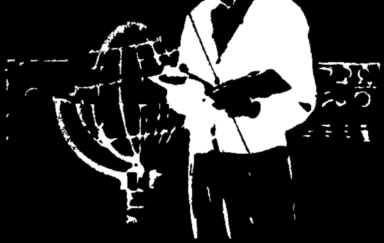
DOCUMENT RESUME

ED 118 577	TM 004 969
AUTHOR	Hansen, Duncan N.
TITLE	Adaptive Testing as a Significant Process in AIM.
INSTITUTION	Memphis State Univ., Tenn. Bureau of Educational Research and Services.
PUB DATE	1 May 75
NOTE	29p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS	*Computer Oriented Programs; *Individualized Programs; Instructional Systems; *Measurement Techniques; Models; Response Style (Tests); Scoring; *Student Ability; *Student Testing; Test Construction; Test Interpretation; Test Reliability; Test Selection; Test Validity
IDENTIFIERS	*Adaptive Testing; Computer Assisted Testing; Flexilevel Tests; Tailored Testing

ABSTRACT

To what degree testing can become adaptive is considered in three ways: from a formal methodological perspective; from a human process, stability, perspective; and from a sub-system or component view within an adaptive instructional system (AIS). With the advent of large computer-based training systems, the opportunity to broadly implement adaptive testing models and contrast them in terms of their adaptive nature has come to its moment of truth. It, therefore, seems appropriate to describe various computer paradigms which are representative of one or more models. This completes the first third of this paper. Testing has long been considered adaptive if the situation is made easier or more relaxing for the student. As this paper illuminates, it is perhaps more important to increase the challenging aspects of the test adaptation, even to stressing characteristics in order to improve both reliability and validity. Adaptive testing can be considered within the context of a total AIS framework. To what degree does it provide for time savings and for enhanced systems improvement? It is in this last area that so little experience and data are available. What little data and conjecture that can be accumulated at this time is presented to complete the overview of adaptive testing. (Author/RC)



ADAPTIVE TESTING AS A SIGNIFICANT PROCESS IN AIM

TM

Prepared by

Duncan N. Hansen

Bureau of Educational Research and Service
College of Education
Memphis State University
Memphis, Tennessee 38152

May 1, 1975

ADAPTIVE TESTING AS A SIGNIFICANT PROCESS IN AIM

by

Duncan N. Hansen

1.0 Introduction

Generic to any adaptive instructional system (AIS) is the testing-evaluation process. Given the goal of adapting the overall instructional learning process, it seems only natural to ask, to what degree can testing become adaptive? For the purposes of this paper, this question can be considered in three ways. First, from a formal methodological perspective; second, from a human process, stability, perspective; and third, from a sub-system or component view within an adaptive instructional system.

In reference to the formal psychometric models it has long been known that many test items (too hard or too easy) provide little or no information concerning the outcome decision to be made about the student. If this is the case, then it seems only natural to find some appropriate way for removing these test items without detracting from either the reliability or validity of the assessment instrument. The vast majority of adaptive testing models formally address only this problem. From a systems point of view, these models have received little or no empirical investigation. With the advent of large computer-based training systems, the opportunity to broadly implement adaptive testing models and contrast them in terms of their adaptive nature has come to its moment of truth. It, therefore, seems appropriate to describe various computer paradigms which are representative of one or more models. This will complete the first third of this paper.

As a student is presented with test items via either pencil and paper or some electronic device such as a CRT terminal, he is involved in a complex behavioral process. The testing itself presents certain kinds of characteristics. It has long been considered adaptive if we can make a situation easier or more relaxing for a student. As this paper will try to illuminate, it is perhaps more important to increase the challenging aspects of the test adaptation, even to stressing characteristics in order to improve both reliability and validity. Thus, the very nature of adaptation as a behavioral process interacting with a dynamic testing algorithm may change our thoughts and views of the environmental conditions for optimality. Fortunately, the indices of reliability and validity directly answer these issues.

Finally, adaptive testing can be considered within the context of a total AIS framework. To what degree does it provide for time savings and for enhanced systems improvement? It is in this last area that we have so little experience and data. What little data and conjecture that can be accumulated at this time will be presented to complete the overview of adaptive testing.

2.0 Adapting Testing Models for Instructional Systems

Adapting testing models (ATM), while interesting from a theoretical point of view are, in fact, only as important as the overall adaptive instructional system (AIS) into which they are embedded. Recognizing that adaptive instruction is to be contrasted with more conventional or individualized approaches, each AIM approach tends to stress characteristics of (1) being adapted to the specific characteristics of each student from both a class and state variable viewpoint; (2) to provide instruction in some systematically contingent fashion; (3) to mediate the information flow so as to optimize the learning rate and outcomes, and (4) to provide empirical feedback, most importantly, to the system so as to allow it to approximate its ultimate state of optimality. As a framework for understanding the role of adaptive testing, Figure 1 presents a flow of how an adaptive instructional system would work. For our testing purposes the critical areas are found in Step 1, Step 8 and most importantly in Step 10. Allow me to elaborate.

