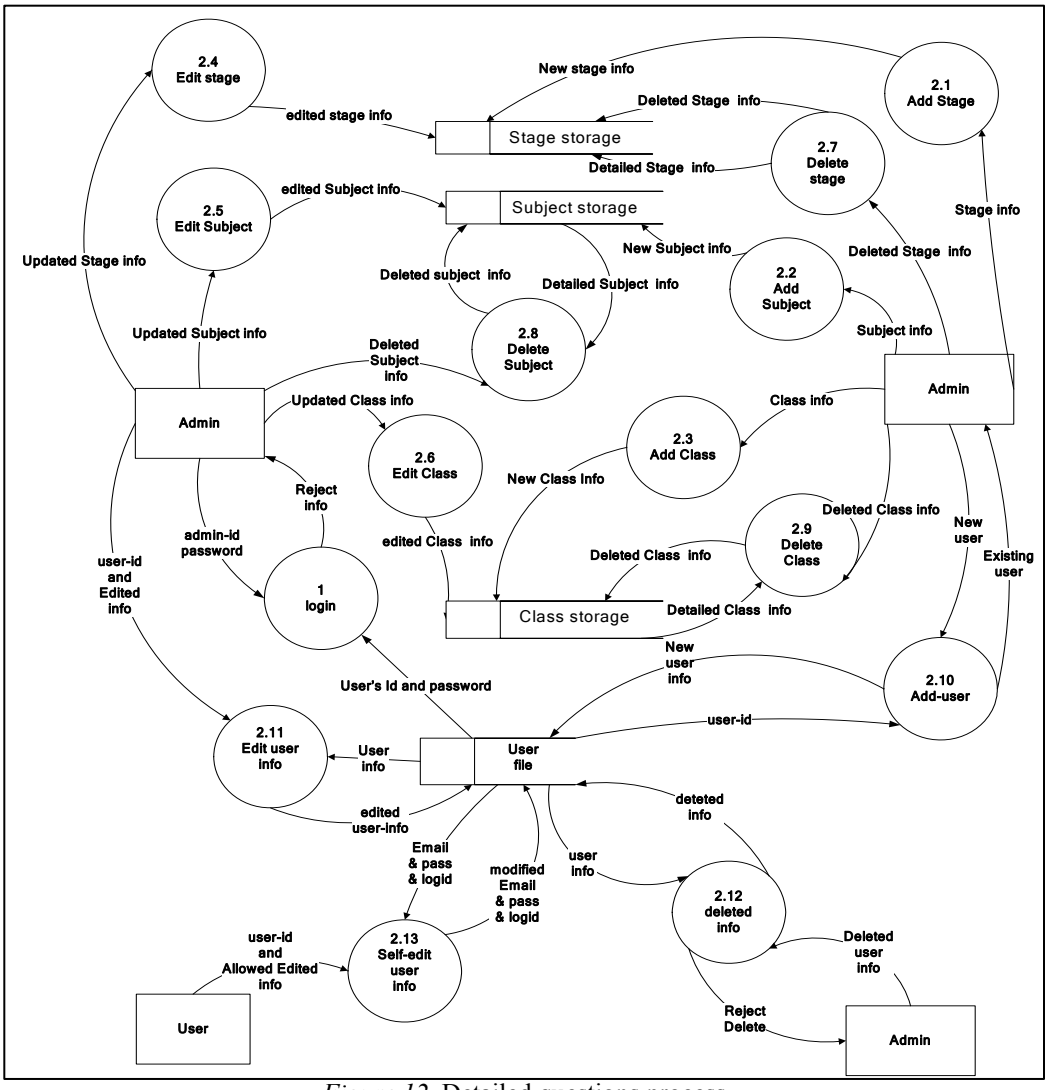

Figure 12. Detailed questions process

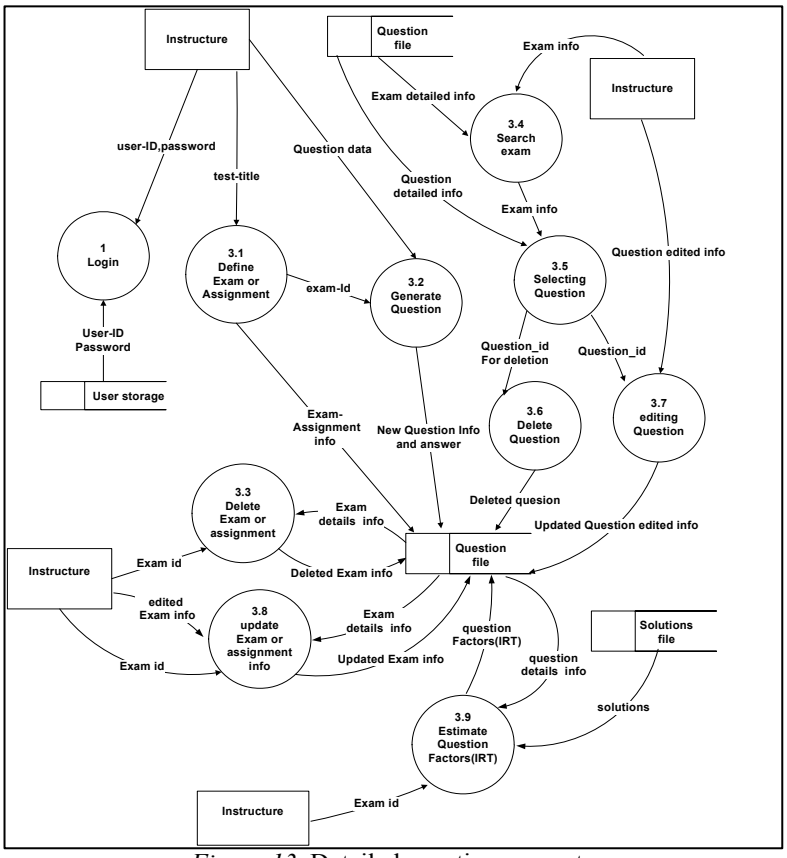

Figure 13. Detailed question generator process

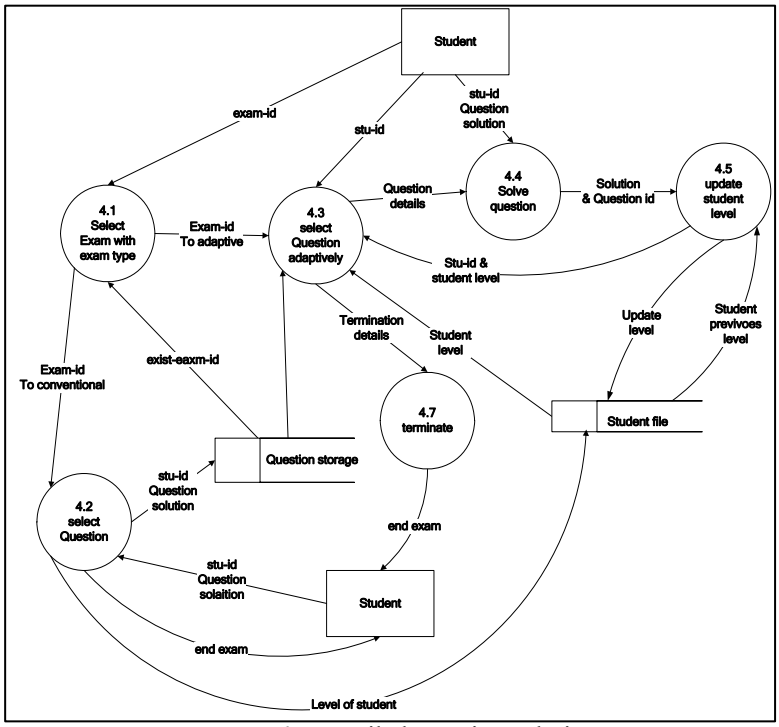

Figure 14. Detailed question solution process

Conclusion

This paper presented a description of adaptive testing based on IRT and experimented with IRT in order to evaluate its applicability and benefits. It also presented enhancements for IRT. Our estimation of IRT questions' parameters was based on the least square method, a well-known statistical method. We demonstrated that it is possible to reduce the number of questions in the adaptive test to reach the final level of the students by modifying the equation for the estimation of the student's ability level. In order to make IRT more realistic and applicable, we incorporated new factors into IRT, namely initial student ability, subject difficulty, number of exercises covered by the teacher, and number of lessons covered by the teacher. Students' attitudes toward adaptive tests and their results were measured by a questionnaire. Our conventional exam results were compared with adaptive results and used in determining IRT parameters. We have presented enhancement and modification of the IRT formula of estimating student ability level and had positive results in minimizing the number of questions in adaptive tests.

References

Al-A'ali, M. (2007). A method for improving adaptive testing by evaluating and improving the item response theory. *WSEAS Transactions on Information Science and Applications 4* (3), 466-471.

Baker, F. (2001). *The basics of item response theory: ERIC Clearinghouse on assessment and evaluation*, College Park, MD: University of Maryland.

Bennett, R. E., & Persky, H. (2002). Problem solving in technology-rich environments. In C. Richardson (Ed.) *Qualifications and Curriculum Authority: Assessing gifted and talented children*, London, England: Qualifications and Curriculum Authority, 19–33.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78* (2) 350–365.

Gouli, E., Kornilakis, H., Papanikolaou, K., & Grigoriado, M. (2001). *Adaptive assessment improving interaction in an educational hypermedia system*, Greece: University of Athens.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*, Newbury Park, CA: Sage Publications.

