Graduate
Management
Admission
Council®

An Approach to Implementing Adaptive Testing Using Item Response Theory Both Offline and Online

Madan Padaki and V. Natarajan

MeritTrac Services (P) Ltd.

Presented at the CAT Research and Applications
Around the World Poster Session
June 2, 2009

Abstract

Many testing organizations conduct more admission and recruitment tests offline by paper and pencil than online on computers. Adaptive testing, when it was formulated in countries where computers were available for many years, did not encounter a problem of implementing computerized adaptive testing (CAT). Therefore, a need was felt for an alternative to CAT in India (particularly in MeritTrac) at a time when they conducted only about 10% of the tests online. An attempt was made to administer an adaptive test without computer delivery so that each examinee could benefit from a smaller number of items in a test matched to their ability. Since the ability of the examinee is unknown at the beginning of the test, the objective was to give a smaller number of test items, each item increasing (or decreasing) in difficulty from the previous item. The response pattern of the examinee for all the items is taken together with a condition that at least one item is answered incorrectly or one item is answered correctly. Ability estimation uses standard IRT methodology. This paper describes an attempt to plan, design, and implement such an adaptive test using paper and pencil. It uses standard IRT software to analyze a parent test of several items administered prior to the adaptive test and analysed for item difficulty of all the items in the parent test. In this case, the parent test had nearly 100 items and the adaptive test between 6 to 10 items. It is hoped that the research and implementation strategy outlined in the paper will help all those testing organizations eager to use adaptive testing without large-scale use of computers in their own organizations.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Padaki, M. & Natarajan, V. (2009). An approach to implementing adaptive testing using item response theory both offline and online. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

V. Natarajan, D-43, Ascent House, Sector – 6 NOIDA, India
Email: drvnatarajan@gmail.com

An Approach to Implementing Adaptive Testing Using Item Response Theory Both Offline and Online

The traditional method of determining the ability of an examinee, synonymous with number-correct scoring based upon his/her responses to a test, is neither accurate nor efficient. The superiority of scoring weight scores over number-correct scores in terms of providing a more accurate result, with very low standard errors, has been demonstrated in judging the ability of the examinee in comparison to much larger standard errors in the case of number-correct scores. Scoring weight scores are a product of classical test theory (CTT) analysis, but with the emerging role of computing technology in mental testing, there was a need to consider a new theory which came in the form of item response theory (IRT) that depended on computer technology. The source of its greater power is in the relationships it establishes between properties of the items and the operating characteristics of the test made up of the items.

The invention of IRT led to its application in various fields including a very important one in the field of assessment called adaptive testing**.** In today's assessment world there is always a need to devise newer, better, and more efficient and accurate methods. Adaptive testing, either in the form of computerized adaptive testing (CAT) or two-stage testing using paper-and-pencil instruments, enables testing time to be reduced to a half or a third of that required for a conventional test of the same precision, but ensures accuracy of the measurements.

With CAT, the examinee's ability level relative to a norm group can be iteratively estimated during the testing process and items can be selected based on the current ability estimate. Examinees can be given the items that maximize the information (within constraints) about their ability levels from the item responses. Thus, examinees will receive few items that are very easy or very difficult for them. This tailored item selection can result in reduced standard errors and greater precision with only a relatively small number of properly selected items.

In adaptive testing it is important to give the examinee a test with a small number of items with items of increasing item difficulty values starting from a very easy item and subsequently increasing difficulty. Normally (borne out of research) a 6- to 10-item test is administered compared to 25 or more items in the parent test.

Modes of Adaptive Testing

Online mode (computer-based mode). Adaptive testing can be put in an online mode using ONTRAC, which can randomly select items from an item bank according to the given specifications, with the help of pre-stored details about the item characteristics in the software. Calculations of an examinee's ability and true score can also be made by the software by addition of a software module enabling the successive approximation application to arrive at the final estimation of ability and, thus, true score. In other words, Student Tracking software is to be suitably modified (for 10 items) and added on as a module to ONTRAC. A copy of the software can be had from MeritTrac.

