DOCUMENT RESUME

ED 042 815 TM 000 103

AUTHOR	Kriewall, Thomas E.; Hirsch, Edward
TITLE	The Development and Interpretation of Criterion - Referenced Tests.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-5-0216
PUB DATE	Feb 69
GRANT	OEC-5-10-154
NOTE	26p.; Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, Feb 1969
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.35
DESCRIPTORS	Arithmetic, Behavioral Objectives, *Computer Assisted Instruction, *Criterion Referenced Tests, *Item Sampling, Measurement Techniques, Probability Theory, *Test Construction, Testing, *Test Interpretation, Test Reliability, Test Validity

ABSTRACT

As an alternative to a classical test theory basis for criterion-referenced test construction, it is proposed that a strict item-sampling model be used. The computer's role in such a model is outlined. The assumptions of the model are carefully defined and its properties reviewed. The relationship between mastery criteria and such sampling plans as single sampling, simple curtailed sampling, and the use of the sequential probability ratio test is discussed. Representative operating characteristic curves for a number of different plans are included. Suggestions are offered for reducing the testing time needed to detect mastery attainment levels which are consistent with the Newman-Pearson theory of hypothesis testing. Applications are indicated, and an example included, in the area of computer-generation and administration of criterion-referenced tests of mastery in selected arithmetic skills. (DG)

THE DEVELOPMENT AND INTERPRETATION OF CRITERION-REFERENCED TESTS

Thomas E. Kriewall
Project Specialist, Project 2013
Computer Managed System of Mathematics Instruction
University of Wisconsin
Research and Development Center for Cognitive Learning

and

Edward Hirsch
Product Research Trainee
Educational Product Research Program
University of California Los Angeles

A paper presented at the annual meeting of the
American Educational Research Association
February 5-8, 1969 Los Angeles, California

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

Acknowledgements

This paper is the result of work begun by the authors in the summer of 1968 at System Development Corporation, Santa Monica, California. Support for the later phases of the study was provided by the Wisconsin Research and Development Center for Cognitive Learning, Madison, Wisconsin.

We wish to recognize especially the following individuals whose interest, encouragement and critical evaluation significantly contributed to the ideas expressed here. Harry Silberman and John Coulson suggested the problem and provided through SDC the environment, facilities, and staff needed to begin the investigation. At the Wisconsin R&D Center, the senior author enjoyed the support of the Center facilities and staff and benefited from especially valuable suggestions offered by M. Vere DeVault, Thomas Romberg, and Thomas Fishbach. At UCLA, T. Husek and W. James Popham provided a number of helpful suggestions for which we wish to express appreciation.

Use of the University of Wisconsin Computing Center was made possible through support, in part, from the National Science Foundation, other United States Government agencies and the Wisconsin Alumni Research Foundation (WARF) through the University of Wisconsin Research Committee.

THE DEVELOPMENT AND INTERPRETATICN OF CRITERION-REFERENCED TESTS

The particular focus of this paper is the relationship between the design of criterion-referenced tests and their use in instructional management systems.

Criterion-referenced tests have been used in a variety of forms over the last three decades as tools for educational measurement (Birnbaum,1958; Hammock,1960; Ebel,1962; Lord and Novick,1968). The different types of criterion-referenced tests vary considerably from one to another in terms of the underlying model of design. Nevertheless, there is a common point of application for such tests, to classify examinees according to higher or lower ability as their observed score exceeds or falls short of a given criterion value. This application is used in instructional management systems to separate learning groups into instructional subgroups in which the instructional treatment can be better fitted to the relatively homogeneous ability exhibited by members of each subgroup.

A second property of criterion-referenced tests of use in instructional management systems is that, when properly designed, such tests can give estimates of an individual's absolute level of proficiency, rather than the relative estimate provided by classical norm-referenced tests.

A third feature of special interest in computer managed systems is the adaptability of the criterion-referenced test to the possibility of computer generation of test items and thereby to the development of relatively economical and efficient decision systems. The model for developing and interpreting criterion-referenced tests that is to be described in this paper suggests a role for the computer similar to that of the industrial quality control sampling inspector.

In rough outline, the idea is to specify classes of problems for which adequately reliable problem solving behaviors are to be developed in the individual by the instructional system. When the "product" (i.e. the individual's specific set of problem solving behaviors) is ready for inspection or test, the computer generates a random sample of problems from a specified population of items, possibly by using item-generation rules. The items are assumed to be of equal importance. They are also assumed to be approximately of equal difficulty, in the sense that the individual has about the same probability of success on any item randomly selected from the population. The absolute level of this probability is assumed to be a function of the degree to which the examinee has developed an appropriate set of problem-solving behaviors. The computer's potential role in this scheme is not only to generate items, check responses, and keep records automatically. It also can effectively administer the sampling plans that define the kind and amount of information required to show that the learning "product" offered by the individual meets "design" specifications set forth in the instructional package.

It may be helpful to consider a concrete example at this point. As part of a recent instructional management experiment conducted at Wisconsin's Research and Development Center, the cooperating teacher was in the process of teaching a review of reduction of fractions to lowest terms. This segment was to be followed by a unit new to the class: adding simple fractions with unlike denominators.

Three criterion-referenced tests were developed independently by different members of the project staff using specified item-generation rules. Test A was designed to measure proficiency in reducing fractions to lowest terms when the outcome is a common fraction. Test B measures a similar proficiency except that the outcome is a mixed number. Test C is a measure of proficiency in adding simple fractions. Each pretest and posttest contained a random sample of five items selected from the item population defined by the item generating rules. These tests together with a brief summary of pre- and posttest results are included in Appendix A as illustrations of one application of the model we are about to describe.

