REFLECTIONS ON ADAPTIVE TESTING

DUNCAN N. HANSEN
Memphis State University

The purpose of this paper will be to reflect on various aspects of the adaptive testing field. Building from our prior Memphis State University and Air Force work in the area, the various issues, alternatives, priorities and ultimate styles of research for adaptive testing will be placed in the context of empirical findings and institutional requirements. The rationale for proposing such a pontifical and extremely challenging task is twofold. First, all our substantive and empirical work was recently reported (Hansen, 1975) and it would seem superfluous to rewrite or try to extend this research prior to more effort; therefore, only the major questions and findings will be summarized in this paper. Secondly, the various characteristics of the adaptive testing field will be reflected on in terms of research productivity and institutional requirements. Having by scholarly necessity been forced to read extensively in this domain over the past five years and, in many instances, to take a pencil in hand to follow a variety of formal derivations, I think it appropriate for me to comment about various purposes and styles of research. This is not done to criticize any of these models but rather to seriously address the question, "Are we moving in the most profitable direction and using the most expeditious procedures?"

MSU Adaptive Testing

Generic to any research in adaptive testing or that relating to the whole educational enterprise is a clear understanding of its purpose. For our group, the purpose is that of facilitating achievement or mastery testing. Within industry and military training it is common to find that testing time and managerial demands, especially for individualized techniques, are now taking upwards of 20 percent of the total training time. Such a training commitment becomes sizable and the systems managers must inevitably ask the question, "Is there a more efficient and effective way of going about it?" For example, the Air Force Advanced Instructional System will ultimately have 700 students aboard for any given training shift (2,100 students per day). If one considers that their day consists of six hours of instruction and that approximately 20 percent of this will be given over to testing, one can see that 72 minutes are being allocated on the average for each student's evalution per day. If such testing time can be reduced by 50 percent, an adaptive testing goal set for our efforts, then effectively 1.5+ million dollars worth of salaried money can be gained by shortening the training time for the 2,100 manpower units flowing in this system. It is precisely this type of monetary achievement that impresses our representatives in Congress concerning the importance of research ideas applied to significant educational problems. As will be suggested later, such specific, operational goals, while unachieved to date, give the best rationale for continued research support in this area.

As a corollary to the efficiency issue, an accompanying objective concerns the efficient application of computer technology to the testing process. In essence, one can demonstrate that adaptive testing falls closer to the drill and practice end of the computer usage continuum (Hansen, et. al., 1973) and certainly is orders of magnitude less demanding on a computer than CAI or simulated training. Our experiences and computer algorithms can be offered to you for your consideration. These document an efficient use of computers, tools which are fast becoming integral to the educational processes within our human institutions.

Finally, adaptive testing should be considered within the context of a total systems effort. For our group, adaptive testing is just one component within an overall adaptive instructional system. As one significantly alters the environment and the sequence of educational elements so as to foster or optimize learning outcomes for a given individual, one can see that testing becomes just one more component in such a stream of events. One should look at it, though, in terms of its contributions to the individual and the institution, be this increasing levels of competency or the educational system itself. Thus, one can contend that theoretical models have little or no value unless placed within such a system context since it is the context which will mold and determine the criteria, values, and operation by which its characteristics shall be judged. Let us turn then, to the specifics of the MSU adaptive testing model.

MSU Adaptive Testing Model

Our adaptive testing approach involves three components, namely, the entry of a student into the test, tailoring the test items for the student, and adaptive scoring procedure. Each of these will be discussed in turn. In reference to the entry and test composition processes, a student is entered at a level commensurate with our prediction of his ultimate performance. Therefore, using linear regression techniques mostly composed of variables from prior test performances, a student is placed into a monotonically arranged test. Such a procedure seems to work quite successfully and has an additional advantage of reducing the number of test items to be presented for any

given student. (How this is done should be obvious given an understanding of the flexilevel algorithm.)

While we have very limited data concerning the efficacy of this procedure, entry to final score correlations tend to be in the low .80 range. These are similar to correlation coefficients reported by Cleary at the University of Wisconsin for students who were placed in a branch test according to a predicted outcome level (a personal communication at an AERA conference in 1969). Thus, the adaptive entry of a student seems to be a positive step forward and should be taken into account by any model working within this field.

