DOCUMENT RESUME

ED 322 729 FL 018 734

AUTHOR	Laurier, Michel
TITLE	What We Can Do with Computerized Adaptive Testing...and What We Cannot Do!
PUB DATE	Apr 90
NOTE	18p.; Paper presented at the Annual Meeting of the Regional Language Center Seminar (Singapore, April 9-12, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Viewpoints (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Foreign Countries; French; Item Banks; *Item Response Theory; Language Tests; *Scoring; Statistical Analysis; *Test Construction; Testing; Test Items; *Test Use

ABSTRACT

Computerized adaptive testing for langauge teaching and learning takes advantage of two properties of the computer: its number-crunching and multiple-branching capabilities. Adaptive testing has also been called tailored testing because it aims at presenting items that suit the student's competence and that are informative, using an item bank and an algorithm for selecting items. The most widely used theoretical framework for these components is item response theory, which allows for the measurement of the trait that corresponds to the subject's ability. Creation of the item bank includes planning, field testing, item analysis, calibration, and item inclusion. Test bank management requires additional features for updating, importing items, listing items, and obtaining item information. The item selection procedure estimates the examinee's ability after an answer and finds the next item that is most appropriate. Test administration involves gathering information about the examinee, a self-rating of general proficiency, and sub-tests using the item selection procedure. Results are available immediately, an advantage appreciated by examinees and administrators. The adaptive procedure allows for shorter tests, and the items are never too difficult or too easy. Limitations include the artificiality of the computer environment, restriction of answer type and test content, non-applicability for small-scale testing, the theoretical assumption of unidimensionality, and cost. (MSE)

Paper to be presented at the RELC Seminar, Singapore, April 1990

What we can do with Computerized Adaptive Testing...

And what we cannot do!

Michel LAURIER

Carleton University (Canada)

Among numerous applications of computers for language teaching and learning there is a growing interest for a new acronym: CAT which stands for Computerized Adaptive Testing. CAT can be seen as the second generation of computerized tests (Bunderson, Inouye & Olsen 1989). The first generation consisted of conventional test administered by computers; further generations will be less obtrusive and will provide constant advice to the learners and teachers. In this paper we shall attempt to explain how CAT works and what is the underlying theory. The various steps involved in implementing an adaptive test will be described with examples from a placement test that we have developed in French.

Principles of adaptive testing

Computers in testing are particularly useful because of two advantages over conventional testing methods:

- number-crunching capabilities: Conventional marking systems often means counting the number of right answers or converting a score with a pre-set scale. Using a computer allows more complex assessment procedures right after the test or even during the test. These calculations may use the data that is available more efficiently. In addition, computers are fast and virtually error-free.
- multiple-branching capabilities: Using "intelligent" testing systems, some decisions can be made during the administration of the test. The computer can analyze students' responses and decide which item will be submitted, accordingly. Therefore, the inherent linearity of a conventional test is no longer a limitation.

CAT takes full advantage of these two properties of the computer.

Let's suppose we want to assign a student to a group that would suit his needs by means of a conventional placement test. We do not know a *priori* at which level the student could be placed; he/she could be an absolute beginner in the language or an "educated native". In this case, the test should probably include some difficult items, as well as some easy ones. In fact, given the student's level, how many of the items of a two hour test are relevant? Probably less than 25%. Some of the items will be too easy, particularly if the student is at an advanced level. From the student's point of view, those items are boring, unchallenging; from the psychometric point of view, they do not bring valuable information because the outcome is too predictable. On the other hand, some items will be too difficult, particularly for beginners who will feel frustrated because they find that the test is "over their heads"; again, there is very little information on the student's level that can be drawn from these items.