Properties of An Item-Sampling Model.

The basis of criterion-referenced test construction proposed here is a strict item-sampling model. One first defines a specific category of problems, either by means of item-generating rules or, if necessary, by simply listing the entire population. This population we call a specified content objective (SCO) inasmuch as it is the intended objective of instruction to develop individually effective sets of problem solving behavior relevant to the SCO.

At any given time, it is assumed that individual a possesses a single proficiency with respect to a specified content objective. A measure of this proficiency is the individual's relative true score on the population of items, which we label ζ_a. According to this definition, proficiency is a parameter which can be interpreted as the probability of the individual's making an acceptable response to any item randomly selected from the specified population.

For present purposes, a criterion-referenced test may be considered to consist of constructed-response, binary items. Assuming local independence the examinee's performance on such a test may be regarded as a series of n independent Bernoulli trials having probability of success ζ_a on each trial, where n = number of test items.

If the examinee were to be repeatedly tested with different random samples of size n or if a homogeneous proficiency group (defined by $\zeta_i = \zeta_j$ for every pair of examinees i and j) were to take the test, the expected distribution of scores, x_a , would be given by the ordinary binomial

$$f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1 - \zeta_a)^{n-x_a}$$

According to the model, each examinee responds to an item as though he were tossing a coin having bias ζ_a.

The following are well-known properties of tests built according to an item-sampling model (Lord and Novick,1968,251). The observed test score, x_a, is a sufficient statistic for estimating ζ_a. Secondly, the error of measurement is given by $\eta_a = x_a - n\zeta_a$. Since the expected value of the test score is also $n\zeta_a$,it follows that the expected error over repeated testings for a given examinee is zero. In other words, the longer the test (within practical limits), the better the true score estimate. Error variance is given by the usual relation

$$\sigma^2(\eta_a) = n\zeta_a(1 - \zeta_a)$$

for which an estimate, unbiased over item sampling is

$$\hat{\sigma}^2(\eta_a) = x_a(n - x_a)/(n - 1)$$

It is interesting to note that the error of measurement is a function only of test length, n, and the examinee's proficiency, ζ_a. Therefcre, if estimation of proficiency is the essential purpose of the test, then classical item selection techniques involving the consideration of such item parameters as the p-value, item-test correlation, and discrimination coefficient are of no use in the design of criterion-referenced tests constructed according to the model proposed here.

Techniques of Quality Control Using Criterion-Referenced Tests.

An individual's proficiency with respect to a specified content objective may also be regarded as a measure of product quality for a given instructional package. Since proficiency is assumed to be a monotonically increasing function of instructional time, a problem of interest to the instructional manager is the estimation of proficiency at given points in time to determine whether or not it meets some minimal criterion of acceptance. A method of handling this decision problem is as follows.

Collecting a population of items into a specified content objective may be compared with the industrial procedure of dividing output into inspection lots. In the simplest sense, one may consider a semester's work as an ordered sequence of specified objectives. All the questions to be asked over the semester are divided into "inspection lots" or SCO's. The quality of individusl test items contained in the lot is judged, again in the simplest case, according to the binary attribute "acceptable" or "unacceptable" depending upon the individual's response to the item.

Each examinee possesses a particular proficiency at a given time for producing "acceptable" items. The criteri n-referenced test is regarded in this view as a random sample of the examinee's potential production on a given inspection lot or SCO. On the basis of the observed score, one must decide whether or not to accept the lot (i.e. judge the examinee a master) or reject the lot (decide that the individual's problem solving behaviors must be improved). This raises the usual questions involved in hypothesis testing concerning what size test and which criterion value should be selected so that the errors of classification are held within specified probability limits.

Assistance in solving this problem can be found in the vast literature dealing with the construction and selection of sampling plans. Time permits our sketching only an overview of the basic ideas here.

A sampling plan may be defined conveniently for our purposes in one of two ways. A single sampling plan is defined simply by selecting a value for test length, n, and an error criterion, c. An alternative method that is more useful when one is considering certain kinds of curtailed sampling plans is to specify the probability of a type I error of classification, α, and the probability of a type II error of classification, β. This method also requires that two additional quantities be specified: the minimum proficiency that sets a lower bound to the mastery range, ζ_1, and the maximum proficiency that sets an upper bound to the nonmastery range, ζ_2. The range of proficiency between $\zeta_1$and ζ_2 is called the region of indifference. Specification of α, β, ζ_1, and ζ_2 is equivalent to specifying n and c and conversely. The equations releting these six quantities are easily derived from the item-sampling model described earlier.

The quality control characteristics offered by a particular sampling plan are revealed by a function called the operating characteristic (OC). The OC simply enables one to compute the probability of deciding in favor of acceptance or mastery as a function of the individual's true proficiency.

The general shape of the OC curve is very similar to the usual item characteristic curves found in classical test theory. In general the probability of an examiner being classified a master on a given SCO is a decreasing function of his error rate, $\zeta_a' = 1-\zeta_a$. If he never makes an error, there is a probability of one that he will be judged a master; if he always makes errors, the probability of his classification as a master is zero. For each error ratio between zero and one, the OC curve shows the probability (S) of a "successful" or mastery decision being made.

