This is to certify that the

thesis entitled

A MODEL FOR CRITERION-REFERENCED MEASUREMENT
AND A COMPARISON OF ITEM ANALYSIS PROCEDURES

presented by

Susan K. Thrash

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

William McBrien
Major professor

Date 9-15-77

O-7639

A MODEL FOR CRITERION-REFERENCED MEASUREMENT AND A COMPARISON OF ITEM ANALYSIS PROCEDURES

By

Susan Kaye Thrash

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling and Personnel
Services and Educational Psychology

1977

ABSTRACT

A MODEL FOR CRITERION-REFERENCED MEASUREMENT
AND A COMPARISON OF ITEM ANALYSIS PROCEDURES

By

Susan Kaye Thrash

The first purpose of this study was to propose a theoretical conception of criterion-referenced testing and to explain two basic item analysis techniques (Cox and Vargas, C-V and Roudabush, R) theoretically with respect to this general model. The second purpose was to determine the adequacy of the C-V and R procedures using the theoretical model. The final purpose was to compare three item analysis techniques, the C-V, R and the Brennan and Stolurow (B-S), using real data.

A theoretical model for criterion-referenced testing was proposed. The model includes 12 parameters that completely described the pretest-posttest situation. The R and C-V indices can be explained in terms of this general model by making certain assumptions.

There were two parts to this study. The first part attempted to determine if the C-V and R indices adequately estimated the true values, if one technique estimated the true values better than the other and if the C-V and R indices were better estimators of the true values for some parameter sets. These questions were considered by simulating data for 21 different sets of parameter values using

the model as the theoretical framework. It was found that for R, when the assumptions were met, the technique provided a more stable and accurate estimate than when the assumptions were not met. It was also found that when the sample size was increased from 50 to 200, the stability and accuracy increased greatly. The C-V technique seemed to provide a reasonably accurate and stable estimate regardless of whether the assumptions were met. The estimates were more stable with larger sample sizes. Also, the C-V technique estimated the C-V true value better than the R technique estimated the R true value.

The second part of the study was designed to determine the comparability of the three item analysis procedures, R, C-V and B-S. C-V and R values were computed for 128 items and B-S values were computed for 64 items. These items were testing 16 objectives from two subject areas, Mathematics and Reading, two grade levels, Middle and Upper, and two treatments, assigned objectives (treatment A) and selected objectives (treatment B).

The major question to be answered was do the C-V, R and B-S item analysis procedures provide comparable results? Three additional questions were also considered:

1. Are the three procedures more comparable for items in Mathematics than for items in Reading?;
2. Does and the comparability of the three procedures depend on the grade level?; and,
3. Are the three procedures more comparable for items given in treatment A than for items given in treatment B?

The Pearson product moment correlation coefficient between the R and C-V indices was significantly different than zero ($r = .80$, $p < .01$). The point-biserial correlation coefficients between the B-S procedure and the C-V index and the B-S procedure and the R index were also significantly different than zero ($r = .70$, $p < .01$ and $r = .36$, $p < .01$, respectively).

The separate analyses of the indices for each subject area, grade level and treatment indicated that the indices were more comparable for Mathematics than for Reading. The indices were also more comparable for treatment B than for treatment A. The correlations between the indices for the grade levels, Middle and Upper, were almost identical.

An analysis of the agreement among the three item analysis procedures showed than when a cut-off of .50 for the R and C-V indices was used for selection of items, there was complete agreement for 39 of the 64 items (61 percent) given on the pretest and retention test.

From the results of the several analyses, it appears that the best item analysis procedure to use for criterion-referenced testing, or pretest-posttest situations, is the C-V technique. This technique provides a reasonably accurate and stable estimate of its true value and gives very similar results when compared to the R index and the B-S procedure.

TO THE W's IN MY LIFE

ACKNOWLEDGMENTS

There are many individuals who have contributed to this work as well as to my professional and personal growth. Dr. William Mehrens, my chairman, advisor and friend helped shape my ideas into a finished product, provided encouragement throughout my studies and gave me advice whenever I needed it. Dr. William Schmidt, who deserves a special thanks, spent a number of hours with me building the framework of this dissertation. Dr. Walter Hapkiewicz has provided me with constant attention throughout my graduate studies. His advice and concern for my educational progress has always been appreciated. Dr. Robert Spira, also a member of my committee, has had a significant impact on my educational achievement. Dr. Spira has had the faith and confidence in me to achieve what at times I was not sure I would be able to do. I will always be indebted to Dr. Spira for his meaningful comments and suggestions for my dissertation, professional goals and personal well-being.

I also wish to thank Joseph Wisenbaker, who assisted me with the computer simulation, four supervisors, Harley Jensen, Dr. Charles A. Pounian, Robert Joyce and Dan Wallock, who provided support and understanding during the trauma of the writing and rewriting of the dissertation, and my mother-in-law, Mrs. Marguerite Thrash, who also has provided me with support and encouragement during the completion of my graduate studies and dissertation.

I have saved for last the one individual who has inspired me the most, helped to build the self-confidence I lacked, listened to my ideas and helped to develop these ideas into a dissertation. I have reserved a very special thanks for this very special person--thank you--William Thrash.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DIAGRAMS

CHAPTER I

INTRODUCTION

Need

Criterion-referenced testing has been an area much discussed and researched in recent years. Much of the research and discussions have focused on the appropriateness of applying classicial measurement theory to criterion-referenced tests and suggestions of new procedures and statistics for the evaluation of criterion-referenced tests. Livingston (1971), for example, developed a new statistic for the estimation of reliability for criterion-referenced tests. Alternative approaches to classical item statistics were proposed by several other individuals (Brennan and Stolurow, 1971; Cox and Vargas, 1966; Roudabush, 1973 to mention a few). In addition, a few studies compared these new item statistics to old statistics (e.g. Cox and Vargas, 1966; Hambleton and Gorth, 1971; and Hsu, 1971).

Many of these new item statistics, however, were not based on a theoretical model. If such a model could be found, it would be easier to explain the item statistics and perhaps possible to develop more powerful statistical techniques. Moreover, little is known about the comparability of the new item statistics to each other. Most of the research has been concerned with the comparison of new with old; few studies have compared the new item statistics to each other. It would seem desirable to compare the new statistics both

empirically and theoretically, with the aid of a general model, to determine what the differences among them actually are and to develop general recommendations for their use.

Purpose

The first purpose of this study is to propose a theoretical conception of criterion-referenced testing and to explain two basic item analysis techniques (Cox and Vargas, and Roudabush) theoretically with respect to this general model.

The second purpose is to determine the adequacy of the Cox and Vargas and Roudabush techniques. If the two techniques can be explained by the general model, then the estimate of each index will be compared to the corresponding true value. In this manner, it may be possible to determine if one technique estimates the item parameters better than the other.

A third approach (Brennan and Stolurow) cannot be explained in terms of the general model due to the nature of the approach. The Brennan and Stolurow technique combines a number of statistics with a set of decision rules. The ultimate outcome is a verdict of revision or no revision for the item and/or the instruction. While the statistics used in the Brennan and Stolurow method do have the traditional theoretical framework, the decision rules have only intuitive appeal. It is not possible to fit the suggested decision rules of the Brennan and Stolurow technique into a theoretical framework.

However, the adequacy of the Brennan and Stolurow technique may be determined by comparison of the three approaches on real data. This, then, is the final purpose of the study--to determine the comparability of the three item analysis procedures (Cox and Vargas, Roudabush, and Brennan and Stolurow).[1] If all procedures provide identical or nearly identical results then it seems reasonable to use the simplest method (in terms of computation and data collection) in the future.

Research Questions

In particular, this investigation will consider the following questions:

1. Can a theoretical conception or a general model of criterion-referenced testing be defined?
   a. Does the C-V technique fit the general model? What assumptions are needed?
   b. Does the R technique fit the general model? What assumptions are needed?
2. Do the C-V and R techniques adequately estimate the true values of the item parameters?
   a. Does one technique estimate the true values better than the other?
   b. Do the C-V and R techniques estimate some true values of the item parameters better than the others?

---

[1] From this point on, Cox and Vargas, Roudabush, and Brennan and Stolurow techniques will be abbreviated C-V, R and B-S, respectively.

3. Do the C-V, R and B-S item analysis procedures provide comparable results?
   a. Are the three procedures more comparable for items in Mathematics than for items in Reading?
   b. Does the comparability of the three procedures depend on the grade level?

Overview

The previous section provided a brief introduction to the ideas and questions pursued in this study. Chapter II will provide a review of the literature relevant to item analysis methods for criterion-referenced tests. Two types of studies are considered--studies which proposed item analysis techniques (new and modifications of traditional approaches) and those which compared new techniques to old.

The third chapter presents a theoretical conception of criterion-referenced testing. The C-V index and the R sensitivity index are described in the context of this theoretical model.

A method for evaluation of the C-V index and the R sensitivity index with respect to the theoretical model is presented in Chapter IV. Procedures for determining the comparability of the C-V index, the R sensitivity index and the B-S method are also discussed in this chapter.

Chapter V presents the results of the evaluation of the C-V and R techniques with respect to the model. The results of the investigation of the comparability of the C-V, R and B-S indices in a practical application are presented in Chapter VI.

Finally, in Chapter VII some implications of the results of Chapters V and VI for test development are discussed, and some recommendations for further research on the proposed theoretical model are given.

CHAPTER II

REVIEW OF LITERATURE

The concept of criterion-referenced measurement in education has initiated many discussions and much research with respect to measurement issues. The main points of interest have been cut-off scores, reliability and item analysis. This review will summarize the literature on item analysis.

The literature can be divided into two categories. One group of studies can be collected under the heading of "proposed item analysis techniques." New techniques have been proposed by some (Brennan, 1972; Brennan and Stolurow, 1971; Cox and Vargas, 1966; Crehan, 1974; Hsu, 1971; Ivens, 1970, 1972; Kifer and Bramble, 1974; Kosecoff and Klein, 1974; Roudabush, 1973; Saupe, 1966) and the use of old (traditional) techniques have been advocated by others (Davis and Diamond, 1974; Ebel, 1973; Hambleton and Gorth, 1971; Harris, 1974; Nitko, 1971; Popham and Husek, 1969). The second category includes research which makes comparisons among the proposed techniques (Cox and Vargas, 1966; Crehan, 1974; Haladyna, 1974; Hambleton and Gorth, 1971; Helmstadter, 1974; Hsu, 1971; Ivens, 1970, 1972; Kosecoff and Klein, 1974; Ozenne, 1971).

Proposed Item Analysis Techniques

New Techniques

One of the earliest item analysis techniques proposed for criterion-referenced tests was suggested by Cox and Vargas in 1966 (Cox and Vargas, 1966). This procedure requires two administrations of the item--before and after instruction. The item statistic is then defined as the difference between the proportion of individuals answering the item correctly as posttest and the proportion of individuals answering the item correctly at pretest; C-V. (The original notation was D_{pp}.) This is the simplest technique to use; however, it has been criticized by Oakland (1972) and Davis and Diamond (1974).

Oakland claims that the C-V technique is limited because it is "more appropriately used to determine the extent to which students may profit from instruction rather than to determine the reliability estimates which apply to a particular CRM" (Oakland, 1972, p. 5). This is a strange criticism, for indeed the intent of the C-V procedure is to select items and not to provide reliability estimates. Oakland also criticizes the use of a statistical technique for item selection without regard to item content. This is a criticism which could be applied to the use of any statistical technique in the selection of items without regard for content.

Davis and Diamond suggest that use of difference scores make the C-V index unreliable. It should be remembered here that the statistic is not based on individual difference scores, but the difference of proportions. They also felt that the use of this statistic

without regard to the content of the items would impair the content validity of the final form of the test.