First, the initial steps indicate how all of the a priori information on a given student is considered and then is matched within the consideration of tasks, instructional alternatives and the students' data profile. From this, an instructional decision rule, sometimes referred to as adaptive instructional model, is selected and applied. This is scheduled and the instruction is prescribed. After it has been implemented it receives an immediate evaluation. This evaluation both provides feedback to the student's learning profile as well as to the overall system as represented in the parameters found in the adaptive instructional models. Thus, Step 1, the student's learning profile, is an update of his immediate prior performance, his learning time, and other associated learning indices, as well as associated behavioral patterns, be they adaptive or personality in nature.

The composition of an instructional prescription is critical in that this represents the point of closure by which the objectives and criterion level are formulated for a student. In Step 10, this information is utilized as entry information into the testing process. The

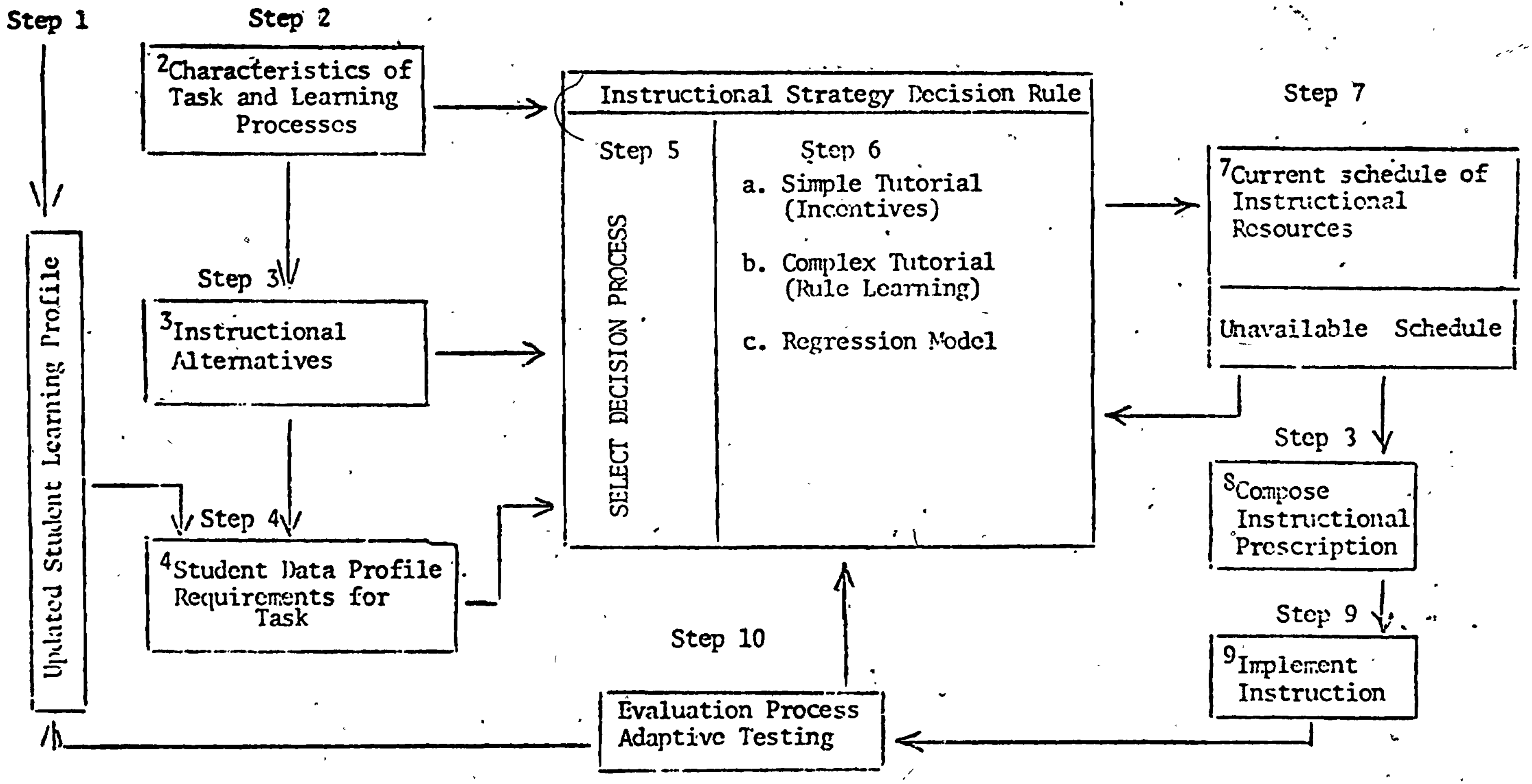

Step 1
Step 2
2Characteristics of Task and Learning Processes
Step 3
3Instructional Alternatives
Step 4
4Student Data Profile Requirements for Task
Updated Student Learning Profile
Instructional Strategy Decision Rule
Step 5
SELECT DECISION PROCESS
Step 6
a. Simple Tutorial (Incentives)
b. Complex Tutorial (Rule Learning)
c. Regression Model
Step 7
7Current schedule of Instructional Resources
Unavailable Schedule
Step 3
8Compose Instructional Prescription
Step 9
9Implement Instruction
Step 10
Evaluation Process Adaptive Testing

Fig. 1. Adaptive Model(s) Program Flowchart.

testing process then consists of the presentation under some appropriate algorithm of a series of items which are scored in real time, and a decision is made. Outcomes of the actual test performance are then utilized to update both the individual student record and the system. We turn now to the details of this adaptive testing process.