In reference to test composition, it can be specified that each student, based on his entry profile, will have a specially developed set of composed items. These composed items may reflect information concerning the student's prior performance on various objectives which form the achievement test. Therefore, if one has information about a student's achievement of these objectives, there is no rationale for presenting the item. It is precisely this concept of test composition that appears so advantageous, although it has not been empirically pursued. One can anticipate that sometime within the next year one of the military training systems will pursue it in greater depth.

Tailored Testing of Items

As indicated, Lord's flexilevel algorithm is utilized for tailoring the presentation of test items. For achievement testing, this approach violates the assumptions as to normality as axiomatically represented within this model, but it can empirically be countered that our findings justify the utilization of the algorithm from a student and systems point of view. This adaptation is precisely the ability to move between very difficult and very easy items while at the same time adjusting cutoff criteria where considered appropriate (up to this point our group always used end of test item cutoff procedures but others could be considered). Achievement and mastery testing, especially in a technical training environment, always tend to yield asymmetric performance score distributions. Such distributions, if better understood, could be more readily adapted to flexilevel testing and yield optimal algorithms. Obviously, no attempt to prove such an assertion has been made at this point.

Scoring

Our views on scoring represent an attempt to remain consistent with the traditional procedures of adding up all correct responses and giving weights to those items that are most difficult. Therefore, we have used the Green procedure (Green, 1970), that is, an averaging of the correct item difficulties achieved by a student. Using the flexilevel algorithm and this scoring process, the overall reliability and validity of the adaptive testing procedure seems reasonably satisfactory as it yields coefficients that vary between .6 and .8 (i.e., alpha coefficients and parallel test coefficients).

In addition, we are making plans to contrast two additional adaptive routines so as to resolve what we perceive as a critical problem, namely, the critical zone performer. In any given training situation, there is a critical criterion zone, typically being between the 70th and 90th percent level which is stipulated as a requirement for the attaining of course mastery. If a student scores close or within this level (consider it being bounded by the standard error of measurement), then one should collect more information prior to judging this student as having achieved the objectives or in need of further remediation. At least two approaches can be considered to resolve this problem. The first is an obvious approach simply involving the presentation of an additional set of items for this zone; this is similar to a branching test. A more promising one, especially given the role of the computer, is Bock's (1972) procedure for item latent structure which makes use of the information contained in wrong alternative answers. The Bock model appears to us to be a far more preferable procedure in terms of ongoing large-flow training situations and it shall be evaluated during the coming year within the AF/AIS context.

Data relating to reduction in testing time indicates that only approximately 31 percent of the items are utilized if individualized entry and adaptive techniques are employed. This yields a 150 percent savings in testing time. The samples unfortunately, were extremely small and our group looks forward to a much more extensive validation study in the AIS military training situation. Similar savings are reported by Tam (1973) in his study of affective adaptive testing although modest ones were reported by Hedl (1971) in his intelligence testing. All in all, the results are sufficiently promising to extend the validation for these approaches as well as explore alternative designs within realistic training situations. These alternatives form the substance of the remainder of the paper.

Issues in Adaptive Testing

As an active reader and investigator in the adaptive testing area over the last eight years, one general observation comes to mind, namely, a classical psychometric approach emphasizing those cherished characteristics of excellence, improved reliability, validity, and consequential individual description, is limited in its systems and institutional view. In essence, our efforts have been to describe each and every individual in reliable, finegrain terms while recognizing the needs to improve the testing system. Given these broader insights, the purpose of this section will be to raise issues and possible alternatives as reflected by priorities concerning objectives for adaptive testing. There are three areas to be considered as reflected

by these queries: (1) What are the possible purposes for adaptive testing? (2) What types of formal models might best be pursued for adaptive testing? and (3) How can our theoretical and procedural methods best be evaluated?

Purposes for Adaptive Testing

The tradition within psychometric research as well as test development has focused on descriptions and decisions concerning individuals. On the other hand, many institutions believe group differences in the testing process should be stressed since it is group data that form the basis of decision making. For example, in the current controversy concerning the contribution of schools and curriculum effects, Rakow (1974) argued that tests have been constructed to maximize on individual discriminations and to minimize group differences. Therefore it is not surprising that one finds no statistically significant group effects for schools or curriculums; the Coleman study (1968) or the Jencks follow-on study (1972) represent this type of outcome. Rakow argues that if one utilizes inter-class correlational techniques, one can find highly significant relationships of a subset of items which distinguish among groups. For adaptive tests that attempt to support large human organizations such as military training, this implies that classifying an individual concerning group membership and the characteristics of this group is of a high priority. This adaptive testing approach would utilize a branching item technique so as to lead to reliable alternative group classifications for an individual. Having achieved this, then the more conventional individual discrimination techniques could be applied. Obviously, the utilization of a flexilevel algorithm based on appropriate individual placement would be preferable. The point of such a two-stage model is to provide for more effective adaptation for group placement and ultimately for maximizing on institutional criteria rather than individual criteria alone. Simply, might it be better to find the correct group for an individual rather than know his "true score" on some ability dimension?