According to Davis and Diamond, test developers should use the same four basic principles that have been in use for 25-30 years. They do caution, however, using the second principle without regard to the content of the item. These principles are:

1. The items in an achievement test should constitute as nearly as possible a representative sample of the population of items that define the domain to be measures . . . .
2. The items in a predictor test, . . . , should constitute the set (drawn from the population of items that define the domain to be tested) which best predicts scores on the designated criterion variable in samples of examinees like those to whom the test will be administered . . . .
3. The items in an achievement test should, within the constraint imposed by principle 1, make up as efficient a measuring instrument as it is possible to produce.
4. Choice-by-choice item-analysis data should be used as a basis for editing and revising items for achievement, aptitude, and selection tests. (Davis and Diamond, 1974, pp. 128-131.)

Of course all these principles are ones that should be considered regardless of the referencing nature of the test. However, it does not necessarily follow that the items will be doing the proper job if these principles are followed.

Ebel (1973) supports the use of the C-V technique when the purpose of the evaluation is to determine the effectiveness of an instructional program. However, he indicates traditional item discrimination indices are appropriate when the purpose is to determine how well an individual has succeeded in a particular course of study.

Ozenne (1971) also recommends the C-V index. In his investigation of a method of measuring test sensitivity, Ozenne suggested that a test composed of items selected on the basis of the C-V index

would have the greatest sensitivity to instruction. Haladyna also recommends the use of the C-V index (Haladyna, 1976 and Haladyna and Roid, 1976). In fact, he feels that the C-V " . . . index comes conceptually closest to measuring CR item discrimination" (Haladyna, 1976, p. 12).

Other individuals have considered the C-V technique as a starting point for further modifications. Brennan (1972) proposed the B index, a variation of the C-V technique and the traditional D. The D statistic is defined as the difference in the proportion of individuals in the upper group answering the item correctly and the proportion of individuals in the lower group answering the item correctly. The upper and lower groups are generally defined as the top and bottom 27 percent of the individuals ranked on the total test. The B index is defined as the proportion of individuals in the mastery group (upper) who answer the item correctly minus the proportion of individuals in the nonmastery (lower) group who answer the item correctly ($B = U/n_1 - L/n_2$). This index differs from D in that different sample sizes in the upper and lower groups are allowed. The evaluator is then able to use one administration, define the upper and lower groups according to mastery or nonmastery or by some similar criterion, and select items on the basis of this index. Brennan also determined the exact distribution of the B index under the null hypothesis, $B = 0$. This allows the evaluator to compute confidence intervals for the item statistic.

Hsu had already suggested an identical procedure in 1971 (Hsu, 1971). He suggested that a predetermined cut-off score be

established which would classify individuals according to mastery or nonmastery. According to Hsu, the difference in proportions of those responding correctly in each group to a given item would be a meaningful discrimination index for items from criterion-referenced tests. This index is identical to the B index.

One of the major problems with this technique is the decision of what defines mastery and nonmastery. Once this problem is solved, then it is possible that there will be too few mastery students in a pilot administration of the item if the group is uninstructed. If the group has been instructed then there may be too many mastery students. In either case, U/n_1 or L/n_2 would provide somewhat less than stable proportions and the value of B may not provide an adequate indication of the item's usefulness.

A modification of the B index (and Hsu index) was introduced by Crehan in his 1974 study (Crehan, 1974). Crehan redefined the upper and lower groups as independent groups of instructed and uninstructed, respectively. This modification basically solves the problem of defining the mastery and nonmastery groups.

The B index as originally proposed by Brennan and Hsu or modified as suggested by Crehan is very similar to the C-V technique and traditional techniques. One advantage for using B is the ability to use a different number of individuals in the upper and lower groups. A second advantage is the ability to test the null hypothesis, $B = 0$. It must be remembered, however, that teachers are the most likely users of criterion-referenced tests. It seems unrealistic to expect teachers to use sophisticated statistical techniques to

select items. A further problem is the availability of probability levels for B. The table of probability levels is available through a computer program which Brennan developed. The other criticisms that were mentioned previously must also be considered in the final analysis of the B index.

A second index that Crehan proposed is defined as the proportion of consistent performances on logically parallel items. In other words, this index equals the number of individuals who fail both items plus the number of individuals who pass both items divided by the total number of individuals. This of course requires the development of logically parallel items which is not necessarily an easy task. In addition, it requires the administration of both sets of items at the same time. For a short test, the time factor would not be a particular problem.

Crehan also employed a third unique technique in his study. The items were ranked by having teachers respond to the question, "Which item would you choose if you were to give a one item test?" (Crehan, 1974, p. 257). This was done until the item pool was exhausted. Compared to all the other item analysis procedures proposed, this approach is the most subjective one.[1]

Another refinement of the C-V method was suggested by Edmonston, Randall and Oakland (1972). For their method consider the two by two table below for a given item:

---

[1]Crehan also used a random ranking of items as an item selection device. See the section on comparison of techniques for the results of Crehan's study.

Table 2.1
Categories for a Given Item

		Posttest	
		Pass	Fail
Pretest	Pass	p_{11}	p_{12}
	Fail	p_{21}	p_{22}

The important pieces of information, they claim, are p_{12} and p_{21}. A high value for p_{21} would indicate a good item. Items that were less discriminating would have high p_{12} values. The refinement seems unnecessary since the C-V index would be $p_{21} - p_{12}$ and provides information of one value relative to the other.

Schooley, et al. (1976) also recommend consideration of the proportion of individuals answering the item correctly (p) on pretest and posttest. They suggest that the proportion should increase from pretest to posttest. In addition, items that supposedly measure the same objective should have similar p values. Those that have inconsistent p values should be looked at and revised if necessary. Their approach is very similar to the C-V method since a comparison of the p values from pretest to posttest would give the same value as the C-V method.

Ivens also considered the C-V technique in addition to two indices of his own (Ivens, 1970, 1972; Ozenne, 1971). Iven's indices require three administrations of the same item to the same subjects. One of the indices is based on the expectation that there would be a

large change in performance from pretest to posttest and a small change from posttest to retest. Ivens calls this Index 2 and it is defined as (p post - p pre) (1 - |p retest - p post|) where p is the proportion of subjects passing the given item on the particular administration. The other index (Index 1) is defined as (1 - pre-post agreement) (post-retest agreement) where the agreement is the proportion of subjects whose item scores (pass or fail) were in agreement across the appropriate administrations.

His final recommendation, however, is that the C-V technique be used for item selection and the information obtained from Index 2 be used for item revision (Ivens, 1970). The two indices defined by Ivens need three administrations of the item. In most situations this would be a definite disadvantage. In addition, if there is a minimum amount of change from posttest to retest |p retest - p post| would be small and 1 - |p retest - p post| would be close to one. In this case, Ivens' Index 2 would be approximately equal to the C-V index.

Ivens' Index 1 is also intuitively appealing. However, Index 1 can have a high value--indicating a good item--and yet be a bad item. For example, if many students pass the pretest, fail the posttest and fail the retest, Index 1 would have a high value. Yet, revision of the item (and probably instruction) should be considered.

Kosecoff and Klein (1974) suggest two indices--an Internal Sensitivity Index (ISI) and an External Sensitivity Index (ESI). For the first index (ISI) consider the following table which categorizes only those individuals who answered Item 1 correctly at the posttest:

Table 2.2
Categories for Individuals Answering
Item 1 Correctly at the Posttest
(ISI)

		Posttest	
		Fail	Pass
Pretest	Fail	n_1	n_2
	Pass	n_3	n_4

where n_1 = observed frequency of students who answered Item 1 correctly on the posttest but failed the pre and posttest; n_2 = observed frequency of students who answered Item 1 correctly on the posttest but failed the pretest and passed the posttest; n_3 = observed frequency of students who answered Item 1 correctly on the posttest but passed the pretest and failed the posttest; and n_4 = observed frequency of students who answered Item 1 correctly on the posttest and passed the pretest and the posttest.

The index ISI is defined as $\frac{n_2 - n_1}{n_1 + n_2 + n_3 + n_4}$, which according to Kosecoff and Klein, provides a measure of an item's ability to discriminate between those who have and have not profited from instruction. Their interpretation of the index does not, however, follow from the definition. It is conceivable that the index could have a high value but all who passed the item at posttest also passed the item at pretest. How does the item then have the ability to discriminate those who have profited from instruction from those who haven't? If all the individuals who passed the item at posttest

also passed the item at pretest, the item could not be said to be sensitive to instruction.

Their second index (ESI) is the Cox and Vargas index. The two indices are identical. Kosecoff and Klein do, however, suggest a "correction for guessing" for the index. They use the Marks and Noll procedure, which is also used by Roudabush in the development of his index, to derive the correction for guessing (Marks and Noll, 1967; Roudabush, 1973). They claim to compute the expected cell frequencies and use these values in the computation of the ESI. However, their expected cell frequencies are true frequencies which are heuristically computed from sample frequencies. This aspect will be discussed in more detail when Roudabush's sensitivity index is presented. (See Chapter III and Appendix I).

A method based on the four possible outcome patterns for an item administered on two occasions was proposed by Popham in 1970 (Kosecoff and Klein, 1974; Ozenne, 1971). The familiar two by two table (see Table 2.3) was used in conjunction with computation of Chi-square values.

Table 2.3
Categories for a Given Item

		Posttest	
		Fail	Pass
Pretest	Fail	f_1 (n_1)	f_2 (n_2)
	Pass	f_3 (n_3)	f_4 (n_4)

First it is necessary to count the number in each category (f_1, f_2, f_3, f_4--following the notation presented in Table 2.3). Secondly, a "prototypic item" is defined by taking the median frequency of each outcome category over all items. Finally, a comparison is made between this prototypic item and the actual frequencies in the four categories for each item. Large Chi-square values would suggest that the item is considerably different than the typical item. One problem with the technique is that the items in the test must be fairly homogeneous to give meaningful results. A second problem is not knowing how large the Chi-square values need to be for one to infer that the item is atypical or bad.

Three other studies have proposed methods totally different from the basic two-way table--Cox and Vargas approach. Kifer and Bramble calibrated a criterion-referenced test using the Rasch model, which is a latent trait model (Kifer and Bramble, 1974). They felt that the Rasch model could determine which items fit the model and which items need revision. However, as in the Popham method, all items need to be sampling one trait; if not, some items may not fit but yet be good items. Item analysis was a subobjective of their study. Their main emphasis was the desire to generalize about the scores and obtain more precision concerning the extent to which a score represents passing a criterion.

Bayesian techniques were applied to item analysis by Helmstadter (1974). Three separate indices of item effectiveness are defined in terms of probabilities. The first is the probability that a subject knows the content given that the correct response was

selected. The probability that a subject does not know the content given that the incorrect response was selected defines the second index and the probability that a correct decision will be made about the examinee's knowledge of the content given the results of performance of that item is the third.

For these indices, P indicates a correct response, $\overline{P}$ an incorrect response, K knowledge and $\overline{K}$ no knowledge. The first index is denoted by $P(K|P)$, the second by $P(\overline{K}|\overline{P})$ and the third by P(correct decision) equal to $P(\overline{KP}$ or KP). Bayes' theorem then implies that

$$P(K|P) = \frac{P(P|K)P(K)}{P(P|K)P(K) + P(P|\overline{K})P(\overline{K})}$$

and

$$P(K|P) = \frac{P(\overline{P}|\overline{K})P(\overline{K})}{P(\overline{P}|\overline{K})P(\overline{K}) + P(\overline{P}|K)P(K)} .$$

Each of the subcomponents, such as $P(P|K)$ were established on the basis of the administration of an item. The probabilities $P(K|P)$, $P(\overline{K}|\overline{P})$ and P(correct decision) were then computed using these pieces of information. There is still the same problem with these indices of determining a cut-off value for the establishment of a knowledge group and a no knowledge group. These indices can use pretest - posttest data or a single administration.