Offline mode (paper-and-pencil mode). Adaptive testing can also be put in a more practical offline/paper-pencil mode by making slight modifications in the process. With the help of item characteristics calculated earlier, a 6-item test with increasing item difficulty can be created on paper. The responses of the examinee can then be entered in the Student Tracking software specially created for this purpose, which will give an estimation of the examinee's true score as

if he has taken the parent 25-item tests or any other parent test.

Since it is not very feasible to conduct an online test everywhere, especially for a firm like MeritTrac which conducts a large and frequent number of offline tests on a pan-India basis, the importance of adaptive testing in offline mode increases many fold. As it only needs a single computer with Student Tracking software and pre-published papers consisting of items whose characteristics have been calculated on the basis of past responses. Thus the offline mode is much more practical and is as accurate as the online mode.

As an alternative, a table can be created that gives true scores for varying number-correct scores, primarily 1,2,3,4,5 correct (as the case may be from 1 to 9 for a 10-item adaptive test) Thus it is a fixed-length adaptive test.

Brief Procedure for Online and Offline Mode of Delivery

The (1,0 X) Excel format is input into BILOG, which generates the following output

1. PH-1 output (CTT analysis)
2. PH-2 output (item difficulty /threshold values)
3. PH-3 output (ability values)
4. PLT output (item characteristic curves)

Based on the above results about the characteristics of the various items in the parent test, many inferences can be drawn about the items of the test. The items are grouped into six groups according to their item difficulty:

Group I: Consists of items with item difficulty in the range of –3 or –4 to –2.

Group II: Consists of items with item difficulty in the range of –2 to –1.

Group III: Consists of items with item difficulty in the range of –1 to 0.

Group IV: Consists of items with item difficulty in the range of 0 to 1.

Group V: Consists of items with item difficulty in the range of 1 to 2.

Group VI: Consists of items with item difficulty in the range of 2 to 3 or 4.

Similarly, there are four categories of examinees used in adaptive testing par excellence, good, average, and poor. In terms of 10-item test, those who get 8 ,9 correct are par excellence , 6,7 good,4 and 5 are average, and below 3 are poor.

Requirements for Adaptive Testing

General requirements. These are the requirements for online as well as offline mode.

1. There should be prior information about the item characteristics of the items used in the 6- or 10-item adaptive tests. That is, there should be data regarding the responses to all the items in the test and that has to be run through BILOG to collect all the information.
2. The reshuffled format, depending on the output of item difficulty parameters, is to be made available.

3. There shall be only six groups of items and four categories of examinees, as illustrated earlier.

Specific requirements depending upon mode of delivery. These are the specific requirements depending upon the mode of delivery of the test.

Online mode:

1. In a traditional online testing of the parent test, a large number of computers would be needed to administer the parent test to a given group of examinees for a longer period of time. It should be noted that for handling the same group, a smaller number of computers would be required as the number of items and thus time taken is reduced considerably. It may also be viewed that at any testing center with a given number of computers, the online delivery of the adaptive test may handle a larger number of examinees for a given time .For example, a 25-computer MAZ (MeritTrac Assessment Zone) handling 100 examinees for a given session can now increase the number of examinees to 500 for the same session.
2. A suitable addition of a module as discussed earlier is to be on ONTRAC.

Offline mode:

1. For offline mode, Student Tracking software is needed in as many laptops or computers as feasible.
2. The coordinator needs to be extra cautious when dealing with Student Tracking software so that he does not make any mistake in entering the values of item numbers in the reshuffled version and the examinee's responses.
3. Alternatively, a table can be constructed which can be used as to read values of ability and true scores corresponding to (1,2,3,4,5 or 1to 9) correct.
4. Appropriate registration form is to be filled in by the examinee.
5. The coordinator has to enter the examinee's responses into the Student Tracking software and generate reports for every examinee.