Hashway, R. M. (1998). *Error-free mental measurements: Applying qualitative item response theory to assessment and program validation including a developmental theory of assessment*, San Francisco, CA: Austin & Winfield.

Hildegarde, S., & Jacobson, Z. (1997). *A comparison of early childhood assessments and a standardized measure for program evaluation*, Ph.D. Dissertation, Virginia Polytechnic & State University, Blacksburg, VA.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (MESA Memorandum No. 69), MESA Psychometric Laboratory, University of Chicago.

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

Rudner, L. M. (2002). An examination of decision-theory adaptive testing procedures. *Paper presented at the annual meeting of the American Educational Research Association*, April 1-5, 2002, New Orleans, USA.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*, Dordrecht: Kluwer Academic, 53-73.

Shute, V., & Towle, B. (2003). *Adaptive e-learning*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Stage, C. (1997a). *The applicability of item response models to the SweSAT: A study of the DTM subtest*, retrieved October 15, 2007, from http://www.umu.se/edmeas/publikationer/pdf/enr2197sec.pdf.

Stage, C. (1997b). *The applicability of item response models to the SweSAT: A study of the ERC subtest*, retrieved October 15, 2007, from http://www.umu.se/edmeas/publikationer/pdf/enr2497sec.pdf.

Stage, C. (1997c). *The applicability of item response models to the SweSAT: A study of the READ subtest*, retrieved October 15, 2007, from http://www.umu.se/edmeas/publikationer/pdf/enr2597sec.pdf.

Stage, C. (1997d). *The applicability of item response models to the SweSAT: A study of the WORD subtest*, retrieved October 15, 2007, from http://www.umu.se/edmeas/publikationer/pdf/enr2697sec.pdf.

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*, New York, NY: Academic Press.

Wu, I.-L. (2000). Model management system for IRT-based test construction decision support system. *Decision Support Systems, (27)* 4, 443–458, Netherlands: Elsevier.

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement, 28* (1), 37-47.