Saupe was concerned with maximizing the reliability of difference scores (Saupe, 1966). He suggested that items possessing certain characteristics would make the maximum contribution to the

reliable measurement of change. According to his analysis, items with the following characteristics should be considered as good items:

1. Items with high item-total score discrimination indices for both initial and final administrations of the test.
2. Items with low item-total score discrimination indices when the total score criterion is from the final administration for items in the initial administration and from the initial administration for items in the final administration.
3. Items with high correlations between initial administration item score and final administration item score (Saupe, 1966, p. 224).

Saupe derived an index that could be used in the selection of items to measure change. Items with high values of this index would be selected and items with low values rejected. The index is based on the correlation of the change in the item score with the change in the total test score;

$$r_{dD} = \frac{r_{xX} + r_{yY} - r_{xY} - r_{yX}}{2\sqrt{1 - r_{xy}}\sqrt{1 - r_{XY}}}$$

where x and y represent item scores and X and Y represent total test scores.

Although Saupe was not directly concerned with criterion-referenced tests, his work has some applicability to it. Obviously items in a pretest-posttest situation are meant to measure change

and the index might have some usefulness in predicting those items which are sensitive to change.

The third criterion, however, seems inconsistent with results of criterion-referenced testing. This criterion specifies that an item with a high correlation between initial item score and final item score is a good item. This high correlation would be achieved only if there is some variance on the pretest (not all individuals fail) and some variance on the posttest (not all individuals pass). In addition a high positive correlation is not obtained if an item is failed by most on the pretest and passed by most on the posttest. This is the situation desired in criterion-referenced testing. A high correlation would not designate items sensitive to instruction. Criterion two suggests that discrimination indices should be low between item and total score using the opposite administration for the criterion. Again these low discrimination values could be obtained and yet the item might be a bad item. For example, a low discrimination value could be obtained with almost all passing the item at pretest and getting low scores on the posttest. A similar situation would result with almost all failing the item on the posttest and obtaining somewhat high scores on the pretest. These results are not desirable in criterion-referenced testing. Items exhibiting these characteristics might not be good items.

As with almost all of these techniques, care must be made to include items that cover all the objectives. Relying on only statistics to select items may result in the exclusion of some important aspects that need to be tested. Nitko, when considering this

problem, suggested "that tests constructed from carefully defined domains of items possess reasonably good psychometric properties without prior statistical selection" (Nitko, 1971, p. 8). On the other hand, Skager felt that "relying solely upon judgments as an index of item quality ought to leave us just as uneasy in the case of criterion-referenced tests as it should be for norm-referenced instruments" (Skager, 1974, p. 53). One of his suggestions was the use of item generation rules; although, he indicated that item selection for criterion-referenced tests is still open to debate.

Hambleton, et al. (1975) also do not advocate the use of empirical techniques exclusively. They feel that items selected should be representative of the domain of items and the empirical methods should be used to detect bad items.

Consideration should also be given to the impact of selecting items that are sensitive to instruction according to some statistic. If items are selected which are sensitive to instruction one might argue that the items, over a number of administrations and revisions, could become very easy or perhaps require only recall of simple facts. Care must be taken to include items that measure all aspects of the domain and to ensure that these items are not only sensitive to instruction but sensitive to the domain.

Another approach similar to the C-V index was presented to Roudabush at the 1973 American Educational Research Association Annual Meeting. It is based in part on a procedure suggested by Marks and Noll (1967). As it was pointed out earlier, Kosecoff and

Klein used a similar technique to develop the "correction for guessing" for their External Sensitivity Index.

Roudabush's technique is based on the familiar two by two table presented earlier as Table 2.3. Roudabush also makes two assumptions. First, he assumes that there is some fixed non-zero probability, p, that a student who does not know the answer to the item will guess the correct answer. This p value is determined by the item only and does not vary from student to student nor from occasion to occasion for the same student. This fixed p value suggests that there is no partial knowledge on the part of the student, and that the student's responses are independent at pretest and posttest when he does not know the correct answer and fails to learn it.

Further, Roudabush assumes that the only possible result of exposure to instruction between pretest and posttest is that the student learns the correct response to an item. This then implies that the non-zero frequency of f_3 is solely due to guessing, further implying that there is no forgetting. This suggests that the "true" value of f_3 is zero.

With these assumptions Roudabush derives a number which serves as an index of the degree to which examinees select the correct response to the item as a function of the instruction received between pretest and posttest. This number is called a sensitivity index by Roudabush. It can be expressed as

$$R = \frac{f_2 - f_3}{f_1 + f_2} .$$

(The original notation was s.) Further clarifications and derivations are presented in Chapter III and in Appendix I.

Traditional Techniques

Traditional item analysis procedures also have been recommended for use with criterion-referenced tests. Most individuals have, however, suggested some modifications in the interpretation of these traditional indices. One of the more detailed procedures is outlined by Brennan and Stolurow (1971).

Their procedure combines traditional item analysis techniques with a set of decision rules. Brennan and Stolurow compute four error rates and two discrimination indices from pretest, posttest and retention test data. The decision rules are then applied to determine the adequacy of the item and of the instruction. The decision rules are similar in context to the first criterion of a good item suggested by Saupe. Further clarifications of this technique are presented in Chapter IV and Appendix II. Their procedure is very complicated and laborious and for this reason, perhaps, has not been investigated further.

Other individuals have also recommended the use of traditional indices. Hsu recommends the use of the phi-coefficient with Right versus Wrong for a given item being one dimension and Mastery versus Nonmastery the other (Hsu, 1971). For this procedure, a cut-off score for each behavior must be established in order to declare a mastery and a nonmastery group. There are other limitations besides the problem of establishing a cut-off score. The phi-coefficient

cannot be used when the item is answered correctly or incorrectly by all or when all subjects are declared masters or nonmasters. Hsu then recommends the use of his upper-lower difference statistic, defined as the difference in proportions of those responding correctly in the mastery and nonmastery groups, or the point-biserial correlation coefficient. Hsu's upper-lower difference statistic was discussed in the previous section.

Hambleton and Gorth (1971) also suggest using traditional item analysis procedures. Items associated with the same objective should have approximately the same value for item difficulty. Items that are different should be modified and tested again. In addition item discrimination indices can be used. Negative indices would indicate a need for revision in the item, instructional materials, and/or teaching. Positive discrimination indices, according to Hambleton and Gorth, more than likely indicate a shortcoming in the instructional program. Items with zero discrimination may be acceptable. Popham and Husek recommended the same interpretations of discrimination indices in 1969 (Popham and Husek, 1969).

If the traditional methods and the interpretations suggested by Hambleton and Gorth and Popham and Husek are used, then the information that is obtained seems to be ambiguous and no definite decision can be made about the item. However, Brennan and Stolurow took these bits of information with other information and a set of rules and have developed a useful guide for item selection for criterion-referenced tests.

Item characteristic curves, another traditional item analysis technique, can also be used for criterion-referenced tests (Hambleton and Gorth, 1971). The parameters (difficulty and discrimination) of the curves supposedly do not change from group to group. This implies that the parameters could be predicted from the pretest administration. An obvious disadvantage in using item characteristic curves would be in the construction and the interpretation of them. This procedure would not be one of the easiest to use or understand.

Harris also suggests traditional item analysis techniques for criterion-referenced tests. However, the test should be used with a sample from a population of instructed students and a sample from a population of uninstructed students. Item difficulties for items for a given objective should be equal within each of the two groups; however, item difficulties should differ between the two groups (Harris, 1974). Woodson's position is very similar to Harris' position. Woodson argues that the item needs to be tested in the proper population. He feels that "items and tests must be evaluated for the range of the characteristic for which they will be used" and if the items and tests give no variability in this population of observation, then the items and/or tests give no information and are not useful (Woodson, 1974, p. 64).

Both of these suggestions are considered when pretest and posttest data are used. The pretest group is generally considered the uninstructed group and the posttest group the instructed group. The B-S decision process includes a comparison of the pretest and posttest item difficulties and the C-V index and R index are

comparisons of the pretest and posttest difficulties. Since most of the other proposed item analysis techniques also consider pretest and posttest data, the Harris and Woodson suggestion of testing the item in a proper population are taken into account.

Summary

The various techniques that have been proposed fall into essentially two categories. One category of techniques contain the C-V technique and its variations (Brennan, 1972; Crehan, 1974; Edmonston, Randall and Oakland, 1972; Hsu, 1971; Ivens, 1970, 1972; Kosecoff and Klein, 1974). The other category contains item analysis procedures generally used for norm-referenced tests with possible alternative interpretations. As is discussed above, these new meanings for old statistics sometimes result in a technique or procedure which is similar to the C-V procedure. Every new technique seems to have as its main purpose, selecting items that are sensitive to instruction. However, there is a need to be alert to the negative implications of selecting items sensitive to instruction. Most individuals recommend using item statistics in conjunction with a review of the domain or objectives and close scrutiny of the instruction. This aspect will be discussed more thoroughly in the final chapter.

Review of the proposed techniques has shown that the C-V index or modifications of the C-V index have been recommended more frequently than any other procedure as an appropriate item analysis technique for criterion-referenced tests. The R technique is a refinement of the C-V technique and, as it will be shown in the

following chapter, makes fewer assumptions than the C-V index. Therefore, the R index may provide a better estimate of an item's sensitivity to instruction than the C-V index.

The B-S procedure combines the best of traditional methods in an attempt to select good items for criterion-referenced tests. All three of these procedures may be considered useful in selecting items that are sensitive to instruction. Most of the remaining procedures are latent trait models. While these are useful they fail to meet the criterion of computational ease which is important in most of the situations where criterion-referenced tests are used.

Comparing Techniques

Several studies have been done to compare new item statistics to old item statistics. Crehan (1974) compared six item analysis techniques using a pool of items constructed by teachers. The procedures he compared were the C-V, a modified Brennan, a teacher rating, a point-biseral correlation between item score and total test score in the posttest situation, a random ranking, and an index which was defined as the proportion of consistent responses on logically parallel items.

Crehan used the concepts of reliability and validity to compare tests composed of items selected by each of the six techniques. Reliability was estimated by $(a + c)/N$ where $N = a + b + c + d$ and a, b, c, d are defined in Table 2.4 below.

Validity was estimated by $(a + c)/N$ where $N = a + b + c + d$ and a, b, c, d are defined differently in Table 2.5 below.

Table 2.4
Categories of Performance (Reliability-Crehan)

		Form B	
		Pass	Fail
Form A	Pass	b	a
	Fail	c	d

Table 2.5
Categories of Performance (Validity-Crehan)

	Uninstructed Group	Instructed Group
Pass	b	a
Fail	c	d

In addition validity was estimated by the point-biserial correlation between test score and a dummy variable representing group membership (instructed group and uninstructed group). The instructed group was a posttest only group and the uninstructed group was a pretest group.

The results of his study suggested that the modified Brennan and C-V methods produced tests with higher test validity. However, the different item selection methods seemed to have no effect on test reliability.

In order to generalize from the results of this study, the definitions of reliability and validity employed by Crehan must be accepted as reasonable. Both definitions are rationally appealing if not theoretically appealing. Reliability could also have been estimated with a phi-coefficient. But with either method the determination of cut-offs is arbitrary and the estimates can increase or decrease with shifts of the cut-offs. Validity could also have been estimated with a phi-coefficient. The same problem exists, however, with determination of cut-offs and assignment to pass or fail groups. The point-biserial, which was also used to estimate validity, does not have the problem of determination of cut-offs.

Two groups of individuals were included in the sample. One group was used to compute item statistics, develop tests and set passing points. The other group was used to determine reliability and validity. The process was reversed and reliability and validity estimates obtained from both groups were averaged. This is unfortunate since it seems reasonable to think of one group as the cross-validation sample. The obtained reliability and validity estimates from both groups could then have been compared and inconsistencies located.

Item statistics were not compared across samples of individuals, even though those data were available. Questions such as how did the item values fluctuate across samples and across subject areas were not considered in this study.

The only conclusion that we can draw from this study is that if the C-V or modified Brennan techniques for selection of items for

criterion-referenced tests are used, the validity, as defined by Crehan, might be better than if some other technique for selection were used.