Adaptive testing has also been called "tailored testing" because it aims at presenting items that suit the students' competence and that are informative. In an open-ended test, this means items in which the chance to answer correctly will be approximately fifty/fifty. This approach to testing problems might bring to mind Binet's multi-stage intelligence tests that were developed at the beginning of the century. For language teachers, it may also resemble recent oral interview procedures in which the examiner is encouraged to adapt the exchange to the examinees' performance (Educational Testing Service 1985). Adjusting the test is in fact a complex process that CAT seeks to replicate. For this task, we need:

- an item bank: a collection of items stored with some specifications and measuring the same ability at different levels.
- a selection procedure: an algorithm which will choose and retrieve the most appropriate item at a given moment, with a given examinee.

Item Response Theory

Although different theoretical frameworks could be applied to set up the item bank and the selection procedure, the most widely used is the Item Response Theory (IRT). Despite its mathematical complexity, IRT is conceptually attractive and very interesting for CAT. The theory was first labeled "latent trait theory" by Birnbaum (1968) because it assumes that a test score or a pattern of answers reflects a single construct that is not directly observable. What the test measures is known as the "trait" and corresponds to the subject's ability. The theory was refined by F. Lord who studied the "Item Characteristic Curve" (Lord 1977). "An item characteristic curve (ICC) is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it" (Hambleton and Swaminathan 1985:25). If we plot the probability of answering correctly against the examinees' ability, the curve should rise as the ability level increases. Thus, the probability of having a right answer at the advanced level will be very high but should be very low at the beginner's level. The ability is expressed in terms of standard deviations and ranges from roughly -3 to +3. Figure 1 shows the curve for an "Intermediate" level item. The inflection point of this ICC is around 0 which corresponds to the sample mean. Since the subject's ability and the item difficulty are expressed on the same scale, we say that the difficulty of the item (the parameter *b*) is 0.