2.1 Test Entry Processes.

The testing process (Step 10) can be characterized by three sub-processes: (a) appropriate test selection and student entry, (b) tailored presentations of the test items and (c) sensitive scoring and diagnosis, interpretation and reporting. For the entry process, it is intuitively and empirically obvious that the test or composited test items should be selected to maximize the accuracy and meaningfulness of the outcome decision. In addition, a student should be entered into the test so as to minimize both trivial items and highly difficult or impossible items while focusing on the presentation of those items that best reflect the student's current learning competencies and provide for appropriate discrimination among the alternatives to be considered within the testing decisions. Therefore, any adaptive test selection and entry process would have to be based on the student's characteristics to be valid.

The research area of computer selected and/or composed tests is practically nonexistent. Wood (1971) reviewed the techniques for computer-composed tests. The Naval CMI project (1973) at Memphis illustrates how students can be routed to specific tests. Adaptive selection of tests remains a highly promising topic for future research. Rasch (1969) provides a model that yields equivalent individual measurement (scores) from sets of items varying in difficulty. Hasang (1972) proposed a procedure for item weighting to achieve invariance of test scores under varying test difficulty levels. Obviously, a large storage capacity, general purpose computer allows for the composition of tests in real time, a near infinitive solution to the problem.

In turn, adaptive entry of a student into a test arranged in a difficulty hierarchy remains unexplored. Owen (1969) has developed a procedure for applying Bayesian concepts to either the appropriate determination of a test or for the tailoring of test items to each student, the methodology being appropriate for each problem. The Bayesian models offer a number of distinct advantages:

1. The step size of difficulty between tests can be of the examiner's choice.

2. The choice of entry is dependent upon previously collected data on each student.

3. The choice of a scoring method is less important and is primarily governed by the choice of a loss function selected by the examiner.

4. All of the test item parameters are permitted to vary. Unfortunately there has been no empirical findings to support these views.

The adaptive entry of a student into a test arranged from a shell hierarchy remains to be investigated. In a more integrated instructional and testing paradigm, Suppes (1968) has provided for individualized entry for well over 50,000 students in a mathematics CAI drill and practice program. The results indicate that students can be given appropriate entry based on the single variable of grade level and find an appropriate performance level within a minimum of one hour of instruction.

It should be observed that each of these programs utilized only one variable (grade level) for the predicted entry placement. If multivariate regression techniques were utilized, it would undoubtedly be true that a much more precise placement could be determined. It should be observed, though, that the evaluation of placement for adaptive testing will have to be determined in terms of the criterion of minimum number of test item presentations, since the behavioral evaluation is elusive at best, and perhaps impossible to answer in terms of student self-ratings.

2.2 Tailored Presentation of Test Items

After the student has been placed in a test, the test item presentations should be designed or tailored so as to match items to the current performance or ability level of students. Simply stated, the student should always be presented with those items that best match his competencies as well as providing the greatest discriminations. As he begins to fail, the test should be terminated as quickly as possible. As an overall point of view, testing should be minimized to that degree which minimizes the risk of error to an acceptable degree within the instructional system. In considering the number of techniques offered for tailoring the tests, reference will be made to a number of reviews by different groups: Weiss (Weiss and Betz, 1973) at Minnesota, Hansen (1973) at Memphis State University, etc.

While Weiss argues actively for a two-stage testing model, serious considerations of a number of factors led our group to consider the flexilevel model developed by Lord (1971). The flexilevel model starts a student with a middle difficulty item and proceeds by presenting the next easier item after each wrong response and the next harder item after each correct response. Testing is stopped after n items where n is defined as $\frac{(N + 1)}{2}$ and N is the total number of items of the test. Lord found through computer simulation studies that the flexilevel model yields highly satisfactory results if the difficulty step size is in the range .033 to .067. This model is quite advantageous for two reasons: first, the reduction in test items is clearly specifiable and potential paper and pencil applications are also feasible. Moreover, the test item pool can be directly implemented from an existing conventional test, a highly important developmental factor.

The various problems raised by tailored testing discussed above are summarized by Lord as follows: "Until now, even some very primitive questions about how to carry out tailored testing did not have even vague answers." If these problems are confusing even to the psychometricians, how can the educational sector have confidence in tailored testing? A mature summary of problems and advantages indicates the wisdom of further research and development.

In some of the studies reported (e.g., Angoff and Huddleston, 1958; Cleary, Linn, and Rock, 1968), as many as 20% of the students were misclassified by the routing test. In the case of conventional testing, misclassification of students is similarly unavoidable, since no training test of today is perfectly valid and reliable. Given equivalent weakness for each approach, the use of improved test development methodology is the best course of action.