If a large number of examinees of given proficiency are tested on the same SCO, some will be classified as masters and some as nonmasters. The OC curve can be derived from the item-sampling model by computing the probability that an individual with any given proficiency, ζ, will make fewer than c errors. This condition for a mastery classification is given by the cumulative probability function:

$$(1) \qquad S = \sum_{w=o}^{c-1} \binom{n}{w} \zeta^{n-w} (1 - \zeta)^{w}$$

where $w = n - x$, the number of wrong responses made on n items. Equation (1) is the OC for single sampling plans of size n and error criterion c. Figures 1, 2 and 3 show representative OC curves for a number of different plans, including one for the illustrative tests included in Appendix A.

Insert Figures 1, 2 and 3 about here

From an examination of these curves, it may be seen that the OC functions somewhat like an item characteristic curve. The value of the criterion c roughly determines the proficiency level at which there is an equal chance of being classified master or nonmaster. This is the region of steepest descent for the OC curve and may be considered as the proficiency level at which the test is maximally discriminating. The test length n determines the steepness of the slope and hence the sharpness with which the test discriminates between "high" and "low" proficiencies.

It should be noted that setting a higher criterion (or lower error criterion) does not of itself improve the proficiency found in those examinees classified as masters. In education, as in industry, quality cannot be inspected into a product. Rather the instructional package must be improved if higher quality is desired in the learning product.

MINIMIZING TEST LENGTH

If the criterion referenced test can be administered via interactive terminals, the model suggested here is well adapted to the study of sampling plans that minimize the number of questions required to classify students with fixed error probabilities α and β.

This may appear contrary to Neyman-Pearson theory, which shows that α and β depend on test length n. Nevertheless curtailment of tests is possible without causing a change in either α or in β.

For example, if a single sampling plan defined by $n = 5$ and $c = 2$ were employed, one could curtail the test as soon as two errors are observed or as soon as 4 correct responses are noted. The final decision in each case is exactly the same as that which would be made if the test ran to completion. In this case, the curtailed plan and the single sample non-curtailed plan have exactly the same OC curve and hence the same error probabilities α, β. What is lost by curtailment is the accuracy of estimation of an examinee's proficiency or true score.

A sampling plan that minimizes the test length for given values of α, β at ζ_1 and ζ_2 exists. This is Wald's sequential probability ratio test or SPRT (Wald, 1947).

Sampling plans that reduce the number of questions required to reach classification decisions without loss of protection against errors of classification are of interest for two counts. Cost of testing is proportional to the number of items as is the length of time taken from instruction for testing purposes. It is highly desirable therefore to minimize test length while still providing for accurate decision making.

Figures 4, 5 and 6 show how the average sample number (ASN), or expected test length, varies with error rate for curtailed tests having the same OC as the single sample plans shown in previous figures.

Insert Figures 4, 5 and 6 about here

The distance from the horizontal line representing the fixed sample to the ASN is a measure of the average saving in test length for these curtailed plans.

The saving that curtailed testing provides together with the growing interest in computer generation of test items warrant further study in connection with research on the design of economical and practical systems of computer-based instructional management.

REFERENCES

Birnbaum, A., Statistical theory of some quantal response models. _Annals of Mathematical Statistics_, 1958,_29_,1284 (abstract).

Ebel, R., Content standard scores. _Educational and Psychological Measurement_, 1962, _22_,15-25.

Glaser, R., Instructional technology and the measurement of learning outcomes: some questions. _American Psychologist_, 1963, _18_, 519-521.

Hammock, J. Criterion measures: instruction vs. selection research. Paper read at the meeting of the American Psychological Association, September, 1960.

Lord, F. and M. Novick, _Statistical Theories of Mental Test Scores_. Reading, Mass: Addison-Wesley, 1968.

Statistics Research Group, Columbia University, _Sampling Inspection_. New York: McGraw-Hill, 1948.

Wald, A.,_Sequential Analysis_. New York: Wiley, 1947.

Appendix A

Three sample criterion-referenced pretests are shown on pages A-1 and A-2. Posttests containing items randomly sampled from corresponding pools are shown on A-3 and A-4. The "fail-safe" box simply permitted pupils who felt they had no proficiency whatever on a given set of problems to bypass the set without embarrassment. Pupils made use of this option mainly on pretest C.

Test results for a class of 19 fifth-graders are summarized on pages A-5 to A-7. The "high" group on each test consisted of pupils who made fewer than two errors. Test reliabilities are relatively high for the total group but become erratic when computed for the relatively homogeneous proficiency subgroups. This is simply an expected consequence of the fact that variation of scores within a subgroup is mostly error variation.

The observed proficiency gains and transition matrices are illustrative of the kind of management information that this type of criterion-referenced test may provide.

The OC curve for the sampling plan employed on these tests ($n = 5$; $c = 2$) is shown on Figure 1 in the main body of this paper. If the tests are expected to discriminate a maximum error rate of $\zeta_1' = .15$ for the high group and a minimum error rate of $\zeta_2' = .65$ for the low group, the OC curve shows that the errors of classification will be approximately $\alpha = .16$ and $\beta = .04$ for this plan. Shifting to an error criterion of $c = 1$ would greatly increase α but only slightly decrease β.

SAMPLE PRETEST

Name ____________________

Instructions: There are three parts on this test. If you decide you don't know how to do the problems in any part, put an X in the "fail-safe" box and go on to the next part.

Part A

Fail-safe: □

For each fraction, find the equivalent fraction in lowest terms.