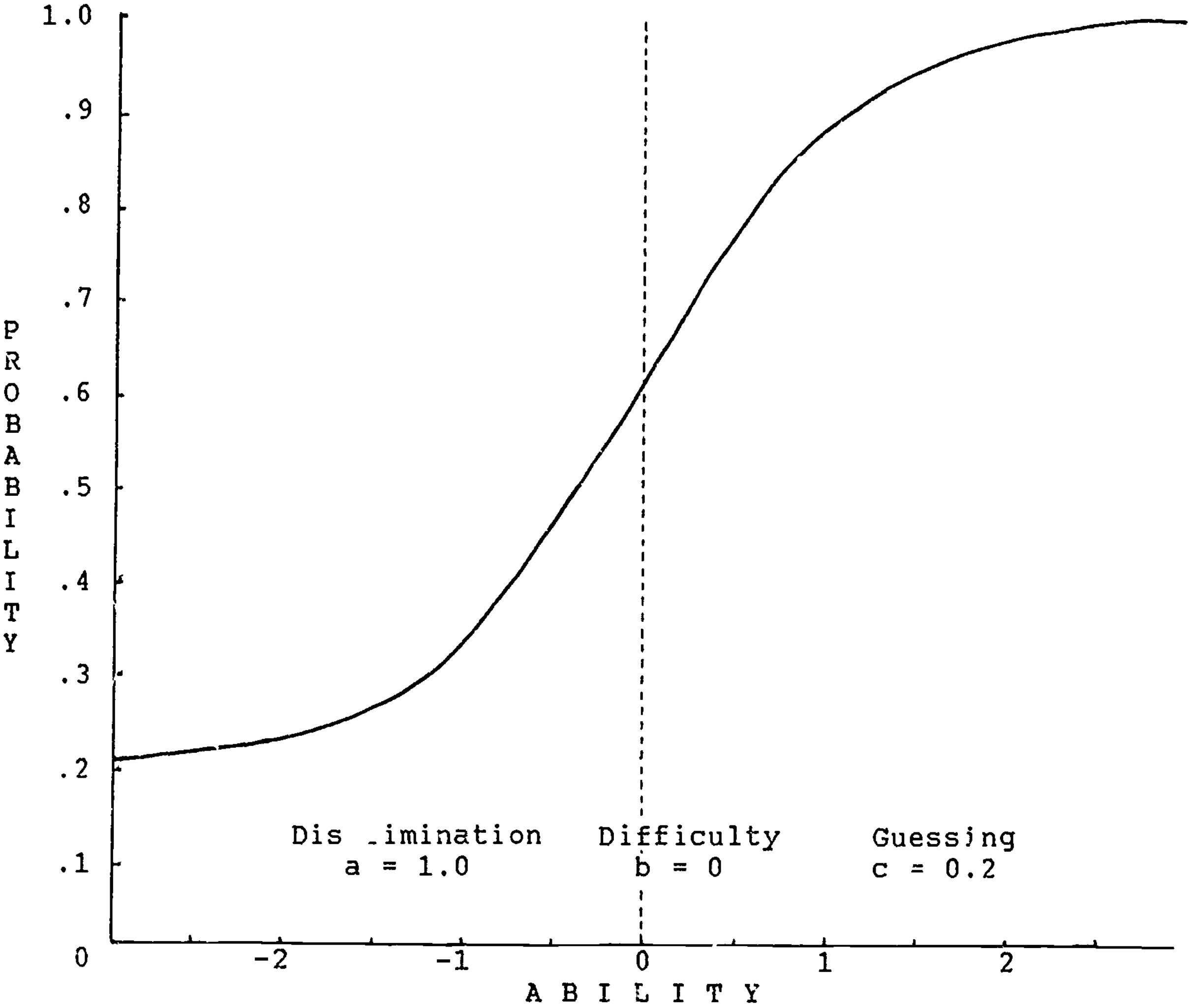

Figure 1: Item Characteristic Curve

If an item clearly separates the advanced students from the beginners the curve should be very steep; if it does not, the curve will be flatter. In other words, the slope of the ICC corresponds to the discrimination (the parameter *a*). An item with a discrimination index of 1 or more is a very good item. Finally, we see that, in this particular case, the curve will never reach the bottom line. This is due to the fact that the item is a multiple choice question which involves some guessing. This is expressed with a third parameter (parameter *c*). A m/c item with five options should have a c around .2. Of course, in reality, such a regular curve is never found. The degree to which the data for an item conforms to an ICC is the "item fit". Misfitting items should be rejected.

Once the parameters are known, we can precisely draw the ICC using the basic IRT formula

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where θ (theta) represents the subject's ability and D a scaling constant set at 1.7. A simpler formula for a less complex but generally less accurate model has been proposed by G. Rasch (1960). The Rasch model is a one-parameter model; it assumes that there is no guessing and that all the items discriminate equally. Under this model, only the difficulty has to be estimated.

The parameter estimation is a complex mathematical procedure that requires a computer. Various programs are available either on mainframe computers (v.g LOGIST, Wingersky, Barton & Lord 1982) or micro-computers (v.g. MicroCAT, Assessment Systems Corp. 1984). To estimate the parameters properly, particularly with the three-parameter model (discrimination, difficulty and guessing) a large sample is needed - about 1,000 examinees. Fortunately, the distribution of the sample does not have to reflect exactly the distribution of the population because the program will try to fit a curve rather than calculate proportions of correct answers. The item calibration is sample-free. This property of IRT models is known as the "invariance of items".

IRT provides also the "invariance of subjects" which means that we get test-free person measurement. This second property is crucial in adaptive testing because it implies that ability estimates can be calculated and compared even though different items have been submitted.

Implementation of the test

The following steps are involved in creating the item bank:

- Planning the bank: Are we measuring more than one common trait? If so, then several item banks should be set up. At this stage, we must also make sure that the items can be administered, answered and marked both with a "paper-and-pencil" format and with a computerized version. Since field testing is expensive, a great deal of attent'on must be paid to the wording of the items. For large item banks, several versions using "anchor items" will be necessary.
- Field testing and item analysis: The items will be tried out on a small sample - 100 to 200 subjects. Classical item analysis using proportions of correct answers and correlations is helpful in order to eliminate bad items from the next version. At this stage, some dimensionality analysis can be conducted to make sure the test (or sub-test) is measuring a single trait.
- Field testing and calibration: The new version(s) is(are) administered to a large sample - 200 to 2,000 depending on the model chosen and the quality of the sample. This data will be

processed so that item parameters and degree of fit will be obtained for each item.