Several other individuals have also compared the C-V index to alternative methods (Cox and Vargas, 1966; Haladyna, 1974; Haladyna and Roid, 1976; Hambleton and Gorth, 1971; Hsu, 1971; Ivens, 1970, 1972; Kosecoff and Klein, 1974). It is interesting to note that of the 11 studies that are reported here which compare criterion-referenced item analysis techniques, eight include the C-V method. This index has to be appealing because of the ease of computation. In addition it seems to fare extremely well in the comparisons with other techniques.

Cox and Vargas (1966) and Hambleton and Gorth (1971) concluded that the C-V index produces results different enough from traditional methods to warrant the consideration of this alternative technique for criterion-referenced test construction. Cox and Vargas compared D to C-V and Hambleton and Gorth compared C-V to the biserial correlation and a modified C-V. The modified C-V was defined as the difference between the proportion of individuals who correctly answered an item on the delayed posttest and the proportion of individuals who correctly answered the same item on the pretest, C-V'. While Hambleton and Gorth found no relationship between C-V and C-V' with the biserial, Cox and Vargas did find significant Spearman rank order correlations between the rank on C-V and the rank on D.

Haladyna, on the other hand, concluded from his study that a point-biserial discrimination index computed on the combined test

results of pre and post-instruction examinees is better than C-V. His conclusion is based on the result of his analysis which indicated that the two statistics give identical information and the point-biserial requires a one-step analysis and the C-V requires a two-step analysis. His argument that the point-biserial is a one-step process is based on the availability of computer programs to compute the correlations. For a classroom teacher C-V has the advantage of being easy to compute as well as "conceptually satisfying" (Haladyna, 1974, p. 98).

Hsu investigated the relationship of a modified C-V (C-V") with r_{pbi} and the phi-coefficient using various samples of individuals (Hsu, 1971). The index C-V" is defined as the difference in proportions of those responding correctly in a mastery and nonmastery group. The mastery and nonmastery groups are established by a predetermined cut-off score. The samples varied with respect to the ability dimension and test score distribution. The results indicated that the relationship of C-V", r_{pbi}, and the phi-coefficient depends on the ability dimension and the test score distribution. When the sample consists of individuals with a wide variety of abilities and the test scores are distributed symmetrically the indices are highly correlated.

Hsu found that a highly discriminating item in one sample may not be a highly discriminating item in another; therefore, he recommended that test items not be tried out in a group with a wide variety of abilities. Items selected on the basis of performance of this group may not be measuring the same kind of performance in a second more homogeneous group.

Ivens also investigated the C-V index (Ivens, 1970, 1972). He found that by choosing items with larger values for C-V for one test and lower values of C-V for a second test, there were marked differences in the quality of the tests. To measure the quality of the test, Ivens considered reliability and validity. He used traditional reliability estimates as well as unique reliability and validity estimates. All statistics computed supported the fact that tests composed of items with higher C-V values were better tests. It should be pointed out that the unique reliability and validity estimates were somewhat related to C-V. For this reason, higher reliability and validity estimates for tests constructed from items with high C-V values would be expected.

The C-V index was again compared to other indices by Kosecoff and Klein (1974). They redefined C-V as ESI and compared this to their ISI, the phi-coefficient and the point-biserial. (ESI and ISI are defined in an earlier section of this chapter.) The results of this study showed that ESI was generally lower than ISI. The values of ISI tended to parallel the values of the point-biserial and phi-coefficient. Of course, the corrected version of ESI resulted in lower values.

After consideration of the data, Kosecoff and Klein determined that there had been too many masters at the pretest. To compensate for this, ESI and ISI were redefined. ESI was defined as $\frac{n_2 - n_1}{n_1 + n_2}$ (Table 2.3 and Table 2.2 notation, respectively). They concluded from the results of the analysis with the redefined

statistics that ISI is sensitive to instruction. The high proportion of prior masters caused the index in the first analysis to be artificially deflated. ESI was found to be an unsatisfactory statistic because the values tended to vary greatly. The values for ESI did correlate significantly with the phi-coefficient and point-biserial values but the correlation coefficients were rather small implying, perhaps, that ESI would not give the same judgment as traditional statistics. Almost all the research that has considered the C-V index (or the ESI) has produced this same result.

Interest in the C-V index remains high as indicated in a recent comparative study conducted by Haladyna and Roid (1976). They compared various Rasch statistics, traditional statistics, the Bayesian indices proposed by Helmstadter (1974), and the C-V index for a total of 17 indices. The results of the study demonstrated a high degree of relationship among four item discrimination measures. These were the z-difference--a Rasch statistic which is an index of the difference of difficulties of pretest and posttest samples, a combined samples point-biserial, the C-V index and a Bayesian index --the probability of having knowledge given that the student gets the item correct. This study provides further evidence that the C-V index may be the most appropriate item index for pretest-posttest situations.

Three comparative studies that did not include the C-V technique are Roudabush (1973), Helmstadter (1974), and Bernkopf (1976). Roudabush and Helmstadter compared their own unique indices to traditional statistics. Unfortunately neither study mentioned exactly

which traditional statistics were being used. Roudabush concluded that his sensitivity index provided different information than the traditional statistics. Helmstadter, on the other hand, found that the "classical discrimination index [he defined it no further than this] comes closest to providing the same item assessment as would the Bayesian probability of making a correct decision . . . ." (Helmstadter, 1974, p. 3). Haladyna and Roid (1976) confirmed Helmstadter's result in their study. On the basis of the analysis, Helmstadter also concluded that "items which are effective indicators that the examinee does know the material are not necessarily the same items which are effective indicators that the examinee does not know the material" (Helmstadter, 1974, p. 3).

Bernkopf compared the point-biserial coefficient using total test score as a criterion (r_t), the phi-coefficient (ϕ_e), and a second point-biserial coefficient using the total score on an essay test as a criterion (r_e). The dimensions of the fourfold table for the phi-coefficient were correct/incorrect for the item and above/below mastery on an independent criterion (the essay test). All three indices were significantly related. As could be expected the correlations between the ϕ_e and r_e were higher than the correlations between ϕ_e and r_t and r_e and r_t.

Summary

The literature reviewed in this chapter has been divided into two categories. The first group of studies reviewed, recommends possible approaches for criterion-referenced item analysis (e.g.

Brennan, 1972; Brennan and Stolurow, 1971; Cox and Vargas, 1966; Crehan, 1974; Hambleton and Gorth, 1971; Hsu, 1971; Ivens, 1970; Kifer and Bramble, 1974; Kosecoff and Klein, 1974; Roudabush, 1973). The second group of studies compares a number of proposed techniques (e.g. Cox and Vargas, 1966; Crehan, 1974; Haladyna, 1974; Hambleton and Gorth, 1971; Hsu, 1971; Ivens, 1970; Kosecoff and Klein, 1974).

Review of the proposed techniques reveals that the C-V index or modifications of this index have been recommended more frequently than any other procedure as an appropriate item analysis technique for criterion-referenced tests. In addition, the majority of the comparative studies included the C-V index along with more traditional indices. The general conclusion is that tests constructed on the basis of the C-V index result in tests sensitive to instruction (Ivens, 1970, 1972; Ozenne, 1971). Another conclusion is that the C-V index results in a different judgment for a given item than traditional statistics (Cox and Vargas, 1966; Kosecoff and Klein, 1974).

Only two studies included more than one new index in their comparisons (Crehan, 1974; Haladyna and Roid, 1976). The C-V index is significantly related to other new approaches--a Rasch statistic and an index recommended by Helmstadter (Haladyna and Roid, 1976) and when used produces tests with higher validity (Crehan, 1974).

Two new approaches to criterion-referenced item analysis have not been researched--one, the R index and two, the B-S procedure. The R index is a refinement of the C-V technique. It makes fewer assumptions and may be a better estimate of an item's sensitivity

to instruction. The B-S procedure combines traditional methods with a set of rules to provide a guide for selecting items which are sensitive to instruction. For these reasons, the C-V index, the Roudabush sensitivity index (R) and the Brennan and Stolurow procedure (B-S) were selected for further investigation.

In the following chapter a theoretical basis for criterion-referenced testing or pretest-posttest situations is provided. It will be shown that the C-V index and R index can be explained in terms of a general model; and, as indicated above, it will be shown that the R index is a refinement of the C-V index which requires fewer assumptions.

CHAPTER III

THEORETICAL DISCUSSION

In this chapter, a theoretical model for the pretest-posttest situation is presented. Two item analysis techniques, R and C-V, which were described earlier, are explained in terms of the general model.

The results of a given item in any test can be represented by the following diagram:

Table 3.1
Categories for a Given Item

		ACTUAL	
		Does Not Know	Knows
OBSERVED	Fail	q_{11}	q_{12}
	Pass	q_{21}	q_{22}

where q_{11}, q_{21}, q_{12} and q_{22} are conditional probabilities with $q_{11} + q_{21} = 1$ and $q_{12} + q_{22} = 1$.

The probability that an individual who does not know the answer to a given item will answer the item incorrectly is denoted by q_{11}. The probability that an individual who does not know the

answer to the given item will answer the item correctly is denoted by q_{21}. Similarly, q_{12} and q_{22} represent the probabilities that an individual who knows the answer will fail or pass the item, respectively.

Now consider a pretest-posttest situation. This can be represented with three diagrams. Table 3.1 can be used to define the pretest results and a similar table with different probabilities (Table 3.2 below) can represent the posttest. These probabilities are defined in the same manner as above.

Table 3.2
Categories for a Given Item

		POSTTEST - ACTUAL	
		Does Not Know	Knows
OBSERVED	Fail	q_{11}'	q_{12}'
	Pass	q_{21}'	q_{22}'

An additional 2 x 2 table (Table 3.3 below) defines the true proportions of the pretest-posttest situation.

Table 3.3
True Proportions for a Given Item

		POSTTEST	
		Does Not Know	Knows
PRETEST	Does Not Know	π_1	π_2
	Knows	π_3	π_4

In Table 3.3, π_1 is the proportion of individuals who do not know the answer to a given item at both pretest and posttest. Similarly, π_2 is the proportion of individuals who do not know the answer to a given item at pretest but learn it by the posttest. π_3 is the proportion of individuals who know the answer at pretest but not at posttest; and π_4 is the proportion who know the answer at both times. These proportions, π_1, π_2, π_3, π_4, sum to one. These are true proportions. They are not the observed results of the pretest and posttest.

The general model is then represented in matrix notation as $P = Q \otimes Q' \underline{\pi}$ where $\otimes$ symbolizes the Kronecker product,

and

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}, \quad Q' = \begin{pmatrix} q_{11}' & q_{12}' \\ q_{21}' & q_{22}' \end{pmatrix}, \quad \underline{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix},$$

and

$$\underline{P} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} \quad \text{or,}$$

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} \otimes \begin{pmatrix} q_{11}' & q_{12}' \\ q_{21}' & q_{22}' \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} .$$

The p_k's, described in Table 3.4, are the observed proportions given the probabilities q_{ij} and q_{ij}' and the true proportions π_k.

Table 3.4
Observed Proportions for a Given Item

		POSTTEST	
		Fail	Pass
PRETEST	Fail	p_1	p_2
	Pass	p_3	p_4

Expanding the model,

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11}q_{11}'\pi_1 + q_{11}q_{12}'\pi_2 + q_{12}q_{11}'\pi_3 + q_{12}q_{12}'\pi_4 \\ q_{11}q_{21}'\pi_1 + q_{11}q_{22}'\pi_2 + q_{12}q_{21}'\pi_3 + q_{12}q_{22}'\pi_4 \\ q_{21}q_{11}'\pi_1 + q_{21}q_{12}'\pi_2 + q_{22}q_{11}'\pi_3 + q_{22}q_{12}'\pi_4 \\ q_{21}q_{21}'\pi_1 + q_{21}q_{22}'\pi_2 + q_{22}q_{21}'\pi_3 + q_{22}q_{22}'\pi_4 \end{pmatrix}$$

This model completely describes the results of a pretest-posttest situation.