EXAMPLE:

Problem	Response
$\frac{16}{40} =$	$\frac{2}{5}$

Problem	Response	Problem	Response	Problem	Response
1. $\frac{10}{36} =$		3. $\frac{27}{60} =$		5. $\frac{35}{42} =$	
2. $\frac{18}{48} =$		4. $\frac{12}{14} =$			

Part B

Fail-safe: □

For each fraction, find the equivalent mixed fraction in lowest terms.

EXAMPLE:

Problem	Response
$\frac{30}{12} =$	$2\frac{1}{2}$

Problem	Response	Problem	Response	Problem	Response
6. $\frac{42}{12} =$		8. $\frac{70}{36} =$		10. $\frac{63}{25} =$	
7. $\frac{63}{54} =$		9. $\frac{42}{9} =$			

Part C

Fail-safe: □

Find the sums.

EXAMPLE:

Problem	Solution	Response
$\frac{1}{6} + \frac{4}{5} =$	$\frac{5}{30} + \frac{24}{30} =$	$\frac{29}{30}$

Problem	Solution	Response	Problem	Solution	Response
11. $\frac{3}{4} + \frac{1}{8} =$			14. $\frac{5}{7} + \frac{1}{2} =$		
12. $\frac{3}{5} + \frac{2}{7} =$			15. $\frac{1}{9} + \frac{4}{7} =$		
13. $\frac{0}{4} + \frac{6}{7} =$					

SAMPLE POSTTEST

Name ______________________

Instructions: There are three parts on this test. If you decide you don't know how to do the problems in any part, put an X in the "fail-safe" box and go on to the next part.

Part A

Fail-safe: □

For each fraction, find the equivalent fraction in lowest terms.

Problem	Response	Problem	Response	Problem	Response
1. $\frac{35}{50}=$		3. $\frac{35}{49}=$		5. $\frac{10}{28}=$	
2. $\frac{15}{35}=$		4. $\frac{21}{30}=$			

Part B

Fail-safe: □

For each fraction, find the equivalent mixed fraction in lowest terms.

Problem	Response	Problem	Response	Problem	Response
6. $\frac{50}{21}=$		8. $\frac{70}{21}=$		10. $\frac{98}{25}=$	
7. $\frac{45}{35}=$		9. $\frac{70}{54}=$			

Part C

Fail-safe: □

Find the sums.

Problem	Solution	Response	Problem	Solution	Response
11. $\frac{6}{7} + \frac{2}{9} =$			14. $\frac{4}{9} + \frac{6}{7} =$		
12. $\frac{0}{9} + \frac{1}{2} =$			15. $\frac{2}{7} + \frac{1}{4} =$		
13. $\frac{1}{5} + \frac{1}{4} =$					

DATA SUMMARY

Test Identification	Mean Proportion Correct						KR-20		
	Group			Group Size			Group		
	Total	High	Low	N	N High	N Low	Total	High	Low
Pretest A	.43	.88	.27	19	5	14	.74	-1.0	.28
B	.42	.85	.31	19	4	15	.70	-.02	.41
C	.00	.00	.00	19	0	19	-	-	-
Posttest A	.63	.98	.38	19	8	11	.82	.04	.48
B	.57	.85	.36	19	8	11	.74	-1.0	.48
C	.39	.92	.20	19	5	14	.88	0.0	.68

n = 5

c = 2

ESTIMATED PROFICIENCY GAIN

	Objective A	Objective B	Objective C
Total Group	.20	.15	.39
High Group	.10	.00	.92
Low Group	.11	.05	.20