- Inclusion to the bank: If the item is acceptable, it will be added to the bank. At least, an identification code, the questions (and the options with multiple-choice items), the right answer and the parameters should appear on an item record. Additional information may be incorporated (Henning 1986).

Of course, a management system will have been previously set up. A management system works like a data base system. Each sub-test is a data base that can be accessed with the management system. Once the user has chosen a sub-test, different operations can be executed:

- Updating the bank: new items may be added, some others deleted. The user should also be able to browse in the bank and modify a item without having to rewrite it.
- Importing items: When a set of items are located in another file, there should be provisions to execute a mass transfer into the bank.
- Listing the items: Each item can been seen individually on the screen. Yet the user can also call a list of the items. Each line will show the identication code of an item, the parameters, and a cue to remind the question. In addition, our system calculates the "Match index". According to Lord (1970), this value corresponds to the ability at which the item is the most efficient.
- Obtaining the item information: Under IRT, one can tell how much information can be obtained at different points of the

ability scale. As the information sums up, at a specific ability point, the estimation becomes increasingly more reliable at this point.

The selection procedure is a method that can be applied in order to estimate the examinee's ability after an answer and to find the next item that is the most appropriate. The concept of item information is crucial as the most appropriate item is the one that brings the most information for a given ability. Tracing the administration of the adaptive test we have designed, will help to understand how the program works. We needed a computerized placement test for English speaking post-secondary students learning French as a second/foreign language in Canada. As a placement test, the instrument attempts to assess the student's general proficiency. It assumes that such a construct exists even though a more refined evaluation should probably divide this general competence in various components such as the grammatical competence, the discourse competence or the sociolinguistic competence (Canale and Swain 1980). The format of the test is affected by the medium, the micro-computer. The three sub-tests contain multiple-choice items because we want to minimize the use of the keyboard and because open-ended answers are too unpredictable to be properly processed in this type of test. The organization and the content of the test also reflect the fact that we had to comply with IRT requirements.

The administration of the test

Within the IRT framework, procedures have been developed to estimate the student's ability, using the answers to the items and the parameters of these items. However, calculating the student's ability is not possible when the program is started since no data is available. This is the reason why, at the beginning of the test, the student will be asked some information about his/her background in the second/foreign language:

How many years did the student study the language?

Did he/she ever live in an environment where this language is spoken?

If so, how long ago?

Then the program prompts the student to rate his/her general proficiency level on a seven category scale ranging from "Beginner" to "Very advanced". All this information, is used to obtain a preliminary estimation that will be used for the selection of the first item of the first sub-test. Tung (1986) has shown that the more precise is the preliminary estimation, the more efficient is the adaptive test.

The first sub-test contains short paragraphs followed by a m/c question to measure the student's comprehension. According to Jafarpur (1987), this "short context technique" is a good way to measure the general proficiency. Figure 2 illustrates how the adaptive procedure works. At the beginning of the sub-test, after an example and an explanation, an item with a difficulty index close to the preliminary estimation is submitted.

Item	U	a	b	c	Score	Theta	Info.	Error
CO23	0	1.212	-0.702	0.230	0/1	-0.750	?	?
CO27	0	0.982	-0.819	0.231	0/2	-0.950	?	?
CO41	1	0.909	-0.930	0.264	1/3	-1.833	0.338	1.719
CO37	1	1.346	-1.109	0.219	2/4	-1.129	1.948	0.716
CO32	1	0.967	-1.109	0.180	3/5	-0.894	2.685	0.610
CO22	0	1.005	-0.568	0.250	3/6	-1.070	2.752	0.603
CO34	1	0.807	-0.905	0.228	4/6	-0.946	3.269	0.553
CO30	0	1.220	-0.809	0.198	4/7	-1.148	3.408	0.542

Figure 2 - Items used in sub-test #1

In the example, the first item was failed (U = 0) and the program then selected an easier one. When at least one right and one wrong answer have been obtained, the program uses a more refined procedure to calculate the student's ability. The next item will one which has not been presented as yet and that is the closest to the new estimation. The procedure goes on until the pre-set threshold of information is reached. When this quantity of information is attained, the measure is precise enough and the program switches to the next sub-test.