For example, consider p_1, the observed proportion of individuals who fail both the pretest and the posttest. Each of the actual proportions, π_1, π_2, π_3, and π_4 can contribute to the observed proportion. In the model $p_1 = q_{11}q_{11}'\pi_1 + q_{11}q_{12}'\pi_2 + q_{12}q_{11}'\pi_3 + q_{12}q_{12}'\pi_4$. If we consider π_1, the proportion of individuals who do not know the answer at pretest or posttest, we can observe that some of the individuals in this category could have guessed correctly at either the pretest (q_{21}) or the posttest (q_{21}') or at both the prestest and the posttest. These individuals would not contribute to the observed proportion p_1, since they would have passed the item at one or both times. However, we can include $q_{11} \times q_{11}' \times \pi_1$ which is the

proportion of individuals who really don't know and didn't learn and failed to guess at either administration. Individuals who did learn the correct response from pretest to posttest can also contribute to p_1. Those contributing would have failed to guess the correct response at pretest (q_{11}) and would have answered incorrectly at the posttest (q_{12}') even though they knew the correct response. Therefore, $q_{11} \times q_{12}' \times \pi_2$ adds to the observed proportion p_1. In addition, individuals who do know the answer at the pretest but don't know the answer at posttest (π_3) contribute to p_1. Ordinarily, we would not expect π_3 to be a very large proportion. Individuals who can be classified in this manner could have failed to respond correctly at the pretest (q_{12}) even though they knew the answer and could have failed to guess the correct answer at the posttest (q_{11}'). Finally, individuals knowing the answer at both pretest and posttest could have answered incorrectly at both administrations ($q_{12} \times q_{12}' \times \pi_4$). Therefore, we can see, intuitively, that p_1 is the sum of parts of each of the proportions π_1, π_2, π_3, and π_4. The observed proportions p_2, p_3, and p_4 can be explained in a similar manner.

It should be noted that π_1, π_2, π_3, π_4 are separated among each of the observed proportions. If, for example, we add all the parts of π_1, which are distributed over p_1, p_2, p_3 and p_4, then $q_{11}q_{11}'\pi_1 + q_{11}q_{21}'\pi_1 + q_{21}q_{11}'\pi_1 + q_{21}q_{21}'\pi_1$ should equal π_1. This can easily be shown by factoring this expression:

$$q_{11}(q_{11}' + q_{21}')\pi_1 + q_{21}(q_{11}' + q_{21}')\pi_1 =$$
$$(q_{11} + q_{21})(q_{11}' + q_{21}')\pi_1 = \pi_1 \text{ since}$$
$$q_{11} + q_{21} = 1 \text{ and } q_{11}' + q_{21}' = 1.$$

It can also be shown that all the parts of π_2, π_3, and π_4, which are distributed over the observed proportions, p_1, p_2, p_3 and p_4, do sum to π_2, π_3, and π_4, respectively.

There are 12 parameters in this model. If these parameters could be estimated, useful information would be available for both the item and the instruction. For example, if π_2, the proportion of examinees who learn the answer, could be estimated, then an evaluation of the quality of the instruction could be made. The estimate of this proportion would also indicate the item's "sensitivity to instruction."

Estimates of the other parameters would also provide useful information. For an objective item, estimates of q_{11}, q_{21}, q_{11}' and q_{21}' can be made after consideration of the number of response choices. For example, a four-choice objective item would ordinarily lead to an estimate of .25 for q_{21} or q_{21}', because an individual who does not know the answer has one chance out of four of choosing the correct response. It is also generally assumed that q_{22} and q_{22}' equal 1.0, because it is very unlikely that an individual who knows the answer will respond incorrectly. However, this may not be the case for a poorly-written item. For example, a distractor for an item may be also a correct response; or, the correct alternative could be worded so ambiguously that even the individual who knows the answer will not choose it. There is also the possibility that an individual will make a clerical error.

Estimates of the q_{ij}'s and q_{ij}''s do provide information about the quality of an item. A bad item would be one where q_{21}

or q_{21}' is high; that is, where the probability of guessing is high. A good item would be one where q_{22} and q_{22}' approach equal 1.0.

Suppose the parameters are considered in a slightly different manner. One could perhaps use the concepts of reliability and validity to describe these parameters. The π_k's represent true values. Estimates of indices defined by the π_k's are estimates of the validity of the item. For example, an estimate of π_2 indicates how many or what proportion of the individuals not knowing at the pretest know at the posttest. The higher this value, or closer this number is to 1.0, the better the item is measuring what it is supposed to measure. In other words, indices based on the π_k's are indicators of validity.

In addition some of the q_{ij}'s and q_{ij}''s can be considered to be estimates of reliability. For example, if q_{11}, q_{22}, q_{11}', and q_{22}' are close to 1.0 then the item is a perfect indicator of knowledge or no knowledge. As these probabilities decrease the item is a less reliable indicator of knowledge or no knowledge.

Assumptions can be made to simplify this conceptualization. In the general model Q does not necessarily equal Q'; different probabilities are defined for the pretest and posttest. It is possible, however, that for any given item these probabilities would be identical; that is, that neither time nor instruction would change these item parameters. One could then assume that $Q = Q'$.

Roudabush simplifies the situation even further. First, he assumes that $\pi_3 = 0$. This implies that there is no forgetting; an individual who knows an item at pretest will know it at posttest.

Second, Roudabush assumes $q_{22} = q_{22}' = 1.0$, ignoring the possibility that someone who knows the answer to an item could fail it.

Under these assumptions the model reduces to:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11} & 0 \\ q_{21} & 1 \end{pmatrix} \boxtimes \begin{pmatrix} q_{11} & 0 \\ q_{21} & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ 0 \\ \pi_4 \end{pmatrix}$$

But $q_{11} + q_{21} = 1$ and $\pi_1 + \pi_2 + \pi_4 = 1$, so

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11}^2\pi_1 \\ q_{11}(1 - q_{11})\pi_1 + q_{11}\pi_2 \\ (1 - q_{11})q_{11}\pi_1 \\ (1 - q_{11})^2\pi_1 + (1 - q_{11})\pi_2 + (1 - \pi_1 - \pi_2) \end{pmatrix}$$

These four equations correspond to equations (1) through (4) presented in Appendix I.

The sensitivity index is defined as $R = \dfrac{\pi_2}{\pi_1 + \pi_2}$.

This is a reasonable sensitivity index, it is the proportion of individuals not knowing the answer at pretest who learn it by the posttest.

Roudabush solves the four equations above using the assumption that the expected observed proportions, p_1, p_2, p_3, p_4 equal the sample proportions, f_1/N, f_2/N, f_3/N, f_4/N respectively and obtains solutions for π_1 and π_2 in terms of f_1, f_2, f_3, and f_4. The f_1, f_2, f_3 and f_4 equal the observed numbers of individuals in each category and N is the total number of individuals. These

solutions are then substituted in the definition of R and an estimate of R is $\frac{f_2 - f_3}{f_1 + f_2}$.

Unfortunately, the general model cannot be heuristically solved since there are seven parameters (unknown) and only three pieces of information. Therefore, we cannot estimate R without Roudabush's assumptions. We can, however, compare the true R and the estimated R for simulated data.

A second index, suggested by Cox and Vargas, can be considered in the same theoretical framework. Cox and Vargas call their index the Pretest-Posttest Difference Index (C-V). This is defined as the percentage of students who pass the item at posttest minus the percentage of students who pass the item at pretest. In terms of observed results, this is $\frac{f_2 + f_4}{N} - \frac{f_3 + f_4}{N}$ or $\frac{f_2 - f_3}{N}$.

The C-V method can be represented as a special case of the general model by assuming that $Q = Q'$, $q_{22} = q_{22}' = 1.0$, and $q_{21} = q_{21}' = 0$. Then,

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \boxtimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_3 \end{pmatrix}.$$

The C-V index can then be defined, using the notation of Table 3.3, as $(\pi_2 + \pi_4) - (\pi_3 + \pi_4)$ or C-V (true) $= \pi_2 - \pi_3$. This is identical to the definition of C-V given by Cox and Vargas except they use the observed proportions as estimates of the actual proportions. This index indicates the sensitivity of the item to instruction. The

closer C-V (true) is to 1.0 the greater the sensitivity and the closer it is to 0.0 the less the sensitivity.

If the equations above are solved heuristically for the true proportions they are found to be equal to the observed proportions. In other words, under these assumptions, the observed proportions are equal to the true proportions. These assumptions, however, are extremely restrictive; they do not even allow for guessing. In fact the C-V approach assumes no misclassification, i.e., no error. C-V is an estimate of C-V (true). Under certain restrictive assumptions C-V would equal C-V (true). We can compare C-V (true) with C-V for simulated data in order to observe the impact of less restrictive assumptions on C-V.

Summary

In this chapter, a theoretical framework is proposed for criterion-referenced testing in pretest-posttest situations. This framework suggests that 12 parameters completely describe the pretest-posttest situation. In addition the Roudabush (R) model and the Cox and Vargas (C-V) technique are explained in terms of the general model.

The design of the research is discussed in the following chapter. The research is considered in two parts. In the first part of the chapter, the design of the simulation study is presented. The simulation study uses the theoretical framework proposed in this chapter to consider the impact of various assumptions on the C-V and the R indices. The design of the comparison of the C-V, R and B-S

techniques with actual data is presented in the second part of the next chapter.

CHAPTER IV

DESIGN

Part A: Design of the Simulation

The purpose of this part of the study is to answer three of the research questions posed in Chapter I. The questions that this part of the study will be directed to are as follows:

1. Do the C-V and R techniques adequately estimate the true values of the item parameters?
2. Does one technique estimate the true values better than the other?
3. Do the C-V and R techniques estimate some true values of the item parameters better than others?

One approach to answering these questions would be to generate hypothetical data with various item values. In other words, one approach would be to design and implement a simulation.

Recall from the previous chapter that the theoretical model is represented by $\underline{P} = Q \boxtimes Q' \underline{\pi}$, where

$$\underline{P} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

are the observed proportions of individuals corresponding to the true proportions,

$$\underline{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} ,$$

(see Tables 4.1 and 4.2 below), $\boxtimes$ symbolizes the Kronecker product, $Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}$ and $Q' = \begin{pmatrix} q_{11}' & q_{12}' \\ q_{21}' & q_{22}' \end{pmatrix}$. The q_{ij}'s and q_{ij}''s represent probabilities and are defined according to Tables 4.3 and 4.4.

Table 4.1
Categories for a Given Item--Observed Proportions

		Posttest	
		Fail	Pass
Pretest	Fail	p_1	p_2
	Pass	p_3	p_4

Table 4.2
Categories for a Given Item--True Proportions

		Posttest	
		Does Not Know	Knows
Pretest	Does not know	π_1	π_2
	Knows	π_3	π_4

Table 4.3
Pretest--Actual

		Does Not Know	Knows
OBSERVED	Fail	q_{11}	q_{12}
	Pass	q_{21}	q_{22}

Table 4.4
Posttest--Actual

		Does Not Know	Knows
OBSERVED	Fail	q_{11}'	q_{12}'
	Pass	q_{21}'	q_{22}'

When the model is expanded, $\underline{P}$ can be represented by the following:

$$\underline{P} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11}q_{11}'\pi_1 + q_{11}q_{12}'\pi_2 + q_{12}q_{11}'\pi_3 + q_{12}q_{12}'\pi_4 \\ q_{11}q_{21}'\pi_1 + q_{11}q_{22}'\pi_2 + q_{12}q_{21}'\pi_3 + q_{12}q_{22}'\pi_4 \\ q_{21}q_{11}'\pi_1 + q_{21}q_{12}'\pi_2 + q_{22}q_{11}'\pi_3 + q_{22}q_{12}'\pi_4 \\ q_{21}q_{21}'\pi_1 + q_{21}q_{22}'\pi_2 + q_{22}q_{21}'\pi_3 + q_{22}q_{22}'\pi_4 \end{pmatrix} .$$

The R procedure defines the sensitivity index to be $\frac{\pi_2}{\pi_1 + \pi_2}$, but for computation uses the sample proportions. Therefore, R is computed by calculating $\frac{p_2 - p_3}{p_1 + p_2}$ where p_k are the sample proportions. In addition the C-V index is defined as $\pi_2 - \pi_3$, but is again computed using sample proportions and is $p_2 - p_3$.