A 5-item test in C++ was created on an organization adopting a DELFI process from an initial 25 items and finally arrived at five items in five important area in C++ with varying item difficulties.

BILOG was used to process the performance of this group of 10 examinees on 5 items and the item difficulties shown in Table 1were obtained

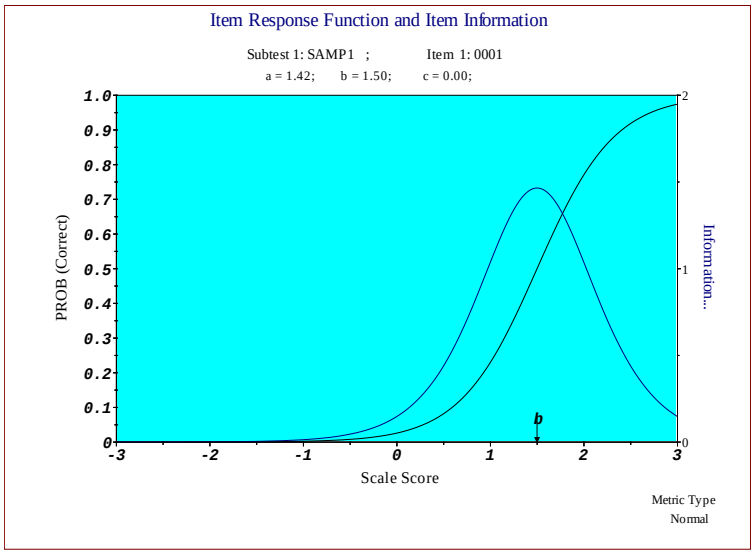
Table 1. Item Difficulties for Five Items

1	+1.499
2	+0.424
3	+2.292
4	+0.920
5	-0.279

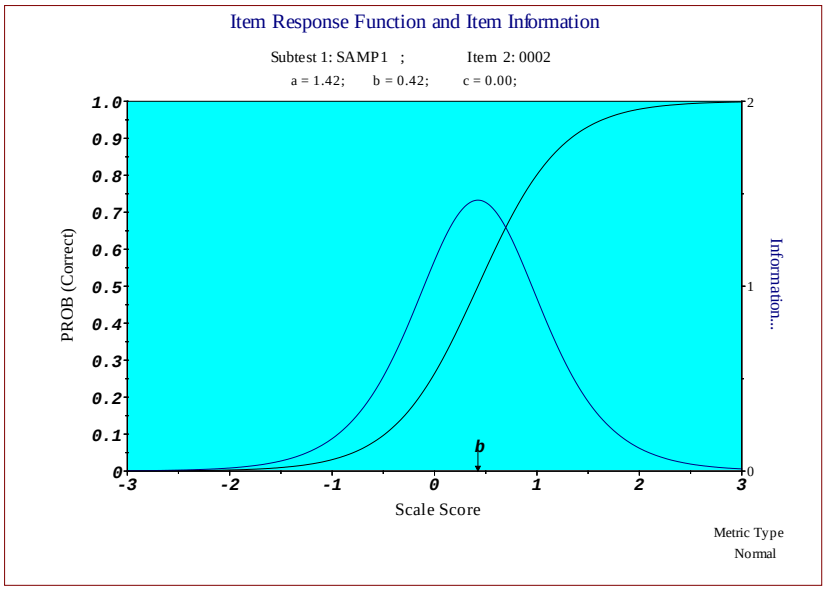
Figure 1 shows the item response functions for the 5 items. The graphs also give the information function for all the 5 items.