The same procedure is used for the second part with the estimation from the previous sub-test as a starting value. On the second sub-test, a situation is presented in English and followed by four grammatically correct statements in French. The student must select the one that is the most appropriate from a semantic and sociolinguistic point of view. Raffaldini (1988) found this type of situational test a valuable addition to a measure of the proficiency. Once we have obtained sufficient information, the program switches to the third sub-test, which is a traditional fill-the-gap exercise. This format is found on most of the current standardized tests and is a reliable measure of lexical and grammatical aspects of the language. Immediately after the last sub-test, the result will appear on the screen. Since a normal curve deviate is meaningless for a student, the result will be expressed as one of the fourteen labels or strata that the test recognizes along the ability range: "Absolute beginner, Absolute beginner +, Almost beginner ... Very advanced +".

Advantages and limitations

Both the students and the program administrators appreciate that the result is given right away. The students receives immediate feedback on what he/she did and the result can be kept confidential. Since there are no markers, the marking is economical, error-free and there is no delay. Individual administration as opposed to group administration is, in some situations, an asset: the students can write the test whenever they want, without supervision. Because of the adaptive procedure, the tests are shorter. In order to reach a comparable reliability with our test, we need a "paper-and-pencil" version that is at least twice as long as the CAT one. Actually, in most cases, CAT will use only 40% of the items of the equivalent conventional test. Finally, the adaptive procedure means that the student is constantly faced with a realistic challenge: the items are never too difficult or too easy. This means less frustration, particularly with beginners. With a more sophisticated instrument than the one we designed, one could even find other positive aspects of CAT. For example, with IRT it is possible to recognize misfitting subjects or inappropriate patterns and therefore detect phoney examinees. Taking advantage of the capabilities of the computer, one could also make the testing environment more enjoyable.

However, there are also very serious limitations with CAT. Even with the fanciest gadgetry, the computer environment will always be a very artificial one. It is always a remote representation of the real world and precludes any form of direct testing. Moreover, the type of answer is restricted because of the machine itself and because of the psychometric model. With the combination of the present technology and IRT, it is hard to imagine how a test could use anything other than m/c items or very predictable questions. The medium, the computer, not only affects the type of answers but also the content of the test. In our test, we wanted to use standard and affordable hardware but some students complained that the test was very poor in assessing oral skills. In spite of recent innovations with videodiscs, audio-tape interfaces, CD-Rom, or even artificial speech devices, the stimulus in CAT is generally written. On the other hand, the model, IRT, not only affects the type of answer but also the practicality of the development. In our test, three parts of fifty items each were administered to more than 700 hundred examinees. This is considered as a minimum and some research shows that even with 2,000 examinees, the error component of a three-parameter calibration may be too large. Using a Rasch model may help to reduce the sample size, usually at the expense of the model fit, but the field testing will always be very demanding. Therefore, CAT is certainly not applicable to small-scale testing.

Perhaps the most formidable problem, is the assumption of unidimensionality. This concept refers to the number of traits that are measured. Under IRT, a common dimension, i.e a single factor, must clearly emerge. Otherwise, applications of IRT may be highly questionable. Even though the calibration procedure is statistically quite robust and most language tests will comply with the unidimensionality requirement (Henning, Hudson & Turner 1985), many testing situations are based on a multidimensional approach of the language competence (Bachman, forthcoming). Multi-dimensional calibration techniques exist but they are not always practical (Dandonelli & Rumizen 1989). One particular type of unidimensionality is the independence of the items. This principle implies that an answer to one item should never affect the probability of getting a right answer on another item. Cloze tests usually do not meet this requirement because finding a correct word in a context increases the chance of finding the next word.

