ASSESSMENT OF SELECTED EFFICIENT MEASUREMENT STRATEGIES IN AN INDIVIDUALIZED CURRICULUM

David T. Morse/Florida State University

Individualized instructional programs, such as Individually Prescribed Instruction (IPI) or the Program for Learning in Accordance with Needs (PLAN) require a great deal of testing. Hambleton (1975) describes the uses of testing in IPI, for example, as including diagnosis, placement, monitoring of progress, practice, and evaluation of student mastery of objectives. Any measurement strategy that could serve to reduce the amount of testing required without affecting decision-making accuracy would effectively free what formerly was time taken by testing for additional instructional time. There has been much development toward a Bayesian-based strategy for mastery decisions in an individualized curriculum. Work by Novick and associates has shown advances toward a usable system (cf. Novick, Note 1; Swaminathan, Hambleton, and Algina, 1975). During the 1940's, the work of Wald and associates with sequential analysis yielded methodologies of hypothesis testing that appear to be potentially efficient in making mastery decisions (cf. Wald, 1947). While there are numerous examples of other uses of a Bayesian or sequential analysis system, there have been no published reports of implementing such systems in an individualized curriculum.

In proposing the usage of an efficient measurement strategy, several pertinent questions arise. Among these are: (a) Do such measurement strategies in fact effect a reduction in the amount of testing in an individualized curriculum? (b) If so, is the reduction sizable enough to offset the additional work required in grading? (c) Finally, since testing has probable value as an instructional tool, does any change in the amount of testing due to an efficient strategy affect how well students retain what they have learned? These questions delineate the scope of this investigation.

Theoretical Framework

Some topics that require discussion prior to describing the investigation are the efficient measurement strategies and parameter selection.

The Efficient Measurement Strategies

The measurement strategies will be described in as brief a manner as possible. The reader desiring further information on the sequential analysis procedure is referred to Wald (1947). The reader desiring additional information on the Bayesian procedure is referred to Novick and Jackson (1974) and Novick (Note 1).

Sequential analysis. By taking into account the sequence as well as the observed values of repeated random sampling of units from a population, the sequential probability ratio test can effect a significant savings in total sample size when compared with classical Neyman-Pearson hypothesis testing. Also, the sequential analysis procedure has a certainty equal to unity that the sampling will eventually terminate in a decision to either reject or fail to reject the null hypothesis under consideration.

As implemented in this study, the sequential approach proceeds as if testing the mean of a binomially-distributed variable (examinee performance level). Each examinee is assumed to be able to answer correctly some unknown proportion, p, of items in a population or domain of items measuring some skill or knowledge.

There are four parameters of importance to be set in the sequential procedure in addition to the criterion score or proportion, p'. Two of the parameters are an upper (p_1) and lower (p_0) bound for the "indifference region" such that for the region $p_1 - p_0$, the decision-maker is indifferent as to how the examinee is classified. The lower bound, p_0, is selected such that $p_0 \leq p'$, and classifying an examinee as having met criterion is considered an error only if $p \leq p_0$. The upper bound, p_1, is selected such that $p' \leq p_1$ and classifying an examinee as not having met criterion is considered an error of consequence only if $p \geq p_1$.

Next, the average error rates are selected. These are analogous to Type I and Type II errors in hypothetical testing. Error rates α and β are set such that:

α = Prob(Classifying examinee as meeting criterion $| p \leq p_0$); the false positive

probability, and

β = Prob(Classifying examinee as not meeting criterion $| p \geq p_1$); the false negative probability. Note that α and β refer to the relative error rates outside the indifference region ($p_1 - p_0$).

Once these parameters are set, the specific mastery and nonmastery values can be determined, as well as the operating characteristic function of the hypothesis test, and the average sample number for the test (Wald, 1947).

Briefly, the assumptions required for the sequential analysis method are:

1. The units (items) are sampled randomly and independently from the parent population of units (items).
2. The items are scored in a binary fashion.
3. All units (items) are of equivalent goodness, e.g., reliability or validity.
4. Observations of success or failure, X, constitute a random variable which is distributed binomially with parameters (n, p).