Figure 1. Item Response and Information Functions for Five Items

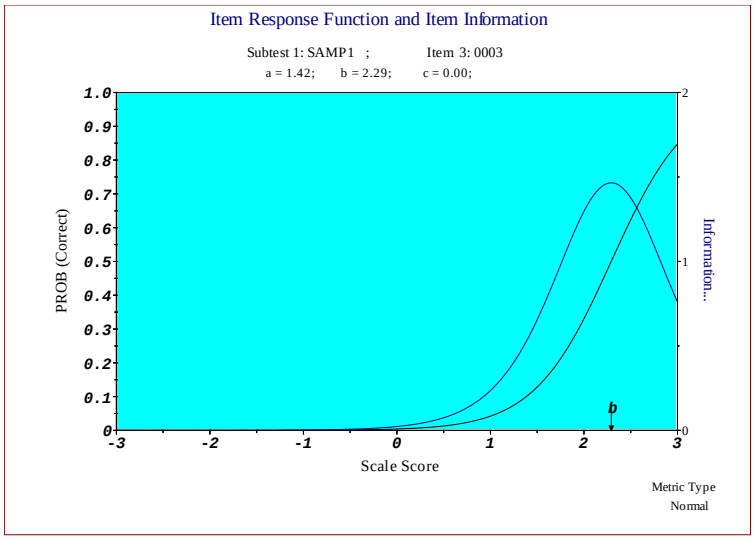
a. Item 1



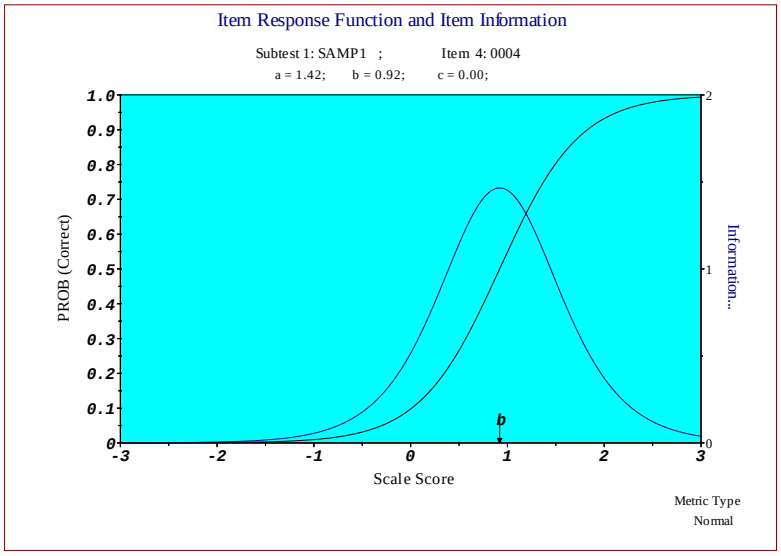
b. Item 2



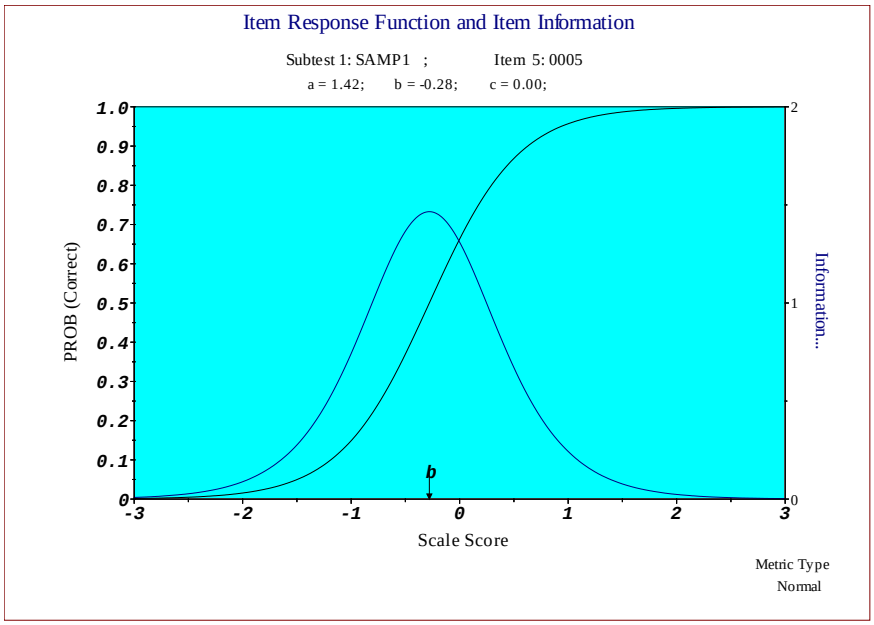
c. Item 3



d. Item 4



e. Item 5



From this 5-item parent test ,a 3-item adaptive test was made and administered to two examinees, X and Y, with X answering 2 out of 3 correct and Y answering 1 out of 3 correct. Their final abilities are estimated as shown in Tables 2 and 3.

Sample CAT θ Estimates

Consider two examinees, X and Y. For X the initial θ estimate was 1.5. This assumption was made on the basis that he answered Item 4 correctly with difficulty value of b = 0.92 and also answered Item 1 correctly with difficulty value of b = 1.5. He/she answered Item 3 incorrectly with difficulty value of b = 2.29. His/her final estimate was calculated as shown in Table 2 and that of Examinee Y in Table 3.

Table 2. θ Estimation for Examinee X Taking an Adaptive Test with 3 Items (Items 4, 1, and 3)

Item No	B	u	θ	–(θ – b)	$e^{-(\theta - b)}$	$p=1/(1+e^{-(\theta - b)})$	q = 1– p	u – p	p × q	Correction Factor	Next Estimate
4	0.92	1	1.5	-0.58	0.560	0.641	0.359	0.359	0.230	0.787	2.287
1	1.50	1		0.00	1.000	0.500	0.500	0.500	0.250		
3	2.29	0		0.79	2.203	0.312	0.688	-0.312	0.215		
								0.547	**0.695**		