Another serious weakness of tailored testing is that although it is better for the extreme ability groups, it provides less accurate measurement for the average individual than that of a "standard" test. Lord gave tailored testing an apparent "fatal blow" in this comment:

> If, for example, 500 items were available for tailored testing, better measurement will often be obtained by selecting, for example, the $\underline{n} = 60$ most discriminating items (highest $\underline{a}$) and administering these as a conventional test, rather than by using all 500 in a tailored-testing procedure. This may actually prove to be a fatal objection to any general use of tailored testing.

This remark would hold if tailored testing is applicable only to normative ability measurement, such as the GRE, or the SAT. However, in reaction to this restricted viewpoint of tailored testing, Green (1970) argued that "the computer's failure to improve on conventional testing in this situation does not foreclose the possibility of computer advantages in other cases." Very similar opinion was also shared by Crick (1972) who reacted: "Lord's restricted view of testing, while certainly a legitimate one, does not exhaust the possible applications of computer-assisted testing."

In discussing the prospects of tailored testing, it seems that the following points are pertinent:

1. One reason for Lord's negative comment on tailored testing is the strategy of comparison with a standard test (i.e., a conventional peaked test). However, in comparing the tailored testing with a

"published" (Lord's definition of a conventional unpeaked test) test, his findings indicated that "the tailored procedure gives more accurate measurement than the unpeaked conventional test for all students regardless of level." Thus, in most instructional contexts, tailored testing is apparently the most effective approach.

2. It has also been shown that tailored testing permits a drastic reduction of test items without much loss in the reproducibility of the total test scores.

3. One novel application was made by Ferguson (1969) who used tailored testing in a hierarchical criterion-referenced measurement situation. Concerning the potential usefulness of tailored testing for this purpose, Crick commented: "Intuitively, tailored testing makes much more sense for a criterion-referenced measure than for a norm-referenced measure since the goal of tailored testing is to adjust the test to the individual."

4. In individualized approaches to instruction, it seems that Lord's flexilevel testing may have wide applicability. In the pretest, every subject would take the easy set of the items; but, in the post-test, the subjects would take the difficult set instead. Thus, the use of the parallel forms of the test can be avoided. Furthermore, since the subjects would not have been exposed to many of the harder items, the carryover effects of testing can be minimized. Although Lord developed the flexilevel testing, he has not emphasized the use of it in this context.

5. Tailored testing is appropriate also in the affective domain of measurement. Tam (1973, a study to be presented later) found that

a flexilevel model yielded reliability and validity indices equivalent to the total conventional test, and an empirically observed stop criterion reduced the test length significantly beyond the 50% level.

The prospects of tailored testing depend on willingness to explore its various uses, and the above list is by no means exhaustive. It is hoped that more rigorous explorations of tailored testing will lead to Green's prediction of the "inevitable computer conquest of testing."

2.3 Scoring and Reporting Procedures

The scoring procedures (right/wrong, average difficulty indices, an average of correct item difficulty indices, etc.) the diagnostic interpretation, and the report (quantitative and/or verbal) should be sensitive to all information obtained from the test completed by the student. For example, a bright student who is having a bad day should be differentially treated from the marginal student who is all but failing in the course. Each factor in this three-process representation of adaptive testing should reflect both individual student data and the requirements of the training system so as to maximize the student's learning rates and mastery performance as well as the efficiency of the training system.

For this highly important third process of adaptive testing, limited research findings (theoretical, simulated, or empirical) have been reported. The reviews above subsumed the preponderance of work to date. Therefore, this section will focus on promising topics of further study.

Most scoring procedures utilized the dichotomous right-wrong summed score. Three promising alternatives appear to be feasible.

First, one could differentially weight items so that the most discriminating items relative to the criterion decision zone rather than the total score have the most decisive influence. Studies of item weight indicate weighting can improve decision making as well as test psychometric characteristics. Thus alternative weighted scoring procedures are promising and feasible given a computer's calculation capacity.

In turn, the aggregating or summation process for total score should be studied. Green (1970) posits that a mean of difficulty indices for correct responses offers the most accurate procedure. Similar composite score procedures that stress minimally acceptable mastery levels should be investigated.

Finally, there is important information in the error responses elicited from students. Bock (1972) proposes an item estimation procedure that yields differential information from error alternatives. Intuitively, a "nearly correct" response is more adaptive than a "dum-dum" response. In turn, these error patterns may yield highly important differential categories of students who have partial knowledge. For one group, the remedial alternative of test item review would be sufficient to achieve mastery while the other extreme group may achieve mastery only through totally new training strategy. Large student flow and a computer are required to implement the Bock model.

In terms of diagnostic requirements, total test scores and item pass-fail indices are far too summarized for instructional inference making. Measurement within instruction should yield an individual performance profile that indicates the structure and "valley" of weakness. Profile techniques could yield insights like "the verbal indices are so low that only a high multimedia with audio training approach will insure mastery," or the "uniform pattern of indices indicates that incentives to enhance motivation will insure fast mastery." While speculative in nature, the individual performance profiles interface directly into an adaptive instructional model at this operational juncture.