Bayesian procedure. As initially described by Hambleton and Novick (1973), a model was proposed that could take into account additional information, termed prior or collateral information, other than an examinee's observed performance on a set of test items. This information could be combined via Bayes's theorem to yield a posterior estimate of the examinee's performance level more accurate than that of the observed test performance alone. Another novel feature was the usage of a step or threshold loss function rather than a squared-error loss function. Misclassification in the model, no matter how much the performance level estimate and true performance level differed, had the same consequence. For a two-action case, the decisions and errors possible might be displayed as in Table 1.

TABLE 1
Outcomes of a Two-Action Decision Model

		Action	
		Advance	Retain
State	Master	0	b
	Nonmaster	a	0

The nonnegative loss a is associated with a false positive error, and nonnegative loss b is associated with a false negative error. Zero loss values reflect correct decisions.

Work on the model has progressed (cf. Novick, Lewis, and Jackson, 1973; Lewis, Wang, and Novick, Note 2) to a process by which the test scores of m - 1 additional examinees on a given test could be combined as collateral information for the calculation of a Bayes estimate of the m*th* examinee's true level. This development, however, would not appear to facilitate implementation in an individualized curriculum, since it requires that all mastery decisions be delayed until an adequate number of students have taken the test in question. Such a procedure would be disruptive to most individualized programs.

The Bayesian procedure, like the sequential analysis approach, requires a preset criterion level of achievement, π_0 • π_0 is analogous to the sequential p′, and reflects the proportion of items from a specified population which the examinee is expected to answer correctly.

One of the simplest models for incorporating prior information that is compatible with a criterion-referenced testing situation is the beta-binomial model. A prior beta distribution specified with parameters p and q (e.g., β[p,q]) can be combined with observed outcomes, such as test item scores, if expressed as a binomial variable with x successes (correct responses) in n trials (items). The combined information, the posterior distribution of achievement level, π, is also of beta form with parameters p + x, q + (n - x) (Novick and Jackson, 1974). Thus, once the prior distribution is specified in terms analogous to those of the likelihood observations (e.g., test results), the calculation of the posterior distribution is simple. If the probability that an examinee's true performance level is greater than or equal to the criterion level equals or exceeds the loss ratio proposition (a/[a + b]), the examinee is declared as meeting criterion. That is, if $[\text{Prob}(\pi_j \geq \pi_0)]/[\text{Prob}(\pi_j < \pi_0)] \geq a/b$, the examinee passes, and fails otherwise.

The assumptions necessary for the Bayesian procedure are the same as those required for the sequential analysis approach.

Parameter Selection

Selection of appropriate parameters is one of the most critical steps in implementing any efficient measurement strategy. The practitioner can glean little information from the literature on how to go about setting these parameters.

For the sequential analysis procedure, the larger the indifference region, or the larger the error rates are set, the smaller will be the number of test items required for either a pass or fail decision. Similarly, the smaller the indifference region, or the more stringent the error rates, the larger will be the number of test items required for either a pass or fail decision.

For the Bayesian procedure, the loss ratio, a/b, need only be determined, rather than a and b individually. The loss ratio expresses the relative degree of "loss" or "cost" associated with a false positive error compared to a false negative error. For instance, a loss ratio of 2/1, or 2, is equivalent to saying that the cost of a false positive error is twice as great as that of a false negative error. In general, the larger the loss ratio, the higher is the effective criterion, or cutoff score, for any test.

The dimensions entering into formulation of loss ratios, the type of loss function, the value of the loss ratio, and the setting of the sequential parameters are all open to research.

Methodology

The Bayesian and sequential strategies were compared to a traditional, fixed-proportion approach in an individualized curriculum at the Florida State University Developmental Research School.

Description of the SCIS Curriculum

The Student Centered Instructional System (SCIS) (Goff and O'Steen, Note 3) is an individualized, self-paced curriculum designed to facilitate the learning of mathematics skills for seventh-grade students. Major subsections of the curriculum are termed components. Each component contains about 10 to 15 individual objectives. Each objective is supported by an instructional segment termed a module. Each module contains embedded review questions for the benefit of the student, as well as explanation relevant to the objective. Students are pretested over objectives in a component, then are exempted from studying particular modules where pretest performance is adequate. The student begins with the first nonexempt module in the component. After studying the material, the student takes a test over that objective. If test performance meets the preset criterion, the student advances to the next assigned module. If test performance is not adequate, the student must review the material, then take an alternate form of the test. A total of three attempts at passing the module test is allowed. The student failing three attempts is advanced to the next assigned module. Upon completion of all assigned modules in a component, the student is given a component test (posttest) that measures acquisition of all the objectives in the component.