4	0.92	1	2.287	-1.37	0.255	0.797	0.203	0.203	0.162	0.027	2.314
1	1.50	1		-0.79	0.455	0.687	0.313	0.313	0.215		
3	2.29	0		0.00	1.003	0.499	0.501	-0.499	0.250		
								0.017	**0.627**		
4	0.92	1	2.314	-1.39	0.248	0.801	0.199	0.199	0.159	0.000	2.314
1	1.50	1		-0.81	0.443	0.693	0.307	0.307	0.213		
3	2.29	0		-0.02	0.976	0.506	0.494	-0.506	0.250		
								0.000	**0.622**		

Table 3. θ Estimation for Examinee Y Taking an Adaptive Test with 2 Items (Items 5, 2)

Item No	B	u	θ	–(θ – b)	$e^{-(\theta - b)}$	$p=1/(1+e^{-(\theta - b)})$	q = 1– p	u – p	p × q	Correction Factor	Next Estimate
5	-0.28	1	0.3	-0.58	0.560	0.641	0.359	0.359	0.230	**-0.152**	**0.148**
2	0.42	0		0.12	1.127	0.470	0.530	-0.470	0.502		
								-0.111	**0.732**		
5	-0.28	1	0.15	-0.43	0.652	0.605	0.395	0.395	0.239	**-0.078**	**0.070**
2	0.42	0		0.27	1.313	0.432	0.568	-0.432	0.245		
								-0.038	**0.484**		
5	-0.28	1	0.07	-0.350	0.705	0.587	0.413	0.413	0.242	**0.000**	**0.070**
2	0.42	0		0.350	1.419	0.413	0.587	-0.413	0.242		
								0.000	**0.485**		

Thus, the ability of the Examinee X and that of test Examinee Y were calculated as 0.070 and 2.314, respectively, after administering the adaptive test to them. Their true scores can be calculated by taking these final ability values to the parent test of 5 items. The calculations are shown in Table 4.