Finally, when all the theoretical problems have been solved some practical problems may arise. For example, for many institutions the cost of the development and implementation of an adaptive test could be too high. Madsen (1986) studied the student's attitude and anxiety toward a computerized test; attention must be paid to these affective effects.

Conclusion

These limitations clearly indicate that CAT is not a panacea. It should never be used to create a diagnostic test that aims at finding weaknesses or strengths on various discrete points because this type of test is not unidimensional. By the same token, it should not be used on so-called "communicative" tests that attempt to measure aspects of the communicative competence without isolating the different dimensions in separate sub-tests. Canale (1986) mentions that the testing environment is so artificial that CAT lacks validity when test results are used to make important decision - for a certification test, for instance.

However if only a rough estimation over a wide range of ability is needed, for placement purpose for example, CAT may be a very adequate solution. It is also appropriate if the trait being measured is unique such as general proficiency, vocabulary, grammar... It could also be a solution to testing problems for some integrative tests of receptive skills particularly if the result will not affect the student's future or can be complemented with more direct measures.

In short, a CAT will always be a CAT, it will never be a watchdog.

NOTES

[1] For an excellent introduction to IRT, see Baker (1985)

[2] An experimental version of this test has been developed at the Ontario Institute of Studies in Education (Toronto) and will be implemented at Carleton University (Ottawa).

REFERENCES

Assessment Systems Corporation (1984) *User's Manual for the MicroCAT testing system*. St.Paul, MN.

Bachman L.F. (1989) *Fundamental considerations in language testing*. London: Oxford University Press.

Baker F.B. (1985) *The basics of item response theory*. Portsmouth, NH.

Birnbaum A. (1968) Some latent trait models and their use in infering an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bunderson C.V., Inouye D.K. & Olsen J.B. (1989) The four generations of computerized educational measurement. In R.L. Linn(Ed.) *Educational Measurement* 3rd ed. (pp. 367-408) New York: American Council on Education - Macmillan Publishing.

Canale M. (1986) The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.) *Technology and language testing* (pp. 29-46). Washington, D.C.: TESOL.

Canale M. & Swain M. (1980) Theoritical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Dandonelli P. & M. Rumizen (1989) There's more than one way to skin a CAT: Development of a computer-adaptive French test in reading. Paper presented at the CALICO Conference, Colorado Spring, CO.

Educational Testing Service (1985) *The ETS Oral Interview Book*. Princeton, NJ.

Hambleton and Swaminathan (1985) *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Henning G. (1986) Item banking via DBase II: the UCLA ESL proficiency examination experience. In C. Stansfield (Ed.) *Technology and language testing* (pp. 69-78). Washington, D.C.: TESOL.

Henning G., Hudson T. & Turner J. (1985) Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2*, 141-154.

Jafarpur A. (1987) The short-context technique: an alternative for testing reading comprehension. *Language Testing, 4*, 133-147.

Lord F.M. (1970) Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance* (pp. 139-183) New York: Harper & Row.

Lord F.M. (1977) Practical application of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Madsen H. (1986) Evaluating a computer adaptive ESL placement test. *CALICO Journal, December*, 41-50.

Raffaldini T. (1988) The use of situation tests as measure of communicative ability. *Studies in Second Language Acquisition, 10*, 197-215.

Rasch G. (1960) *Probabilistics models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Tung P. (1986) Computerized adaptive testing: implications for language test developers. In C. Stansfield (Ed.) *Technology and language testing* (pp. 11-28). Washington, D.C.: TESOL.

Wingersky M.S., Barton M.A. & Lord F.M. (1982) *LOGIST* user's guide. Princeton, NJ: Educational Testing Service.