Sample

The potential sample consisted of approximately 60 seventh-grade students eligible for the SCIS program. However, participation in the SCIS program necessitated the mutual decision of both the student and the instructor. Hence, the actual sample was smaller and fluctuated during the study (30-34 students at any given time). Each of the students was randomly assigned to one of the three measurement strategies (traditional, Bayesian, or sequential).

Parameters

The criterion or mastery score selection was made after examining the existing passing scores for the module tests in four of the SCIS components (3-6) and after discussion with the instructor. Since most of the tests had a proportion of about .80 as the criterion, and the instructor verified that this was in fact the intent, the criterion or mastery proportion was set at .80.

The loss ratio was also selected after discussion with the instructor concerning implications for test length as well as relevant dimensions. An initial loss of a/b = 2/1 was selected, implying that it was twice as bad or costly to advance a nonmaster as it was to retain a student who had mastered the material. Although the guidelines suggested by Davis, Hickman, and Novick (Note 4) were followed, selection of the loss ratio was a rather arbitrary process.

The selection of the sequential parameters was also based on discussion with the instructor. Some 48 possible sets were considered. Final selection was made to satisfy three arbitrary requirements: (a) the total number of items required to satisfy minimum mastery or nonmastery requirements be relatively small (e.g., < 8); (b) the region of indifference ($p_1 - p_0$) be as small as practical; and (c) potential for missing items and still passing the test exist within 10 items. Given these restrictions and the implications of parameter selection, the mutual decision of the instructor and the investigator was to select the parameters $p_1 = .90$, $p_0 = .70$, $\alpha = .20$, and $\beta = .40$. This was called the "original" parameter set.

Tests and instruments. All module tests to be used were designed to be of standard length, 10 items or responses. All tests were approved by the instructor prior to implementation.

To assess student attitude toward the testing procedure, a 15-item questionnaire was developed. Each item stimulus consisted of a four-point, Likert-type response scale, selected for simplicity and ease of comprehension for the seventh-grade students. An example of the question type is:

I had to take too many tests in Components 3-6. Strongly agree___ Agree___ Disagree___ Strongly disagree___.

Finally, a seven-item, forced-choice instrument was designed to aid in assessing how students ranked the relative importance of false positive and false negative outcomes. The instrument was designed to obtain a measure of the loss ratio that students would feel to be realistic. An example of the type of item in the instrument is:

Which is WORSE (choose one):
a) Passing a test when you don't know the material___
OR
b) Failing a test when you do know the material___.

Implementation. The Bayesian model used was the beta-binomial. The prior distribution was initially expressed as a beta function equal in weight to 10 test items, and was calculated from the proportion of module tests passed on first attempt in the previous component. Observed test performance was incorporated with the prior distribution to yield the posterior distribution. If the equation: $[\text{Prob}(\pi_j \geq \pi_0)]/[\text{Prob}(\pi_j < \pi_0)] \geq a/b$ was met, the student was advanced. Otherwise, the student had to review the material and take an alternate form of the module test. The resulting posterior distribution was then expressed as a prior equal in weight to 10 items for the next module test.

The sequential analysis strategy involved grading the items on any given test in a random order and checking the results against a table of compiled mastery-nonmastery scores for the original parameter set. It was often the case that not all the test items had to be checked before the mastery decision could be made using the sequential method.

The traditional, fixed-proportion strategy simply compared test results to the preset criterion of .80. If the total score met this proportion, the score was considered passing, and was considered failing otherwise.

Students were randomly assigned to the measurement strategy treatments, and the study covered students' progress through four components (3-6). After the majority of the students finished the last component covered in the study, they were administered the attitude and forced-choice instruments. The study covered a five-month period from Fall, 1975 to Spring, 1976.

Results

Comparisons of mean number of initial module tests taken by students during the study were nonsignificant. Hence, the random assignment of students to measurement strategy did not appear to incorporate any systematic bias. Table 2 summarizes the results of the number of alternate tests taken by students.