TRANSITION MATRICES

Test A

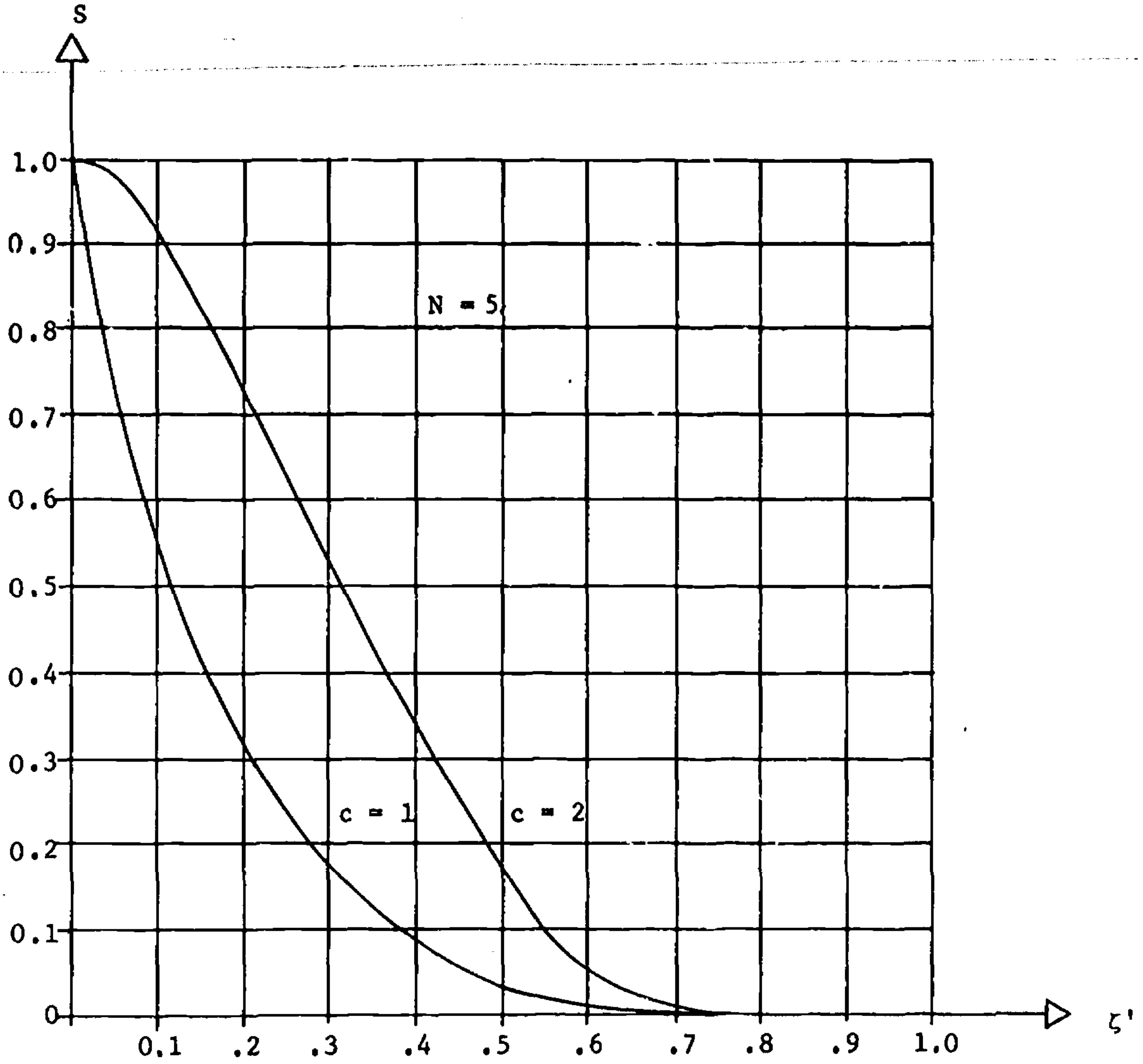
		From Level		
		High	Low	Final Totals
To Level	High	4	4	8
	Low	1	10	11
	Initial Totals	5	14	19

Test B

		From Level		
		High	Low	Final Totals
To Level	High	3	5	8
	Low	1	10	11
	Initial Totals	4	15	19