Table 4. Calculation of True Scores of Examinees X and Y

Examinee	θ	Item 1 b = 1.5	Item 2 b = 0.42	Item 3 b = 2.29	Item 4 b = 0.92	Item 5 b = -.028	True Score
X	0.07	0.193	0.443	0.136	0.313	0.644	1.731
Y	2.314	0.693	0.568	0.152	0.317	0.645	2.374

Illustration of Adaptive Testing With the Analytical Ability Test

Over a period of time, more items were added to the bank and nearly 100 items were taken and the responses of 1,000 examinees were entered into BILOG and item difficulty values were generated. Finally, 93 items were relevant (7 items were discarded) and the parent test of 93 items eventually emerged. The items were grouped into six groups based on the item difficulty, as described above, and ten items were selected (one item each from Very Easy and Easy, and two items from Below Average, Above Average, Difficult, and Very Difficult), as shown in Table 5.

Table 5. Item Difficulty Ranges and Item Numbers

Description	Item difficulty Range	No. of items used in creating a form	Item Numbers
Very Easy	-4 / -3 to -2		Nil
Easy	-2 to -1	2	1,8
Below Average	-1 to 0	2	6,3,7,13
Above Average	0 to 1	2	16,18,17,4,20,15,14
Difficult	1 to 2	2	10,24,19,25,9,5,11,12
Very difficult	2 to ¾	2	21,23,2,22

- Combination 1 – 1,8,6,3,16,18,10,24,21,23
- Combination 2 – 1,8,7,13,17,4,19,25,2,22
- Combination 3 – 1,8,3,13,18,14,19,11,21,22
- Combination 4 – 1,8,6,13,17,20,5,12,23,22
- Combination 5 – 1,8,3,7,20,15,25,9,21,2
- Can create more such combinations

Several sets of 10-item adaptive tests are selected and administered to the examinees. Their responses to 10 items were categorized in terms of 9, 8, 7, 6, 5, 4, 3, 2, and 1 correct and a table, as shown in Table 6, generated from which ability and true scores can be read. Student Tracking software also could be used.

Table 6. Response Patterns, Ability Estimates, and True Scores for Ten Items

Responses	Final Ability	True Score
1,1,1,1,1,1,1,1,1,1	Not applicable	Not applicable
1,1,1,1,1,1,1,1,1,0	3.139	21.502
1,1,1,1,1,1,1,1,0,0	2.158	18.313
1,1,1,1,1,1,1,0,0,0	1.432	15.276
1,1,1,1,1,1,0,0,0,0	0.7896	12.393
1,1,1,1,1,0,0,0,0,0	0.1821	9.709
1,1,1,1,0,0,0,0,0,0	-0.4172	7.276
1,1,1,0,0,0,0,0,0,0	-1.037	5.114
1,1,0,0,0,0,0,0,0,0	-1.735	3.208
1,0,0,0,0,0,0,0,0,0	-2.685	1.518
0,0,0,0,0,0,0,0,0,0	Not applicable	Not applicable

Limitations of Adaptive Testing

Whatever is the mode of delivery, there are certain general limitations in adaptive testing, and some of these are

1. Balancing the content. Since there is no referencing of content, it is not possible to have a representative sample of the content. This can be overcome by having an item bank with both content sorting and difficulty sorting and combining them together into groups, e.g., content domain C1 will have the same grouping as item difficulty. When items are to be chosen from groups I, II, III, etc., then choose the items from the set of groups, and it will take care of both the difficulty and content.
2. Limitations on termination rule. Even though there are many practices for terminating adaptive tests, namely the number of test items, the time available at the disposal of the examinee and administrator, and method of calculating the final ability and true score, this will act as a limitation.
3. A very high level of training for the coordinator is required and the number of coordinators required for an event may increase slightly.

Specific limitations of the online (computer-based) mode:

4. Through its software, ONTRAC should be able to choose items of increasing difficulty successively.
5. ONTRAC should be able to do built-in calculations taking into account items taken

and responses made by the examinee. BILOG and Student Tracking software must be integrated with ONTRAC.