TABLE 2
Comparison of Mean Alternate Tests Taken by Component

	Component			
Statistic*	3	4	5	6
One-Way ANOVA	F=4.839	F=0.951	F=5.267	F=3.999
df	(2, 31)	(2, 29)	(2, 27)	(2, 28)
p	p=.015	p=.398	p=.012	p=.030

Combined probability using Fisher method: $X^2_{8df} = 26.15$ ($p < .001$)

*All contrasts show Bayesian-group mean > sequential-group mean > traditional-group mean.

There were uniform differences in alternate tests taken. The Bayesian-group students consistently took more alternate tests than the sequential-group students, who in turn took consistently more alternate tests than did the traditional-group students. The mean number of total alternate tests required, on a per-component basis, was 6.20, 3.63, and 2.55 for the Bayesian, sequential, and traditional groups, respectively. Thus, the Bayesian-group students more consistently failed to meet the Bayesian criterion for passing (determined by the loss ratio) than did the other students with their respective criteria.

Mean scores on the component posttests were used as a measure of retention. There were no significant differences in mean posttest scores across the groups, indicating that the additional tests taken by the Bayesian- and sequential-group

students did not yield any significant practice or retention effects.

The stability of the attitude questionnaire and forced-choice instrument scores was assessed by administering the instruments twice to those students who were not in the individualized portion of the SCIS program. The interval between administrations was one week. Stability was expressed as the proportion of items with the same response over occasions for each student averaged over students. There were 22 paired-occasions data sets. Ninety-five percent confidence intervals for these proportions were .63-.67 for the questionnaire, and .80-.84 for the loss ratio (forced-choice) items. No significant differences were noted between the groups on mean total scores on either of the instruments after adjusting item scores for congruency of direction.

Analysis of the loss ratio questions was revealing. Each choice from the set could be interpreted as either a false positive or false negative preference. For example, the item listed earlier has one alternative (a) which indicates that the respondent feels a false positive outcome is worse. Selection of alternative (b) is the same as saying that a false negative outcome is worse. Further, items could be classified as either dealing with actual misclassification, as with the sample item listed earlier, or with the outcomes or consequences of misclassification, such as the following item.

Which is WORSE (choose one):
(a) Restudying material in a module even if you feel you know it.___
(b) Going on to another module before you are ready.___

Alternative (a) represents the false negative outcome, and alternative (b) represents the false positive outcome.

The student loss ratios were formed by calculating the ratio or the number of false positive choices to the number of false negative choices. The ratios varied as a result of the type of question asked, as summarized in Table 3.

Interestingly, if the outcomes of misclassification are considered, the students would apparently accept a ratio of 2/1 as viable, indicating that a false positive outcome is twice as bad as a false negative outcome. For actual misclassification (e.g., passing or failing a test), the preference was clearly for a ratio like 1/2, indicating that a false negative error is twice as undesirable as a false positive error. An overall weighted ratio of 1/1 would appear to represent the students' preference without respect to item type. Thus, what students view as realistic varies with the consequences involved.

TABLE 3
Students' Loss Ratios

Ratio	Individualized Students (n=37)	All Students (n=60)
Actual misclassification (4 items)	0.51	0.47
Oucomes of misclassification (3 items)	1.89	2.59
Overall	0.90	0.97
Overall, weighted for items	0.98	1.08

Reanalysis. The original results were reanalyzed for the nontraditional groups, using parameter sets which had lower effective criteria or mastery scores. For the Bayesian set, loss ratios of 1/1, 3/5, and 1/9 were selected. The first two were selected for their resemblance to the different student loss ratios, and 1/9 was selected as a very liberal (in terms of lowering the effective mastery score) loss ratio. For the sequential set, four parameter sets were selected such that the minimum mastery number changed from one item to four items. These sets were: α=.4, β=.4, p_1=.90, p_0=.60 (parameter set one); α=.4, β=.4, p_1=.90, p_0=.70 (parameter set two); α=.35, β=.2, p_1=.90, p_0=.70 (parameter set three); and α=.3, β=.3, p_1=.90, p_0=.70 (parameter set four). Note that the parameter set number expresses the minimum number of correct responses for a mastery decision. That is, parameter set two requires a minimum of only two correct responses initially in a test to declare the examinee a master.