Interpretation of adaptive tests can be viewed as an "actuarial to clinical" challenge. As sufficient test data bases are collected, refined classification techniques (discriminant analysis) and statistical decision models can be constructed so as to improve the predictive aspects of the interpretation. While a futuristic form of research, the ultimate requirement should be investigated so as to have the full potential of adaptive training (instruction and testing) achieved.

In regard to reports, the recurrent problem of understanding numerical or statistical outputs by instructors, supervisors, etc., are still present. Graphical and verbal reports should be considered and studied. The sufficiency of information for instructional decision making and monitoring is critical. As cited in the Hansen, Hedl, and O'Neil (1971) review, automation of the report process is both feasible and desirable in terms of cost and resource utilization. A consumer survey methodology could be profitably employed at this stage. Obviously, adaptive tests will only be useful to the degree that their results are utilized in a sound, rational manner.

From a modeling viewpoint, the need for empirical research far outdistances our ability to generate ideas or psychometric models. We turn now to a computer based paradigm for implementing this approach.

2.5 Flexilevel Mastery Test

Using the concept of a three-phased adaptive testing process, an approach to adaptive mastery testing has been developed in both Tutor language for the Plato system at the University of Illinois as well as the Sigma 9 system at Memphis State University. From a student point of view, the procedure runs as follows: the student (a) signs on the computer terminal, (b) enters control processing, (c) the system selects the test and entry level for him, and (d) executes the adjusted flexilevel item presentation which will assess his performance. After he has completed the adaptive portion of the test, all remaining items are presented. If he has demonstrated an acceptable level of performance, the system then decides whether to (a) assign the next flexilevel test, reenter the student in control processing and once again begin the flexilevel sequence, or (b) sign him off, an option available to the instructor for acceleration. Figure 2 presents a flowchart of a student moving through each of these answers. A more detailed description follows.

In signing on, the student enters his name and the computer executes a security check designed to limit system accessibility and assure test security. Once he has completed the required sign-on activities, the computer system checks his performance record and aptitude profile to determine which of the tests he is ready for. The system also determines his entry level in the chosen test. Thus, the student is provided the most timely entry test point in terms of his recorded performance, aptitudes, and current in-course status.