In turn, one can look at training systems and recognize that there is a trade-off between training load vs. standard error effects. In essence, as the training load absorbs more and more of the readily available resources, an improvement in the testing process with an associated reduction in standard error is superfluous since all the remaining individuals will have the same minimal treatment. In essence, each student is likely to spend long waiting times and not be able to pursue any kind of optimum course of instruction. Under such circumstances, it is therefore critically important to identify those individuals who can pursue self-study where appropriate. Moreover, it might also be highly important to have adaptive tests that better detect those individuals who seem to have aptitudes for transfer, so that when branched forward or back for review within a normal sequence of instruction, they will receive facilitating effects rather than negative ones.

In turn, as the training load on resources diminishes, one should expect the test length to increase so as to reduce errors of measurement. Thus, one can see that a systems approach to adaptive testing tends to reflect a far more dynamic procedure which might change the criteria, the test length, and the algorithms depending on the state of the training system.

Finally, to be optimally adaptive, one should recognize that our clientele and their institution basically do not understand the concepts, methods, or models of adaptive testing. To them, the quantification, especially as represented by our psychometric models, tends to defy understanding. Allow me to illustrate. MSU has been teaching a measurement course on base at NAS, Memphis. Two of the students were commanding officers of Navy technical training schools and have direct responsibility for supervising the measurement processes within these schools. After completing an eight-week course, each volunteered that they had, prior to the course, never understood any of the quantitative test item statistics or reports other than those concerning students passing or failing, the all important attrition rate. To be adaptive the system should provide the commanding officers, instructors, students, and other concerned people with verbal reports rather than quantitative reports; thus, a client-oriented product approach would vastly enhance the acceptance of adaptive testing. The work of Fowler (1969) with the MMPI successfully demonstrates that psychiatrists readily desire and understand verbal interpretations rather than quantitative reports of the 13 MMPI subscales. These observations about institutional effects hopefully will stimulate your interest in thinking about your clientele as well as your model when you formulate some of your priorities for future research. As cited in the introduction, adaptive testing research must be scholarly, diligent, and of the highest quality while reflecting a form of institutional adaptation which can be appreciated and supported by the clientele who provide the resource support for all research.

Psychometric Models for Adaptive Testing

Within the tradition of adaptive testing research, one reads numerous reports that focus on the comparative merits of alternative psychometric models for adaptive testing. It shall be the thesis of this section that pursuit of an optimal adaptive testing model is likely to be ineffective and the adaptive testing domain needs a strategy for identifying selection criteria that chooses among the many existing models. Optimization studies, especially from a formal point of view, have been pursued for the last 30 years in different contexts with surprisingly similar indifferent results. For example, during the 1940's many statisticians pursued within analysis of variance models the issue of optimal *a posteriori* mean difference tests. After better than a decade and a half of effort, John Tukey (1962) observed that one could not really argue for the one best *a posteriori* test because each varies according to the

decision criteria of the investigator. In essence, it is the characteristic of the research which determines which one of the many tests is the most appropriate.

In turn, the area of mathematical learning models offers a similar finding. Within the context of research on the all-or-none vs. incremental learning processes during the early 1960's, one notes a flurry of research, all of which ended with the conclusion (Atkinson, et. al., 1965) that each mathematical learning model has a set of task characteristics which allows it to be optimal provided that the *a priori* task characteristics are sufficiently matched.

Recently a great deal of effort has gone into the investigation of adaptive instructional models from an optimization point of view. Generalized approaches include various regression models. While these regression models are clearly non-optimal, they have proven significantly successful in facilitating the process. On the other hand, fairly specific models, be these Markoff processes or dynamic programming structures, provide an elegant theoretical explanation (Hansen, et. al., 1973) but rarely fit the data or facilitate learning. Thus, one is led to the view that an array of models for the instructional area will be necessary in order to fit the rather diverse nature of the learning process.