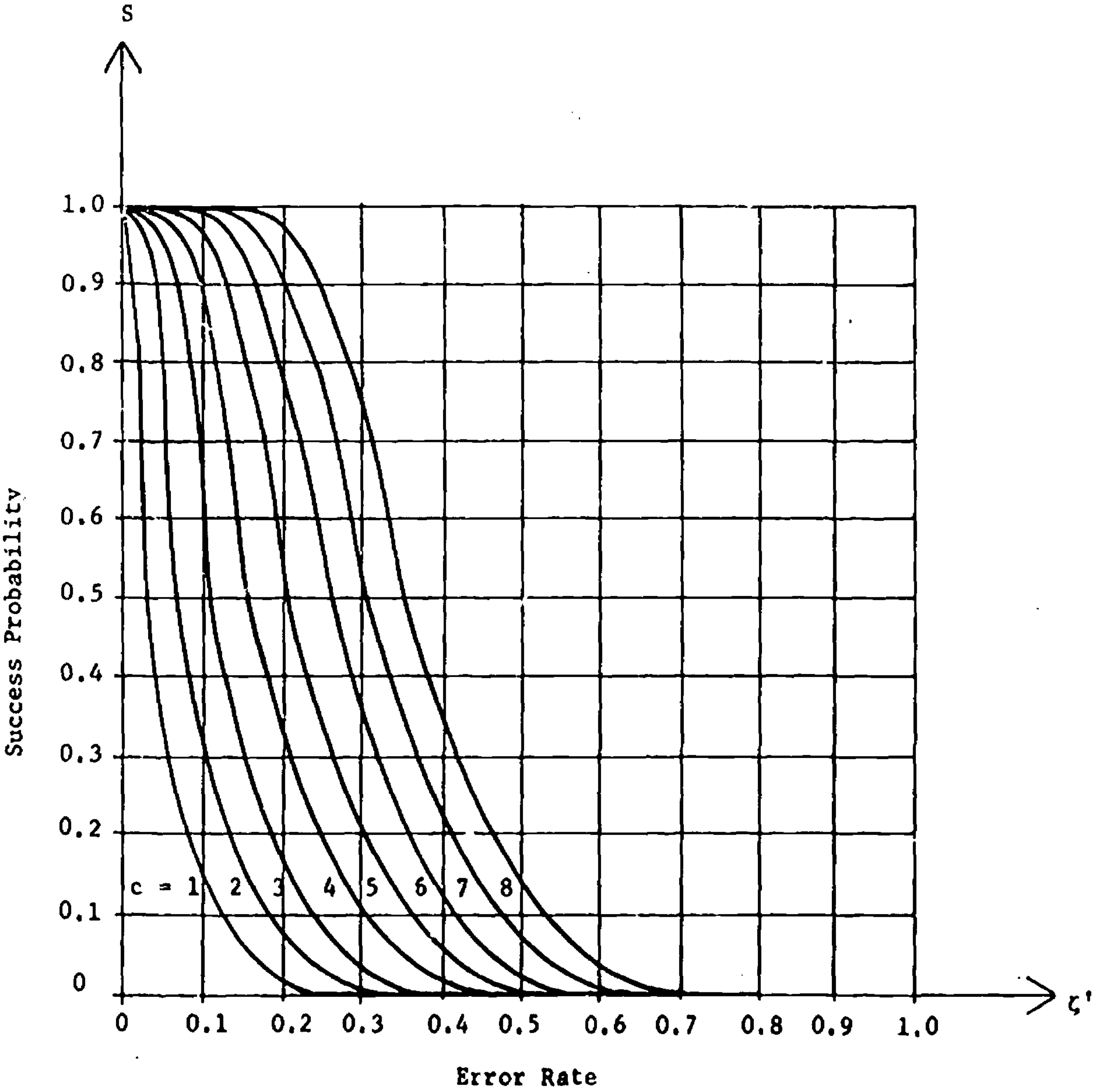
Test C

		From Level		
		High	Low	Final Totals
To Level	High	0	5	5
	Low	0	14	14
	Initial Totals	0	19	19



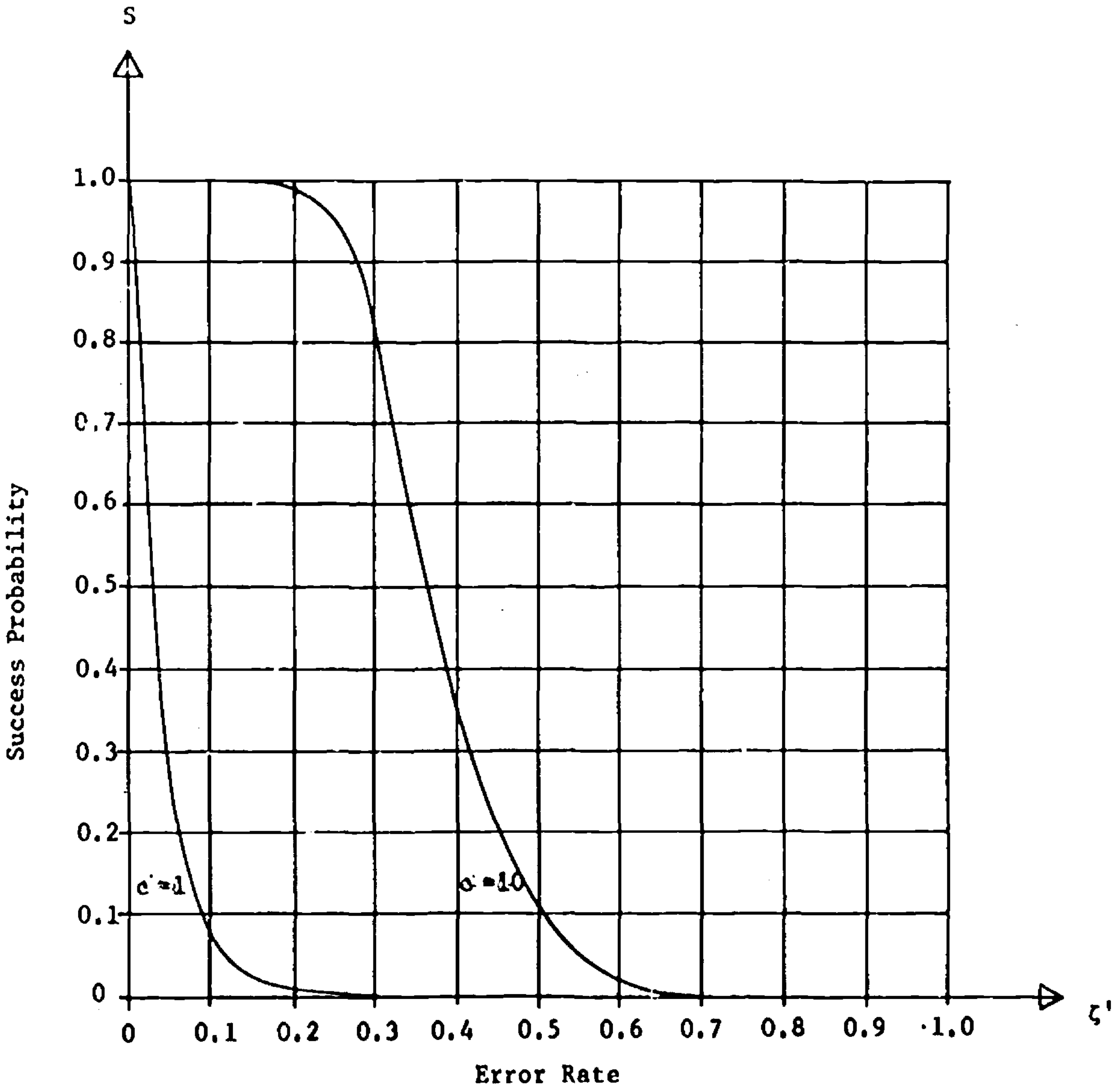
O-C Curves for N = 5; c = 1, 2.