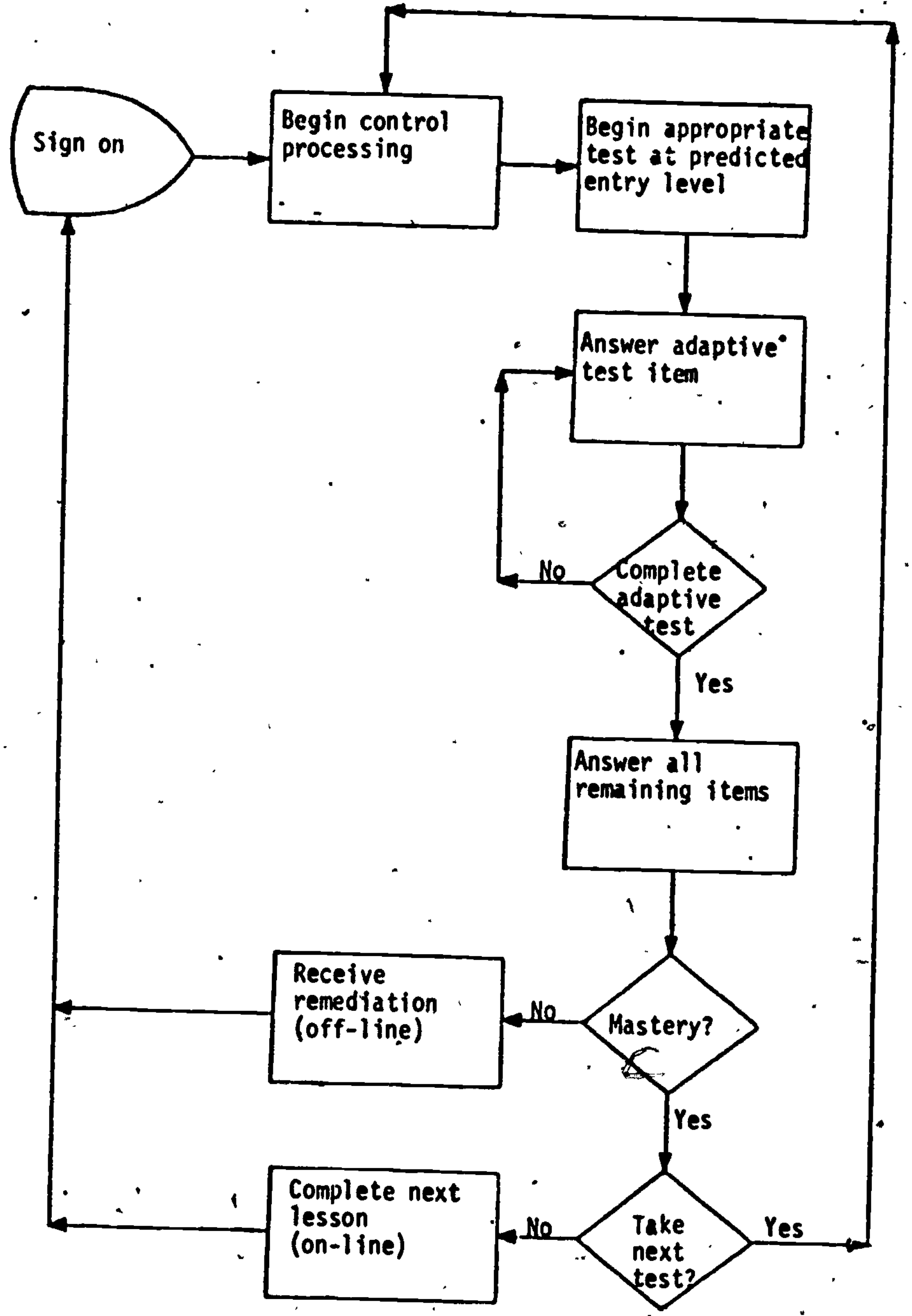

Sign on
Begin control processing
Begin appropriate test at predicted entry level
Answer adaptive test item
No
Complete adaptive test
Yes
Answer all remaining items
Receive remediation (off-line)
No
Mastery?
Yes
Complete next lesson (on-line)
No
Take next test?
Yes

Figure 2. Flowchart of student progress through flexilevel testing program.

Student readiness indices would include previous instructional activities, courses completed, formal education, and other objective learning indices. His aptitude profile might include his test scores on the standardized college entrance tests. His current instructional status identifies how far along he has come in the course. Together these data enable control processing to almost instantaneously compute a predictor equation based on these variables.

Once the predictor equation is determined, the computer system translates it into an appropriate flexilevel test whose difficulty and scope are adjusted to the student's predicted performance. He is therefore provided an evaluation experience individually tailored to his current status. He executes this test on the computer terminal, a useful medium not only because of its rapid response but also because of its transitory display, which augments test security.

The student enters the test at the difficulty level that has been predicted appropriate. If he misses an item, he continues down the difficulty scale until he gets one correct. This establishes his in-test performance base, from which subsequent flexilevel items originate.

When he has completed the adjusted flexilevel test, the remaining test items are presented and the student responses are evaluated (see Figure 1). Green's scoring procedure will be used to evaluate the flexilevel portion of the test, while the entire test will be evaluated using performance criterion scoring procedures. Thus, for each student a tailored test score and a conventional test score will be available. If full mastery, based on the entire test score, is achieved, the student is provided the opportunity to take the next lesson. If he elects to, he then reenters control processing and begins the same sequence in the next assigned flexilevel test.

In the case of test failure, the student goes offline for course remedial activities keyed to his learning deficiencies. Following remediation, all students reenter control processing and restart the flexilevel testing cycle.

After the student attains performance mastery, as a result of either the initial or postremedial test score, the system then decides if he should continue to the next test. If time permits, he most likely will be routed to control processing for a performance prediction update and subsequent testing. If further testing is not prescribed, he is signed off.

Other paradigms have been implemented. Weiss (1974) and his colleagues have a two stage fortran based program. Ferguson (1969) and MSU have elaborate hierarchical skills test paradigms. The Bock procedure for critical zone analysis has been implemented at MSU. We turn now to some empirical results that substantiate these models and computer paradigms.

3.0 Adaptive Processes and Validation

As was presented in the introduction, ATM's should allow not only for systems adaption but also for significant behavioral adaptation for the student. As indicated before, the amount of empirical work to assess this adaptation especially from a reliability and validity point has been exceedingly limited. There is now sufficient starts made in this endeavor to indicate what some of the likely trends appear to be, namely, ATM's provide equivalent or slightly improved reliability and validity measures. This tentative finding appears to hold for asymmetrical score distribution as found in criterion or mastery testing.

3.1 Hedl Study of Intelligence Testing

The first study which examined a computer based adaptive test was performed by Hedl (1971). The Slasson Intelligence Test (CB-SIT) was designed to operate with an IBM 1500 computer instructional system. The test items were presented individually as commonly found in all individualized intelligence testing via a CRT terminal. Students enter their answers for immediate computer evaluation which was based on various key word answer algorithms. For the reliability and validity study, 48 undergraduate students were individually tested with the WAIC, SIT, and the CB-SIT. As can be seen in Table 1, the modified split half reliability correlations for the computer based intelligence tests were lower but essentially equivalent to that of the human administered tests. Table 2 presents indices concerning the concurrent validity which yield moderately strong concurrent relationships. Perhaps most impressively, Table 3 presents the multiple regression analysis on grade point average; and surprisingly, the computer based test proved to be the superior predicator. As indicated in Table 4 computer based testing led to significantly heightened anxiety as well as a decrease in the positive attitude toward the testing. Thus, one can interpret this finding as indicating that computer based adaptive testing may

Table 1

Modified Split-Half Reliability Coefficients (Hedl Study)