Based on these examples, the proliferation of psychometric models for adaptive testing is likely to have limited productivity. Our efforts to focus on the criteria to be used for the selection of a given adaptive testing model and a better description of how to test the model's fit with the given behavioral phenomena would seem to be a more desirable direction in which to move.

Validation Procedures

As has been observed by each of the reviewers in this area, the amount of empirical work is modest at best. If one considers critical topics, namely, sample size and design techniques, one is even further impressed by our modest beginnings. For example, in reference to sample-size there are those such as Bock (personal communication) who would advocate that at least for his latent item structure model, a sample size of 2,000 students would be required. While pursuing some of the test data for the Air Force with a sample of 1,000 plus airmen, the groups were divided into samples of 200 each and then the usual reliability and validity analysis was performed. In addition, each sample was progressively aggregated into the next. It is fairly clear that the parameter convergence process was still taking place after the sample size had increased to 800. Therefore, it can be argued that it is important to consider maximizing on sample size and to develop techniques by which both item and test parameters converge on their appropriate group and individual values.

In turn, our review of the designs for validation is consistent with that proposed by Tam (1973), namely, that one has to consider a within-test as well as a between-test validation procedure. This can be achieved simultaneously if one notes that one can present adaptive testing as a variation within total test procedure. In turn, this can be contrasted with a parallel form presentation. The two statistics, correlation between the two adaptive and total test scores and the correlations between the two parallel forms, yield a comprehensive representation of the validity. While this may seem excessive to some, such validation procedures provide more substantial empirical results which clearly indicate the justification for reducing total test items.

Summary

This review and reflection has run on in a rather extensive manner. Furthermore, it seems inappropriate to have reflections on reflections. Therefore, this summary will state a final point of view, namely, adaptive testing is sufficiently dynamic that multiple concepts and hypotheses can be incorporated in a design sequentially so as to determine their effect on the efficiency and effectiveness of the assessment process. This extensive review of a number of neglected topics should not be taken as a set of imperatives for research. Rather, these topics and suggestions can best be considered as potential variations within experimental designs of the future. They are offered to you under the assumption of collegial productivity and a firm commitment to the human and societal benefits from adaptive testing. Of all the evaluational techniques available to us at this time, adaptive testing offers that chance to humanize our assessment processes. Such an eventuality, especially in terms of shortening high-stress situations commonly found in testing, cannot be minimized in terms of its benefits.

BIBLIOGRAPHY

Atkinson, R. C., Bower, G. H., Crothers, E. J., *An Introduction to Mathematical Learning Theory,* John Wiley and Sons, Inc. 1965.

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.

Coleman, J. S., et. al., *Equality of Educational Opportunity,* (Washington, D.C.: United States Government Printing Office, 1966).

Fowler, R. D. "The Current Status of a Computer Interpretation of Psychological Tests." *American Journal of Psychiatry* 125 (Supp. 1969): 21-27.

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

Hansen, D. N., Brown, B., Merrill, P., Tennyson, R., Thomas, D., Kribs, H. D. *The analysis and development of an adaptive instructional model(s) for individualized technical training – Phase 1,* Technical Report AFHRL-TR-72-50(1), Tallahassee, Fla.: CAI Center, Florida State University, 1973.

Hansen, D. N., Merrill P. F., Tennyson, R. D., Thomas, D. B., Kribs, H. D., Taylor, S. T. and James, T. G. *The Analysis and Development of an Adaptive Instructional Model(s) for Individualized Technical Training.* Technical Report for Contract No. F33615-71-C-1277, Air Force Systems Command. Tallahassee: Florida State University, 1973.

Hansen, D. N., Ross, S. *A Proposal to Study Additivity for Adaptive Instructional Treatments: A Theoretical Issue, Report for NPRDC, San Diego, Calif.* Memphis State University, 1975.

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

Jencks, C. S., "The Coleman Report and the Conventional Wisdom" in *On Equality of Educational Opportunity,* ed. by Frederick Mosteller and Daniel P. Moynihan, New York: Random House, Inc., 1972.

Rakow, E. A., "Evaluation of Educational Program Differences Via Achievement Test Item Difficulties", paper for American Educational Research Association, Chicago, Illinois, April, 1974.

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

Tukey, J. W. The Future of Data Analysis. *The Annals of Mathematical Statistics,* 33, 1-67. 1962.