Figure 1



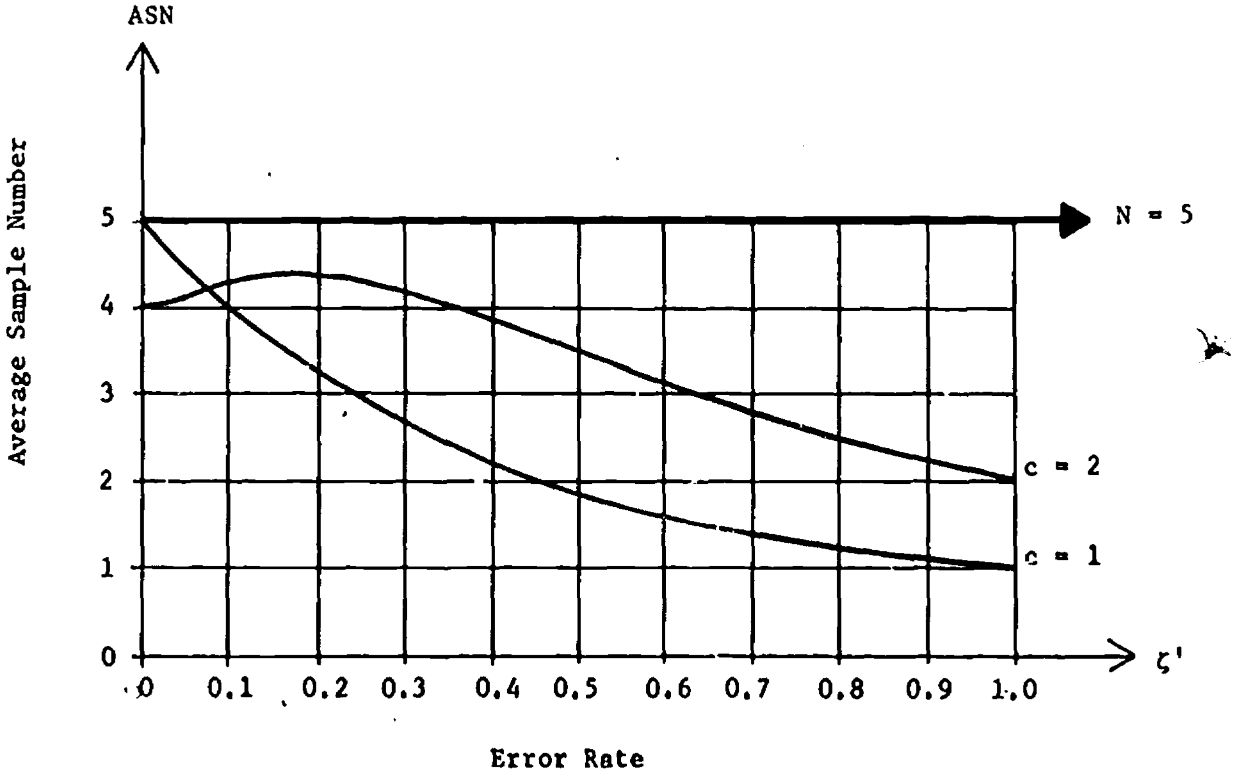
O-C Curves for N = 20
And Error Criteria c = 1, 2, ..., 8