	CB-SIT	CB-SIT*	SIT	SIT
Total Group (N = 48)	.66	.79	.79	.88

* Adjusted for test length

Table 2

Coefficients of Correlation (Concurrent Validity) Hedl Study

		WAIS		
CB-SIT	SIT	VIQ	PIQ	FS-IQ
	.75	.55	.32	.54

Table 3

Multiple Regression Analyses (Hedl Study) with GPA

	R	R^2
GPA = -.57 -.66 (Sex) + .03 (CB-SIT)	.66	.44
GPA = .40 -.65 (Sex) + .02 (WAIS)	.56	.32

Table 4

Means For STAI A-State Scores and Attitude Scores

	CB-SIT		SIT		WAIS	
	Pre	Post	Pre	Post	Pre	Post
Anxiety Means	10.7	12.1	9.5	9.2	9.1	9.8
Attitude Means	73.2	66.9	70.0	71.5	70.9	75.5

increase the stress and, consequently, anxiety reactions toward the assessment situation. There is a reasonably consistent pattern for improved validity and acceptable levels of reliability.

3.2 MSU Study of Adaptive Mastery Testing

Our group at Memphis State University has implemented the computer paradigm of adaptive testing described in Section 2.5. As a follow up to this, a quasi-individualized modularly adapted course in beginning graduate level statistics and research methodology was utilized as the context for assessing the reliability and validity of computer based adaptive testing as opposed to conventional paper and pencil testing. Utilizing two different groups, the students were presented with a paper and pencil version of the test and a computer version presented over either a CRT or teletype terminal. Varying predictor variables were used for the entry techniques; the students' grade point average and their running average scores on modules were the main determiners. As presented in Table 5, the mean performance on either version of the test tended to be asymmetrical in nature with this being more pronounced for the module test which can be thought of as complex multi-lesson test. In turn, the reliability coefficients were within the acceptability range. In passing, it should be noted that a modified odd-even technique was utilized; this is similar to that employed in the Hedl study. Unfortunately, this estimation technique tends to underestimate the reliability but is the only available one for assessing adaptive testing sequences, given that they vary as to length and precise item equivalence. Simply, there is a need for new reliability estimation procedures for the adaptive testing situation.

Table 5

Means, Reliability and Validity
(Convention with Adaptive) Coefficients For MSU Study

	Mean Percent	Rel rtt	Validity
Final Exam (N = 28) Paper and Pencil	36	.84	
Adaptive Test	82	.87	.91
Module Tests (N = 33) Paper and Pencil	92	$\overline{.77}$	$\overline{.87}$
Adaptive Test	90	$\overline{.71}$	

Table 6

Reliability and Validity For Affective Adaptive Tests
(Tam Study) of Three Levels of Perceived Teaching

	Reliability				Validity C			
	TH	TA	TL	Tpool	TH	TA	TL	Tpool
Flexilevel	.89	.97	.95	.97	.91	.99	.96	.98
Branched	.57	.88	.85	.92	.60	.88	.85	.92
Two-stage Flexiblock	.90	.93	.79	.94	.91	.87	.65	.91

The total scores were then correlated to yield the validity coefficient. As can be seen in Table 5, these are not only significant but quite substantial. Thus, one finds a fairly reasonable outcome for adaptive testing, that is, it tends to yield reliability and validity coefficients equivalent to that found for conventional testing. This study is continuing and ultimately shall reveal validity measure relating to projects and instructor ratings. Additionally, the Air Force Human Resource Laboratory is contracting to further replicate and extend these paradigms and findings.

3.3 Adaptive Testing On Affective Behaviors

Tam (1973), while at Florida State University, performed an assessment of the reliability and validity for adaptive testing of an affective domain, namely, a Thurstone scale of students' attitudes toward teaching effectiveness. Utilizing a within subject design, Tam presented items which varied from very negative to very positive. A student was allowed to move among the flexilevel adjustments according to the prescribed Lord algorithm. He was terminated once he had agreed three consecutive times, be this at a positive or negative point in the scale. All entries were made at the midpoint in the scale as suggested by Lord. For the purposes of comparison, Tam compared three independent groups, one under flexilevel algorithm, a second under a branching algorithm and a third under a two stage flexiblock algorithm. As can be seen in Table 6, the flexilevel adaptive test yielded substantially the best reliability and validity coefficients. The three groups considered were those teachers who were rated high, average and low, these being pooled over a number of teachers. As can be seen by the magnitude of the coefficients, one can judge affective adaptive testing to be a highly reliable and valid activity.

Perhaps more interestingly, Tam assessed in a posteriori fashion the actual test length required, if a student had been appropriately placed at the positive or negative end of the continuum based on the prior known means for the teacher. Under these conditions it was found that three or less items had to be presented in order to start the Thurstone match required by the design. Such efficient identification of the affective state of the student is quite impressive.

3.4 Summary of Behavioral Studies

While limited in number and scope, these studies indicate a trend, namely, adaptive testing yields equivalent or slightly superior reliabilities and validities given a significant reduction in test items. There is some indication that more stress, anxiety, and perhaps, reality is found in adaptive testing. While no one likes stress, it may be a precursor to improved validity. Given the range of test content, the findings appear to have robust generalizability.