If numerical values of π_k, q_{ij} and q_{ij}' are chosen, then the expected observed p_k can be computed. Random numbers can be generated

and then based on the values of p_k the number of cases in categories 1, 2, 3, and 4 can be determined. (Categories 1, 2, 3, and 4 follow the same pattern as the notation for the π_k and p_k.)

For example, suppose $p_1 = .1125$, $p_2 = .5075$, $p_3 = .0375$ and $p_4 = .3425$. Suppose also that a random number is generated. This random number is from a uniform distribution and is between 0.0 and 1.0. If it is less than .1125, then the number of cases in category 1 would increase by 1. If the random number is less than .6200 (.1125 + .5075) but greater than or equal to .1125, then the number of cases in category 2 would increase by 1. If the number is less than .6575 (.6200 + .0375) but greater than or equal to .6200, then the number of cases in category 3 would increase by 1. And finally, if the number is less than 1.00 but greater than or equal to .6575, then the number of cases in category 4 would increase by 1. Any random number generated would be counted in one and only one category. In this manner, simulated frequencies for the fail-fail group (category 1), fail-pass group (category 2), pass-fail group (category 3), and pass-pass group (category 4) are obtained.

For this simulation sample sizes of 50 and 200 will be considered. The sample size of 50 was selected because in most actual situations, 50 is the maximum number of individuals available. Some parameter values will be repeated in the simulation with a sample size of 200 in order to consider the stability of the indices.

For each set of parameter values 1000 samples will be generated. For each sample, the R and C-V indices will be computed. Of course, the true values remain the same for all 1000 cases. A number

of descriptive statistics will be computed based on the 1000 samples. These will include the means and the variances for the R and C-V indices and the largest and smallest values for each. In addition, skewness and kurtosis will be computed for each. The simulation is designed to consider a range of parameter values in order to see how close the estimate of the R and C-V indices are to the actual values.

Consider Tables 4.2, 4.3 and 4.4. The probability that an individual knows the answer yet fails to answer the item correctly, q_{12} or q_{12}', is probably quite small. Since $q_{12} + q_{22} = 1$ and $q_{12}' + q_{22}' = 1$, this assumption would imply that q_{22} or q_{22}' is large. In addition, the probability that an individual can guess the right answer (q_{21} or q_{21}') can be estimated by the number of options offered in the item. For example, a good estimate of q_{21} for a true-false item would be .50. For a multiple-choice item with four options a good estimate would be .25. The probability (q_{21}') that the correct answer could be guessed given some instruction may stay the same as q_{21} or it may decrease or increase. All possibilities were considered in the selection of the values of q_{21}'.

Table 4.5 lists the 21 different sets of parameter values that were selected for the simulation. Sixteen sets designate the probability of guessing (q_{21}) to be .25 (four--option multiple-choice item). Eight of these retain this estimate for the posttest ($q_{21}' = .25$). Seven of these sets increase the probability of guessing for the posttest to .50 ($q_{21}' = .50$). This makes the logical assumption that instruction may improve the individual's chances of guessing the correct answer by eliminating two of the possible options. For

Table 4.5
Selected Parameter Values for the Simulation

Parameter Set Number	N	q_{11}	q_{21}	q_{12}	q_{22}	q_{11}'	q_{21}'	q_{12}'	q_{22}'	π_1	π_2	π_3	π_4
1	50	.75	.25	.10	.90	.50	.50	0.0	1.0	.3	.5	.0	.2
2	50	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.3	.5	.0	.2
3	50	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	.3	.5	.0	.2
4	50	.50	.50	0.0	1.0	.50	.50	0.0	1.0	.3	.5	.0	.2
5	50	.75	.25	.10	.90	.50	.50	0.0	1.0	.2	.5	.1	.2
6	50	.75	.25	.10	.90	.50	.50	0.0	1.0	.8	.1	.05	.05
7	50	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.8	.1	.0	.1
8	50	.75	.25	.10	.90	.50	.50	0.0	1.0	.1	.8	.05	.05
9	50	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.1	.8	.0	.1
10	50	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	.2	.5	.1	.2
11	50	.75	.25	.10	.90	.50	.50	.10	.90	.2	.5	.1	.2
12	50	.75	.25	.10	.90	1.0	0.0	0.0	1.0	.2	.5	.1	.2
13	50	.75	.25	.10	.90	.75	.25	.10	.90	.2	.5	.1	.2
14	50	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.2	.5	.1	.2
15	200	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.3	.5	.0	.2
16	200	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	.3	.5	.0	.2
17	200	.75	.25	.10	.90	.50	.50	0.0	1.0	.2	.5	.1	.2
18	200	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	.2	.5	.1	.2
19	200	.75	.25	.10	.90	.50	.50	.10	.90	.2	.5	.1	.2
20	200	.75	.25	.10	.90	.75	.25	.10	.90	.2	.5	.1	.2
21	200	.75	.25	0.0	1.0	.75	.25	0.0	1.0	.2	.5	.1	.2

one set, the value of q_{21}' is set equal to zero, implying that after instruction the individual has no chance of guessing the correct response.

One set designates the probability of guessing (q_{21}) to be .50 (true-false item). This estimate is retained for the posttest ($q_{21}' = .50$). The remaining four sets satisfy the assumptions of the C-V index. These assumptions include assuming the probability of guessing is 0.0 for pretest and posttest ($q_{21} = q_{21}' = 0.0$) and assuming the probability of getting the item right when knowing the answer is 1.0 for pretest and posttest ($q_{22} = q_{22}' = 1.0$).

Based on the assumption that the probability that an individual who knows the answer yet fails to answer the item correctly (q_{12} or q_{12}') is quite small, q_{12} was designated to be 0.0 11 times and .10 the remaining ten times. The value for q_{12}' (posttest probability) was set at 0.0 for all but four parameter sets. For these, the value of q_{12}' remained equal to q_{12} which had been set at .10.

The values of π_1, π_2, π_3, π_4 were selected to represent reasonable situations. Two basic sets of values were chosen with π_1, π_2, π_3, π_4 equal to .3, .5, 0.0, .2, and .2, .5, .1, .2 respectively. Four sets of values were selected to consider the impact of extreme values on the indices. These sets (6, 7, 8, 9 of Table 4.5) considered the possibility that the majority of individuals would fail the pretest and pass the posttest (8 and 9).

As previously stated, for each set of parameter values 1000 samples will be generated. For these 1000 samples, the C-V and R

indices will be computed. Means, variances, highest and lowest values, skewness and kurtosis will also be determined for the C-V and R values.

In an attempt to answer the research questions, the data will be considered in a number of ways. All descriptive statistics--menas, variances, skewness and kurtosis--will be considered for each of the 21 parameter sets for both indices. Means of each of the C-V and R values will be compared to their true values and variances of these indices will also be considered in an attempt to answer the question of adequacy. For any given parameter set a mean value close to the true value in conjunction with low variance would imply some degree of adequacy.

The second question, "Does one technique estimate the true values better than the other?", will also be answered by consideration of the data. One approach will be to consider how close the values are to the true value for each set of parameters for each technique. The variance, skewness and kurtosis values will also be considered. A comparison of the correlation coefficients between the true values and the means for each technique might show whether or not one technique estimates the true values better than the other technique. However, some caution will be used in the interpretation of the correlation coefficients and the comparison.

The final question will also be handled descriptively. The actual data will be considered and an attempt will be made to locate values that are not estimated as well as others.

All questions will be considered descriptively. Each parameter set with results will be discussed with respect to the three basic research questions. Summary statistics will be presented in order to facilitate the understanding of the techniques and the conclusions reached about them.

Part B: Design of the Comparison Study With Actual Data

The purpose of this part of the study is to determine the comparability of three item analysis procedures (C-V, R and B-S). Data were obtained from the Michigan Middle Cities Project. One hundred twenty-eight items were chosen from two subject areas, Reading and Mathematics. Two levels were considered--Middle and Upper. (These levels generally refer to grades three and four, and five and six respectively.)

Each item was written for a particular objective. Each objective was tested by four items on a pretest, four different items on a posttest and all eight items on a retention test. The retention test was given approximately 40 days after the posttest.

There were also two treatment groups where item data were collected. In one treatment, teachers were assigned objectives (treatment A). In the other treatment, teachers were allowed to choose objectives (treatment B). Sixteen objectives were chosen to complete the design which is represented in Diagram 4.1.

The major question to be considered is "Do the C-V, R and B-S item analysis procedures provide comparable results?" The analysis of this question will primarily be descriptive. To determine the

Subject	Level	Treatment	Objective Number	N	Items
Reading	Middle	A	142	31	1-8
			116	59	9-16
		B	112	21	17-24
			120	20	25-32
	Upper	A	145	66	33-40
			199	57	41-48
		B	182	30	49-56
			166	18	57-64
Mathematics	Middle	A	108	52	65-72
			111	43	73-80
		B	107	42	81-88
			109	37	89-96
	Upper	A	198	22	97-104
			176	46	105-112
		B	187	16	113-120
			167	17	121-128

Diagram 4.1
Design of Administration of Items

comparability, of the indices C-V and R, a Pearson product moment correlation will be computed between the C-V and R values.

The B-S procedure (see Appendix II) does not allow for a single index. The B-S procedure involves the computation of four error rates. TER (theoretical error rate) is defined as (J-1)/J where J is the number of possible answers to an item--or it is simply the expected proportion of students answering a pretest item incorrectly. The Base Error Rate (BER) is the observed proportion of students answering a pretest item incorrectly. The Posttest Error Rate (PER) is the observed proportion of students answering a posttest item incorrectly. In this situation the data used as the posttest data will be from the retention test. The Instructional Error Rate (IER) is the proportion of students answering incorrectly on a terminal test item which is administered to students who have been exposed to instruction. This last error rate is not included in any of the decision rules related to item revision.

In addition two discrimination indices are computed, the Base Discrimination Index (BDI) and the Posttest Discrimination Index (PDI). These are computed using the total score on the appropriate test as the criterion. For BDI, the criterion will be the pretest and for PDI, the criterion will be the posttest. Again in this situation the data used will be from the retention test. Two separate statistics will be used to compute the discrimination indices, the phi-coefficient and the B index. The B index equals $B/(B + D) - A/(A + C)$ where A, B, C and D are defined in Table 4.6.

Table 4.6
B--Index

		Total Test Score	
		Nonmastery	Mastery
	1	A	B
Item Score			
	0	C	D

Mastery for the items on the pretest and retention test data was set at three out of four items.

These five pieces of information, TER, BER, PER, BDI and PDI are then used in conjunction with some rules to determine the adequacy of the item. Appendix II provides a description of these rules.

Since the B-S procedure does include several statistics, the comparison of the three indices will be done in the following manner. First the individual statistics, TER, BER, PER, BDI, and PDI, which are necessary for the B-S procedure, will be computed. The appropriate rules will be applied and a decision will be made about the quality of the item; that is, does the item need to be revised? Each item can then be assigned a "0" or a "1" depending on the outcome of the application of the rule. A "0" would indicate non-acceptance or revision required; a "1" would indicate acceptance or no revision required.

There is a limitation in this procedure. The B-S process requires that the evaluator set various cut-off points. For example, the evaluator must decide an appropriate cut-off point between a high

and low error rate. Comparison of B-S with the R and C-V indices will be influenced by the selected cut-off values. To minimize the effect of this limitation various cut-off values will be set and several comparisons will be made. Point-biserial correlations will be computed between the B-S value of "0" or "1" and the C-V value and the R value.