6. An initial estimate should be made about the examinee's ability (either through self-rating or depending upon the number of incorrect responses , as detailed above).

Specific limitations of the offline (paper-and-pencil) mode:

7. There should be absolute control of an event by the coordinator in terms of getting the registration form filled properly and entering the data properly in the Student Tracking software, as detailed out in the registration form.

8. The time taken to respond by the examinee and the time taken by the coordinator to enter data will be a determining factor in cost saving and time saving in the operation.

9. There may be a necessity to have more coordinators per event in comparison to the online mode.

Areas of Application of Adaptive Testing in the Future

Achievement testing. Following increasing demand for very accurate measurement, evaluation, and assessment, school boards, universities and other certifying organizations are driven to the use of shorter tests with increased accuracy and efficiency, paves the way for the application of IRT and adaptive testing.

Recruitment testing. Recruiting agencies in India are being compelled to make use of shorter tests and shorter testing times which at the same time yield better and more accurate results. They are also increasingly introducing computer-aided adaptive testing modules and other online instruments. Even though its application in India to recruitment has just begun, there are increasing opportunities for recruitment in BPO and ITES sectors where a very large number of aspirants compete for a smaller number of positions.

Mastery testing. It is found in many assessment scenarios that in certification or categorization of achievers in a dichotomous fashion , such as traditional pass/ fail, selected/ rejected and master/ non-master, mastery testing is used. This is a result of Carroll and Bloom's research work for years that yielded the concept of mastery testing and mastery learning. This is a particular case of an achiever who can be certified as a master or non-master. A usual level of mastery is prescribed as 90/90 which indicates 90% of examinees will score 90% on the test. There are also situations where 100/100 is required, particularly in a nurses certification test where 100% mastery is required (for example, a nurse is to be certified for distinguishing between poisonous and non-poisonous materials).

Scholarships and other award testing. IRT calibrated items in relation to the item characteristics of item difficulty, item discrimination, and item guessing provide a platform to constitute a special test for award of scholarship or any other excellence awards. It has to be understood that items in a bank calibrated with IRT parameters will enable sorting out the items in the order of increasing difficulty and discrimination and, as far as possible, we can sort out items with high difficulty values ranging from 2 to 2.5 and with very high to perfect discrimination so that they can serve the requirements for a scholarship test. Those who perform very well on these items, and their ability estimates and true scores are beyond the accepted cut off score for award or scholarship may be selected for such awards. IRT thus provides a test with items aimed at a cut off of difficulty and discrimination required for such awards. This is an

application which is worth attempting for such awards as award of foreign scholarships.

Diagnostic testing. Educators and trainers all over the world are increasingly providing feedback to students and trainees on the strengths and weaknesses of their performances in terms of content areas, abilities, and skills tested and levels of difficulty of items. It is, therefore, possible with IRT calibration and coding adopted for items in terms of content, ability cluster, and difficulty levels, to sort out items from different content areas, from different clusters, and from different levels of difficulty. It shall then be possible to generate a feedback that will list strengths and weaknesses with respect to selected contents, selected clusters, and selected levels of difficulty. Thus, weaknesses in these areas can be diagnosed and on the basis of these remedial steps can be recommended. This is an area of application that needs to be tried at all levels of education and training, in particular by teachers and trainers, on a continuous basis.

Conclusions

1. One basic conclusion is that adaptive or tailored testing is an extremely promising application area of IRT.
2. No doubt, there is a set of pre requisites ,which are: a parent test, calibration of test items through IRT (BILOG), rearranging the items by difficulty and content, choosing items for an examinee when he/she rates himself, or administering a pre-designed 6- or 10-item adaptive test, which are nothing but salient features of adaptive testing.
3. It is imperative that many clients, who are unable to allocate a longer period for testing and assessment, will genuinely accept adaptive testing. Of course, it is vital to bring the salient features of adaptive testing to the clients and thus the necessity to demand increased per examinee cost.