4.0 System Factors and Adaptive Testing

As presented in Figure 1, the adaptive testing process not only provides feedback to the pass-fail decision process for the student but ought to assist in the cybernetic growth of the system; a wonderful concept but yet to be realized. This section shall review the time saving in adaptive testing and move on to the proposed paradigms for systems feedback.

As is to be expected, there is a significant saving under flexilevel item presentation. The MSU groups indicated that only 31% of the items are utilized given individualized test entry. This yields a 153% saving in testing time. The Tam study indicated that 6.3 items as opposed to 16 items yield reliable and valid results. The Hell study did not have an adaptive entry or termination but post hoc analysis indicates a 10% time savings. Given that most individualized AIM systems commit up to 20% of student time to testing, these 50% or greater values are highly significant in a systems efficiency sense. Further replications over systems and test types are obviously required.

4.1 Systems Cybernetic Effects

Surprisingly, there are few suggestive conjectures relating to systems feedback. While mentioned frequently since Stolurow's use of the concept, the operationalization of feedback tends to be a null set for AIM. Let us consider some concrete possibilities.

First, the flow logistics and management of the system is paramount. Overload is the most frequent cause of AIM failure. Test pass-fail rates and time consumptions are highly idealized indices of the system. These can be used to monitor the system and seek quasi-optimal states. Most importantly, modeling of these may be a first step to optimizing the system in a rigorous quantitative manner.

Concurrently, the opportunity to reduce testing time should provide time for the class-state measurement. Class is a concept that relates via cluster analysis common student characteristics to optimal instructional treatments. To know how many group treatment relationships are necessary represents the rank of the system. In turn, individual state to state variations is at the heart of AIM. Therefore, these class-state indices are the core for forming profiles and prescriptive algorithms.

Finally, adaptive testing allows for system adaptation in that shifts in criterion levels by manpower loads or test-remediation subprocesses by pipeline flow are at the heart of system readjustments. The concept of readjustments are hardly new but rarely approached in a dynamic process manner.

References

Angoff, W. H., and Huddleston, E. H. The Multilevel Experiment: A Study of a Two-Stage Test System for the College Board Scholastic Aptitude Test, Statistical Report 58-21. Princeton, N.J.: Educational Testing Service, 1958.

Bock, R. D. "Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories." Psychometrika 37 (March 1972): 29-51.

Cleary, T. A., Linn, R. L. & Rock, D. A. "An Exploratory Study of Programmed Tests". Educational and Psychological Measurement 18 (Summer 1968): 345-360.

Crick, J. E. "A Critical Review of Computer-Assisted Testing." Unpublished qualifying paper, University of Massachusetts, 1972.

Ferguson, R. L. "The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction". Unpublished Ph.D. dissertation, University of Pittsburgh, 1969.

Green, B. F. "Comments on Tailored Testing." In W. Holzman, ed., Computer-Assisted Instruction, Testing, and Guidance. New York: Harper & Row, 1970.

Hansen, D. N., Johnson, Barbara F., Fagan, Robert L., Tam, Peter, Dick, Walter. "Computer-Based Adaptive Testing Models for Air Force Technical Training". Prepared under Contract No. F41609-73-C-0013, Air Force Human Resources Laboratory, Lowry Air Force Base.

Hansen, D. N., Hedl, J. J., & O'Neil, H. F. Review of Automated Testing. Technical Memo No. 20. Tallahassee: Florida State University CAI Center, 1971.

Hedl, J. J., Jr. An Evaluation of a Computer-Based Intelligence Test. Technical Report 21. Tallahassee: Florida State University CAI Center, 1971.

Lord, F. M. "The Self-Scoring Flexilevel Test." Journal of Educational Measurement 9 (Fall, 1971): 147-151.

Masang, B. "Item Weighting: An Approach to Invariance of Test Scores under Varying Test Difficulty Levels." Unpublished preliminary paper, Florida State University, 1972.

Owen, R. J. A Bayesian Approach to Tailored Testing. Research Bulletin 69-72. Princeton, N. J.: Educational Testing Service, 1969.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmark Paedogogische Institut, 1969.

Suppes, P., Jerman, M., & Brian, D. Computer-Assisted Instruction: Stanford's 1956-66 Arithmetic Program. New York: Academic Press, 1968.

Tam, P. T. "A Multivariate Experimental Study of Three Computerized Adaptive Testing Models for the Measurement of Attitude Toward Teaching Effectiveness." Unpublished Ph.D. dissertation, Florida State University, 1973.

Weiss, D. J. Strategies of Adaptive Ability Measurement. Research Report No. 74-5. Prepared under contract No. N00014-67-A-0113-0029 NR No. 150-343, Office of Naval Research. Minneapolis: University of Minnesota, 1974.

Weiss, D. J., & Betz, N. E. Ability Measurement: Conventional or Adaptive? Research Report No. 73-1. Prepared under Contract No. N00014-67-A-0113-0029 NR No. 150-343, Office of Naval Research. Minneapolis: University of Minnesota, 1973.

Wood, R. "Computerized Adaptive Sequential Testing." Unpublished doctoral dissertation, University of Chicago, 1971.