There is also a limitation with the data used in the computation of the various indices. Retention data are substituted for the posttest data generally used in the computation of C-V and R indices. There may be some additional forgetting not normally found in a more immediately given posttest. However, since both R and C-V are computed using the same data, the effect on the comparison of the two should be minimal. The observed frequency of f_3 (those who forget from pretest to retention test) might appear to influence C-V more than R since this frequency is included in the denominator of C-V as well as the numerator $\left(\frac{f_2 - f_3}{f_1 + f_2 + f_3 + f_4}\right)$. R only includes f_3 in the numerator $\left(\frac{f_2 - f_3}{f_1 + f_2}\right)$. However, since the only difference in the two indices is the addition of f_3 and f_4 in the denominator of C-V, and if f_3 gets larger due to the longer time frame, then f_1, f_2 and/or f_4 would get smaller. Since $f_1 + f_2$ are the same in both and the total $f_1 + f_2 + f_3 + f_4$ is N, a constant, then the effect on both indices should make little difference in the comparison of the two. This same argument holds for the impact of a decrease or increase in f_4 on the comparison of the two indices. A similar argument can be made for the individual statistics of the B-S procedure.

Two additional questions that are of interest are as follows:

Are the three procedures more comparable for items in Mathematics than for items in Reading?; and,

Does the comparability of the three procedures depend on the grade level?

It is anticipated that the procedures will be more comparable for Mathematics than for Reading. Items for Mathematics are constructed more easily than Reading items because the subject area is more structured. In addition the items are generally of higher quality. The Reading items may be more ambiguous than the Mathematics items. It is also anticipated that the correlations among the indices would be almost identical for items given in the Upper grades and for items given in the Middle grades. There is no reason to expect the correlations to depend on grade level.

Each of the two questions will be analyzed in two steps. A comparison of the C-V and R values will be considered separately then a comparison of the B-S with the C-V and R will be made.

The first question can be analyzed in the following manner. First, a Pearson product moment correlation will be computed between the C-V and R values for items in Mathematics and separately for items in Reading. Then a comparison can be made between these two correlation coefficients. The null hypothesis can be expressed as H_o: $\rho_R = \rho_M$ with the alternative hypothesis being H_1: $\rho_R \neq \rho_M$, where ρ_M = the population correlation of C-V and R for Mathematics and ρ_R = the population correlation of C-V and R for Reading. A

Fisher's z- transformation will be made for each of the sample correlations and a z-test will be applied.

Secondly, the point-biserial correlation will be computed between the B-S and C-V values for Reading and Mathematics and between the B-S and R values for Reading and Mathematics.

The second question will be considered in the same manner only correlations will be computed between the various indices for the two grade levels separately.

A final question which may be considered is "Are the three procedures more comparable for items given in treatment A than for items given in treatment B?" The analysis of this question is similar to the analyses proposed above. The prediction for the correlations among the indices is that they will be higher for treatment B than for treatment A. This is due primarily to the fact that teachers were assigned objectives in treatment A. Instruction may not have been needed for these specific objectives or may not have been given adequately, so the item response data for treatment A may be unstable. Items from treatment B should more closely fit the ideal criterion-referenced situation, i.e., no knowledge on pretest and knowledge on posttest.

Summary

There are two parts to this study. Each part is designed to answer different questions. The first part, the simulation, will attempt to determine if the C-V and R indices adequately estimate the true values, if one technique estimates the true values better

than the other and if for some parameter sets the C-V and R indices are better estimators of the true values. Data will be analyzed descriptively for the 21 sets of parameter values used in the simulation. Additional questions, such as the stability of estimates for different sample sizes, will also be considered in the analysis of the data.

The second part of the study is designed to determine comparability of the three item analysis procedures, R, C-V, and B-S. C-V and R values will be computed for 128 items and the B-S values will be computed on 64 of the 128 items. The relationships among the indices will be determined using correlation coefficients. Additional questions pertaining to the comparability of the indices with respect to subject matter, grade and treatment also will be considered.

CHAPTER V

RESULTS OF THE SIMULATION

The simulation was designed to answer three questions:

1. Do the C-V and R techniques adequately estimate the true values of the item parameters?
2. Does one technique estimate the true values better than the other?
3. Do the C-V and R techniques estimate some true values of the item parameters better than others?

The results of the simulation for the 21 sets of parameter values (see Table 4.5) are presented in Table 5.1.

The C-V Index: Adequacy and Stability

Assumptions Met

Consider the statistics of parameter sets 3, 10, 16 and 18 (see Table 5.2). For these parameter sets, the assumptions for the C-V index are met. Recall that the assumptions for C-V include no guessing ($q_{21} = q_{21}' = 0.0$) and an individual who knows the answer will not fail to answer correctly ($q_{22} = q_{22}' = 1.0$). (The parameter sets 16 and 18 are identical to 3 and 10 respectively except N = 200.)

Table 5.1
Descriptive Statistics for Each Parameter Set

Parameter Set	True R	True C-V	Mean R	Mean C-V	Absolute Deviation Squared R (Bias)	Var R	Absolute Deviation Squared C-V (Bias)	Var C-V	SD R	SD C-V	Kurtosis R	Kurtosis C-V	Skewness R	Skewness C-V	Range R	Range C-V
1.	.6250	.5000	.7553	.4722	.016978	.0074	.000773	.0070	.0860	.0837	.4655	-.0672	-.4433	-.1495	.3478 to 1.0	.16 to .70
2.	.6250	.5000	.6209	.3760	.000017	.0112	.015376	.0076	.1059	.0869	.3512	-.1108	-.4369	-.1469	.1852 to .8929	.1 to .62
3.	.6250	.5000	.6231	.4995	.000004	.0059	.000000	.0049	.0765	.0697	-.3022	-.1825	.1022	.0938	.4103 to .85	.32 to .72
4.	.6250	.5000	.6093	.2487	.000246	.0221	.063152	.0070	.1488	.0835	1.7626	-.0070	-.8651	-.1138	-.0769 to 1.0	-.02 to .5
5.	.7143	.4000	.7252	.4072	.000119	.0101	.000052	.0076	.1004	.0873	.2062	-.1417	-.4417	-.0480	.3684 to .9667	.14 to .68
6.	.1111	.0500	.3730	.2592	.068592	.0133	.043765	.0083	.1152	.0909	.1349	-.1096	-.4259	-.1217	-.0690 to .6486	-.04 to .52
7.	.1111	.1000	.1061	.0762	.000025	.0147	.000566	.0072	.1213	.0850	-.0186	-.1160	-.2974	.0197	-.3 to .4390	-.18 to .36
8.	.8889	.7500	.8911	.6102	.000005	.0032	.019544	.0059	.0565	.0771	.2398	-.0950	-.5329	-.2082	.6765 to 1.0	.36 to .84
9.	.8889	.8000	.8876	.6005	.000002	.0033	.000018	.0061	.0576	.0781	.6876	.1005	-.5607	-.1962	.6 to 1.0	.34 to .82
10.	.7143	.4000	.5700	.4042	.020822	.0108	.000018	.0088	.1039	.0936	.3832	.0377	-.4005	-.0631	.08 to .8684	.04 to .7
11.	.7143	.4000	.6012	.3398	.012792	.0160	.003624	.0089	.1264	.0944	1.2161	-.0101	-.7574	-.2693	.0385 to .9	.02 to .56
12.	.7143	.4000	.4481	.2535	.070862	.0190	.021462	.0088	.1377	.0941	.0600	-.2646	-.3595	.0335	-.1 to .7586	-.04 to .52
13.	.7143	.4000	.4596	.2614	.064872	.0194	.019210	.0094	.1392	.0968	.5078	-.0691	-.4648	.0064	-.2 to .8611	-.06 to .62
14.	.7143	.4000	.5654	.3013	.022171	.0171	.009742	.0086	.1309	.0929	.1224	-.1925	-.4544	.0002	.0667 to .871	.02 to .56
15.	.6250	.5000	.6247	.3763	.000000	.0024	.015302	.0016	.0485	.0404	-.0189	-.2957	-.1462	-.0300	.4679 to .776	.255 to .505
16.	.6250	.5000	.6265	.5005	.000001	.0015	.000000	.0013	.0386	.0367	-.4356	-.2182	.0242	-.0090	.525 to .7365	.385 to .615
17.	.7143	.4000	.7282	.4047	.000193	.0026	.000022	.0019	.0507	.0440	.2324	-.1345	-.2784	.0009	.5347 to .8783	.27 to .54
18.	.7143	.4000	.5728	.4018	.020022	.0027	.000003	.0022	.0522	.0469	.0584	.0071	-.3137	-.1255	.3571 to .7133	.225 to .56
19.	.7143	.4000	.5999	.3332	.013087	.0036	.004462	.0022	.0598	.0471	.2745	.0125	-.3962	-.1211	.3053 to .7481	.145 to .49
20.	.7143	.4000	.4657	.2593	.061801	.0047	.019996	.0023	.0685	.0481	-.0871	-.1115	-.2247	-.0187	.2472 to .6829	.11 to .42
21.	.7143	.4000	.5689	.2998	.021141	.0040	.010040	.0021	.0630	.0458	-.1075	-.0877	-.1492	.0056	.3535 to .7456	.175 to .46

Table 5.2
Parameter Sets Where Assumptions for C-V Are Met

	True C-V	Mean C-V	Absolute Deviation	Var C-V	Kurtosis	Skewness	Range of C-V
3.	.5000	.4995	.0005	.0049	-.1825	.0938	.3200 to .7200
10.	.4000	.4042	.0042	.0088	.0377	-.0631	.0400 to .7000
16.	.5000	.5005	.0005	.0013	-.2182	-.0090	.3850 to .6150
18.	.4000	.4018	.0018	.0022	.0071	-.1255	.2250 to .5600

Comparison--Assumptions Met Versus Assumptions Not Met

The average absolute deviation from the true C-V value for the mean C-V for the sets where the assumptions are met is .0018. In comparison for the remaining parameter sets, where the assumptions for the C-V index are not met, the average absolute deviation from the true C-V for the mean C-V is .1096, a considerable difference. The variances for the parameter sets where the assumptions are met range from .0013 to .0088, but the variances for the sets where the assumptions are not met, range from .0016 to .0094. Ten of these 17 have variances equal to or greater than .0070. The variances are lowest for those parameter sets (15 through 21) which have sample sizes of 200 (.0013 to .0023).

The kurtosis values for the distributions of the C-V index range from -.2957 to .1005. Only four values are positive; two of these are for those parameter sets where the C-V assumptions are met. Since the kurtosis values are not very large or very far from zero, it seems reasonable to describe the distributions as mesokurtic.

The skewness values range from -.2693 to .0938. Fourteen values are negative. The skewness values are also not very large or very far from zero, so the skewness for any parameter set is slight. If the skewness and kurtosis values are considered together, then the distributions for all the parameter sets can probably be described as normal.

A comparison of the averages of the absolute deviations from the kurtosis value of zero for parameter sets with N = 50 and for

parameter sets with N = 200 reveals that the latter average is slightly larger (.11 for N = 50; .12 for N = 200). A similar comparison of the averages of the absolute deviations from the skewness value of zero for parameter sets with N = 50 and for parameter sets with N = 200 reveals that the latter are less skewed (.11 for N = 50; .04 for N = 200). The values, however, for N = 50 and N = 200 do not differ enough for one to infer that the greater sample size provides a more normal distribution.

Comparison of the average ranges for parameter sets with N = 50 and N = 200, .54 for N = 50 and .29 for N = 200, does demonstrate that the C-V estimates are more stable with larger sample sizes. For those parameter sets that do not meet the C-V assumptions the average range is .47, N = 17. For those parameter sets that do meet the assumptions, the average range is .41, N = 4. There appears to be slight differences in the averages when sample sizes are also considered. See Table 5.3 below.