If efficiency for the nontraditional methods is defined as the difference between the number of initial tests passed given that the traditional criterion was not met and the number of initial sets failed given that the traditional criterion was met (P|TM - F|TM), then this index can be used to compare the nontraditional methods. Using the

TABLE 4
Component Means of Bayesian and Sequential Data Reanalyses Using New Parameters

	Component							
	3		4		5		6	
Bayesian Parameter Set	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency
1/1	3.38	-0.08	5.00	-0.73	7.30	-0.30	4.00	0.00
3/5	2.31	0.61	3.09	0.55	4.50	2.00	3.00	0.80
1/9	0.53	1.62	0.63	2.18	1.50	4.20	0.40	2.40
	Component							
	3		4		5		6	
Sequential Parameter Set	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency	Total Alternates Required	Efficiency
No. 4	1.63	-0.64	4.00	-0.55	5.30	-0.80	3.36	-1.18
No. 3	1.09	-0.18	2.27	1.09	3.30	0.80	1.81	0.09
No. 2	1.09	-0.18	2.63	1.09	3.30	0.80	1.91	0.09
No. 1	0.63	0.09	2.09	1.27	3.40	1.40	1.36	0.73

original nontraditional parameters, both the sequential and Bayesian methods showed negative efficiencies. Table 4 summarizes the reanalysis of the new data using the different parameter sets.

As the data in Table 4 indicate, uniformly positive efficiencies are not obtained for the nontraditional methods until the Bayesian loss ratio changes to 3/5, and the sequential parameter set one is used. Therefore, for the Bayesian method, had the actual outcomes of misclassification student loss ratio of 1/2 been used in the study, the method would have shown a uniformly positive efficiency in comparison with the traditional method. By the same token, the overall student loss ratio of 1/1 would not have shown a positive efficiency in the same comparison. For the sequential method, had parameter set one been used in the study, the method would have shown a uniformly positive efficiency in comparison with the traditional method.

Conclusions

From the data presented in the results section, it is readily apparent that the selected nontraditional measurement strategies used in this study can in fact show a reduction in the average number of items required to make a mastery determination over a traditional, fixed-proportion criterion.

The data from this study, while somewhat sparse, yields some valuable information. First, as mentioned above, careful consideration must be given to the selection of parameters prior to the implementation of any nontraditional strategy. The additional time and effort required for grading test papers requires that nontraditional measurement strategies must effect a reduction of practical significance in order to have utility in the classroom. From the reanalysis of the data, it was shown that such reductions are possible by selection of appropriate parameters.

Second, relatively small differences in total tests taken would not appear to have practical consequences in terms of retention test performance. This result should be tempered by keeping in mind the small sample sizes, however.

Third, and unique to the Bayesian method, is the outcome of students' loss ratios. Students are clearly more reluctant to accept the outcomes of misclassification than the actual misclassification itself. The question arises of what role the student should have in the determination of such parameters as the loss ratio. Guidance available in the literature is clearly deficient for most potential users of nontraditional measurement strategies. Further, whether parameter sets which are required for efficiency will be acceptable to instructors has not yet been investigated.

The potential impact of nontraditional measurement strategies depends, in large part, upon how well the models can be adapted to the existing classroom situation. Work on tailored testing models has been extant for many years. Few day-to-day applications, however, exist. The move towards a more systematic, individualized approach to instruction is a reality. Nontraditional measurement strategies have the potential for reducing the amount of student time taken up by testing. This potential can only be realized upon further investigation and refinement of the model.

Reference Notes

1. Novick, M. R. New statistical techniques to evaluate criterion-referenced tests used in individually prescribed instruction. Final Report, Project No. 2-0067. Iowa City: The American College Testing Program, 1973.
2. Lewis, C.; Wang, M.; and Novick, M. R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City: The American College Testing Program, 1973.
3. Goff, M., and O'Steen, J. *SCIS mathematics.* Tallahassee, Florida: The Developmental Research School of the Florida State University, 1975.
4. Davis, C. E.; Hickman, J.; and Novick, M. R. A primer on decision analysis for individually prescribed instruction. ACT Technical Bulletin No. 17. Iowa City: The American College Testing Program, 1973.

References

Hambleton, R. K., and Novick, M. R. Towards an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement,* 1973, *10,* 159-170.

Novick, M. R., and Jackson, P. H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

Swaminathan, H.; Hambleton, R. K.; and Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement,* 1975, *12,* 87-98.

Wald, A. *Sequential analysis.* (Reprint Edition) New York: Dover, 1973. (Out of print, New York: Wiley, 1947)