Figure 2



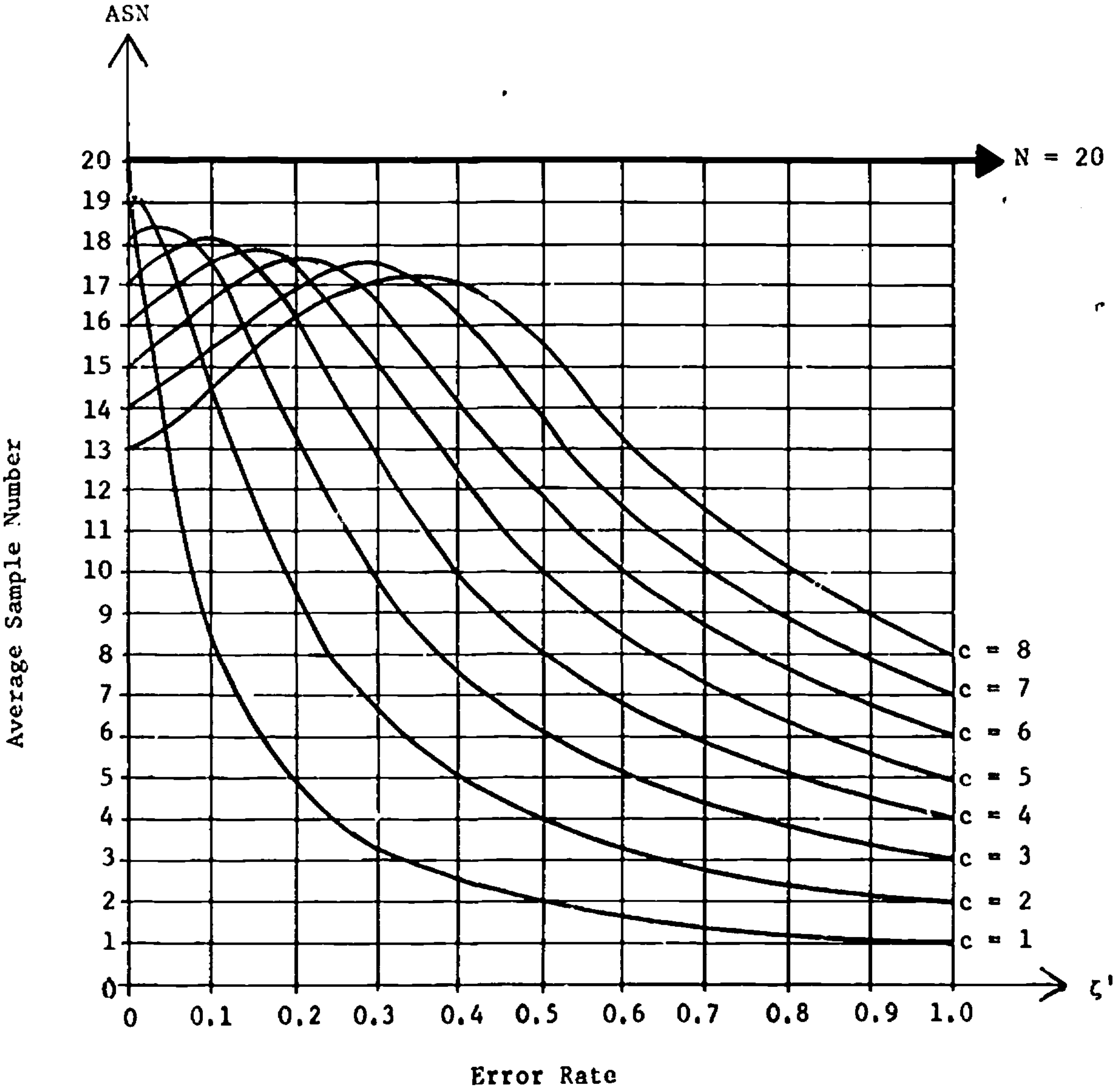
O-C Curves for N = 25
And c = 1, 10

Figure 3



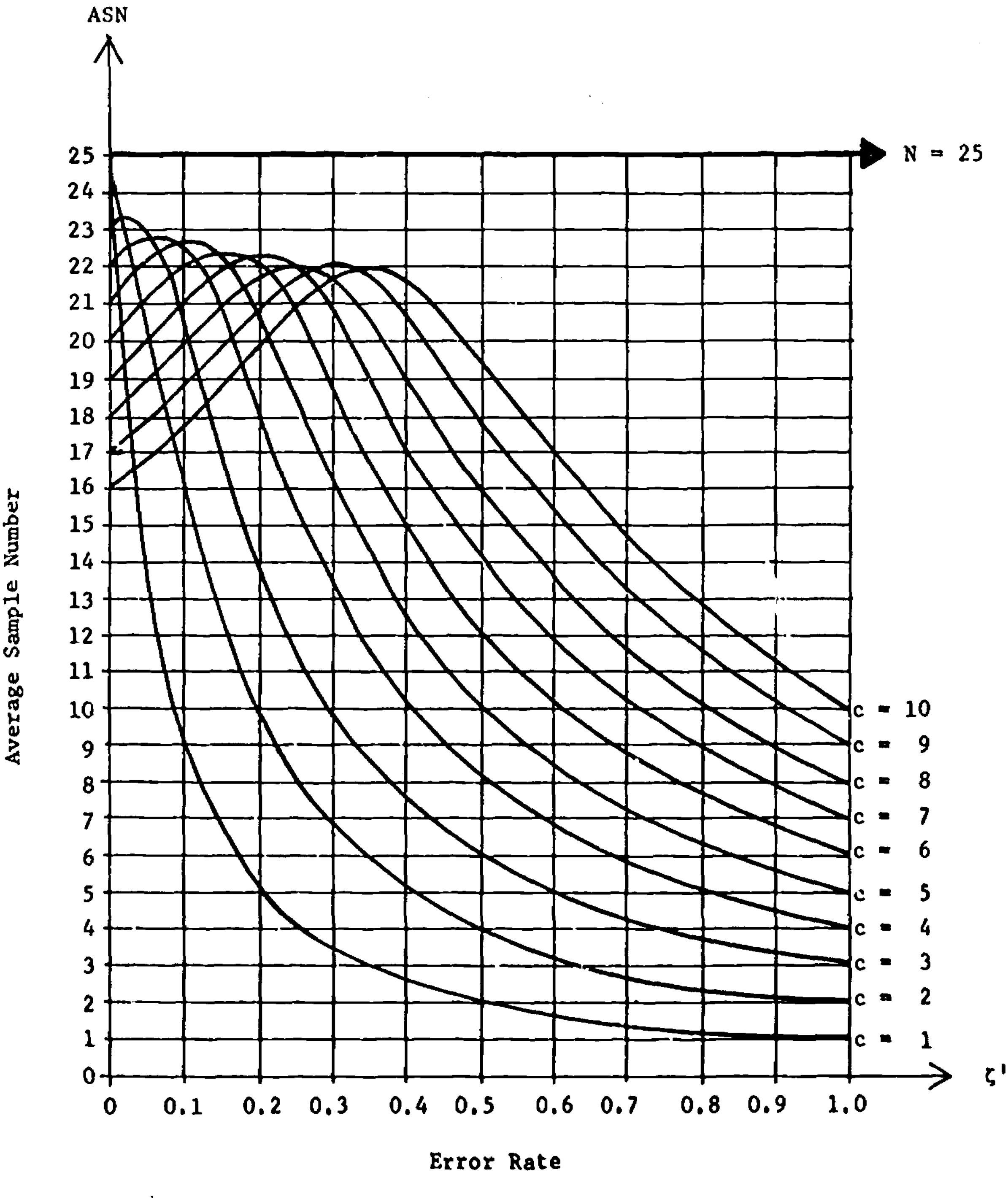
Average Sample Number for N = 5
And Error Criteria c = 1, 2

Figure 4



ASN Curves for N = 20
And Error Criteria c = 1, 2, ..., 8

Figure 5



ASN Curves for N = 25
And Error Criteria c = 1, 2, ..., 10

Figure 6