Table 5.3
Average Ranges for the C-V Estimates

	All	Sample Size = 50	Sample Size = 200
Meet Assumptions	.41 (N=4)	.53 (N=2)	.28 (N=2)
Does Not Meet Assumptions	.47 (N=17)	.54 (N=12)	.29 (N=5)
All	.46 (N=21)	.54 (N=14)	.29 (N=7)

Final evidence of the adequacy is the correlation between the true C-V value and the mean C-V value. This correlation is .800 (N=21, $p < .001$). From the evaluation of the other statistics: ranges, kurtosis and skewness values and variances, in addition to the correlation cited above, one can infer that the C-V technique provides reasonable estimates of the true C-V value and these estimates are distributed normally. However, the technique does provide a more stable estimate for larger sample sizes (N=200).

The R Index: Adequacy and Stability

Assumptions Met

Consider the statistics of parameter sets 2, 3, 4, 7, 9, 15 and 16 (see Table 5.4). For these parameter sets, the assumptions for the R index are met. Recall that the assumptions for the R index are that guessing is the same for the pretest and the posttest ($q_{21} = q_{21}'$), an individual who knows the answer will not fail to answer correctly ($q_{22} = q_{22}' = 1.0$), and an individual who knows the answer on the pretest does not forget it on the posttest ($\pi_3 = 0$). (The parameter sets 15 and 16 are identical to 2 and 3 respectively except N = 200.)

Comparison--Assumptions Met Versus Assumptions Not Met

The average absolute deviation from the true R for the mean R for the sets where the assumptions are met is .0043. In comparison, for the remaining parameter sets where the assumptions of the R index are not met, the average absolute deviation from the true R for the

Table 5.4
Parameter Sets Where Assumptions for R Are Met

	True R	Mean R	Absolute Deviation	Var R	Kurtosis	Skewness	Range of R
2.	.6250	.6209	.0041	.0112	.3512	-.4369	.1852 to .8929
3.	.6250	.6231	.0019	.0059	-.3022	.1022	.4103 to .8500
4.	.6250	.6093	.0157	.0221	1.7626	-.8651	-.0769 to 1.0000
7.	.1111	.1061	.0050	.0147	-.0186	-.2974	-.3000 to .4390
9.	.8889	.8876	.0013	.0033	.6876	-.5607	.6000 to 1.0000
15.	.6250	.6247	.0003	.0024	-.0189	-.1462	.4679 to .7760
16.	.6250	.6265	.0015	.0015	-.4356	.0242	.5250 to .7365

mean R is .1426. The variances for the R values where the assumptions are met range from .0015 to .0221 but the variances for the remaining R values only range from .0026 to .0194. The reverse might have been expected. It would seem more likely for the R values to be more stable when the assumptions of the index are met. This does not seem to be the case; although, the differences in the ranges of the variances are slight.

Other evidence of the stability or lack of stability of the estimates of the R index can be obtained by consideration of other distributional statistics; such as skewness, kurtosis and range.

There are 15 total parameter sets where the kurtosis is positive. A positive value implies that the curve is leptokurtic (peaked). Two of the positive kurtosis values are near zero (parameter set #12, K = .0600 and parameter set #18, K = .0584). For these two parameter sets the distributions can probably be described as mesokurtic. The remaining six parameter sets have negative kurtosis values; three of these are near zero (parameter set #7, K = -.0186; parameter set #15, K = -.0189; parameter set #10, K = -.0871). A negative kurtosis value generally indicates that the curve is platykurtic (flat); however, the three curves whose values are near zero could be considered mesokurtic. The largest value is 1.7626 for parameter set #4. This is one set where the assumptions of the R index are met. Ideally, the 1000 R values should be concentrated in a narrow range about the true value.

There are 19 total parameter sets where the distribution is negatively skewed. Only two parameter sets have positively skewed

distributions. Parameter set #16 has a skewness value close to zero (Sk = .0242) which might indicate a non-skewed distribution. When the kurtosis value is also considered (K = -.4356), it appears that the distribution is slightly flat. However, the kurtosis value is not very large so the interpretation of the two statistics could be that the distribution of R values for parameter set #16 is fairly normal. Parameter set #16 has a sample size of 200. Parameter set #15, also with a sample size of 200, has a small negative kurtosis value and a small negative skewness value. Again one might infer that the distribution is fairly normal. Perhaps, for parameter sets that meet the assumptions of the R index and have sample sizes of 200, the R values are distributed more normally.

If the other five parameter sets with N = 200 (17, 18, 19, 20 and 21) are also considered, the skewness values are greater than the values for parameter sets #15 and #16. However, the skewness value for parameter set #21 (Sk = -.1492) is not very different than the value for #15 (Sk = -.1462). Also the kurtosis value is fairly small (K = -.1075). The assumptions for the R index are almost met in #21 except π_3 does not equal zero. The kurtosis values for these five sets, are all small with three positive values and two slightly negative. It is interesting to note that the highest kurtosis value (absolute value) of the seven sets with N = 200, is set #16. A comparison of #15 and #16 with the remaining five parameter sets with N = 200 seems to show that if the assumptions of R are met (or nearly met) the distribution is more nearly normal.

If the absolute deviations from the kurtosis value of zero are averaged for the parameter sets with N = 50 and for the parameter sets with N = 200 and these two values (.46 and .17, respectively) are compared, then further evidence is obtained that the distributions of R values for larger sample sizes are more nearly normal.

A similar consideration of the absolute deviations from the skewness value of zero reveals that the average for parameter sets with N = 50 (.47) is larger than the average for parameter sets with N = 200 (.22). It seems then that for any given parameter set as the sample size increases the distribution of R values approaches a normal distribution.

Now consider the ranges of the R values for the 21 parameter sets. For parameter sets with sample sizes of 50, N = 14, the average range is .72. For parameter sets with sample sizes of 200, N = 7, the average range is .36. The ranges, then, were decreased on the average by one-half when the sample sizes were increased. For those parameter sets that do not meet the assumptions and with sample sizes of 50, N = 9, the average range is .74. For parameter sets with sample sizes of 200, N = 5, the average range is .39. For those parameter sets that do meet the assumptions, the average range is .67 for sample sizes of 50, N = 5, and .26 for sample sizes of 200, N = 2. There is some reduction in the ranges when the assumptions are met; however, just the increase in sample size without meeting the assumptions has a marked effect on the stability of R.

Final evidence which might be considered in answering the question of adequacy is the correlation of the true R values with

Table 5.5
Average Ranges for the R Estimates

	All	Sample Size = 50	Sample Size = 200
Meet Assumptions	.55 (N=7)	.67 (N=5)	.26 (N=2)
Does Not Meet Assumptions	.62 (N=14)	.74 (N=9)	.39 (N=5)
All	.60 (N=21)	.72 (N=14)	.36 (N=7)

the mean of the estimated R values for each parameter set. This correlation is .759 ($N = 21$, $p < .001$). Even though this correlation is significant, it must be remembered that for any given parameter set there were many R values which greatly differed from the true R. Consideration of ranges, variances, skewness and kurtosis values reveals that the R technique more adequately estimates the true R when the sample size is larger, i.e. N = 200. In addition the R technique more adequately estimates the true R when the assumptions of R are met. The differences in the adequacy are far more dramatic, however, when the sample size is increased than when the assumptions are met.

The C-V Technique Versus the R Technique

When the distributions of the estimates of the C-V and R values are compared it appears that, in general, the C-V estimates are distributed more normally than the R estimates. The R values tend to be higher than the C-V values and there seems to be a ceiling

effect, i.e. the R distributions are generally skewed negatively and in almost every case approach the upper bound (1.0). The C-V distributions while skewed negatively in 14 cases seem to span a middle range of values.

Consider the summary statistics for the C-V and R distributions provided in Table 5.6 and the correlation matrix in Table 5.7.

Table 5.6
Summary Statistics Comparing R to C-V

	Average Absolute Deviation From True Value	Range of Variances	Range of Kurtosis	Range of Skewness	Average Range of Values
C-V	.0891	.0013 to .0094	-.2957 to .1005	-.2693 to .0938	.46
R	.0965	.0015 to .0221	-.4356 to 1.7626	-.8651 to .1022	.60

Table 5.7
Correlations

	True C-V	Mean R	Mean C-V
True R	.820	.759	.608
True C-V		.855	.800
Mean R			.891

One can infer from these statistics that the C-V technique estimates the true C-V values better than the R technique estimates the true R values. The variances of the distributions of the C-V

values are smaller. The largest variance for the C-V values is .0094 while the largest variance for the R values is .0221. The range of the kurtosis values and the range of the skewness values for the C-V index are considerably smaller than the ranges for the R index. The average range of values for the parameter sets is smaller for the C-V index than for the R index. Finally, if the correlations of the mean index with the true value are considered, the C-V technique provides a closer estimate of the true C-V value ($r = .80$ for C-V compared to $r = .76$ for R). It is interesting to note that the means of the estimates of the R index are more closely related to the true C-V value (.86) than they are to the true R values (.76) or than the means of the estimates of the C-V index are related to the true C-V values (.80).

In interpretation of the correlations, it must be remembered that the means of the estimated values for a given parameter set (over 1000 values) are correlated with the true values. Means are more stable than the actual estimates. The other statistics, range, kurtosis, variance, and skewness, must be considered in the evaluation of the adequacy of the techniques. When all statistics are considered, the C-V technique seems to provide a more stable estimate of the true value than the R technique and the distributions of the C-V values seem to be more normally-shaped than the R values for any given parameter set.

Consideration of C-V and R Techniques By Parameter Set

The conclusion from the analyses cited in the previous sections is that the C-V technique provides a more stable estimate of the true value than does the R technique. Now consider the parameter sets individually. Perhaps one technique is a better estimator than the other technique under certain conditions. If so, what are these certain conditions?

Consider, first, the parameter sets where the assumptions for the index are met. Table 5.8 gives the summary statistics for R and C-V.

It is apparent from these data that, over 1000 samples, the mean estimate for either index, is better when the assumptions are met than when they are not. (See column one of Table 5.8.) One perplexing fact is that the variances for those parameter sets where the R assumptions are met, span a larger range than do those parameter sets where the R assumptions are not met. However, if the size of the samples is also considered and only those parameter sets with N = 200 are compared, then the variances are less when the R assumptions are met. Interestingly, the same unexpected result occurs if the variances of the C-V estimates are considered. Here, for sample sizes of 50, the range of the variances is slightly greater when the assumptions are met than when they are not met. This is not true, however, for sample sizes of 200. Caution must be used in interpreting these results, since the number of parameter sets used is quite small (see column six of Table 5.8).

Table 5.8
Summary Statistics for R and C-V With Consideration of Sample Size and Assumptions

	Average Absolute Deviation From True Value	Range of Variances	Range of Kurtosis	Range of Skewness	Average Range of Values	Number of Parameter Sets
			C-V			
Assumptions Met						
N = 50	.0024	.0049 to .0088	-.1825 to .0377	-.0631 to .0938	.53	2
N = 200	.0012	.0013 to .0022	-.2182 to .0071	-.1255 to -.0090	.28	2
All	.0018	.0013 to .0088	-.2182 to .0377	-.1255 to .0938	.41	4
Assumptions Not Met						
N = 50	.1189	.0059 to .0094	-.2646 to .1005	-.2693 to .0335	.54	12
N = 200	.0872	.0016 to .0023	-.2957 to .0125	-.1211 to .0056	.29	5
All	.1096	.0016 to .0094	-.2957 to .1005	-.2693 to .0335	.47	17
All						
N = 50	.1022	.0049 to .0094	-.2646 to .1005	-.2693 to .0938	.54	14
N = 200	.0626	.0013 to .0023	-.2957 to .0125	-.1255 to .0056	.29	7
All	.0890	.0013 to .0094	-.2957 to .1005	-.2693 to .0938	.46	21
			R			
Assumptions Met						
N = 50	.0056	.0033 to .0221	-.3022 to 1.7626	-.8651 to .1022	.67	5
N = 200	.0009	.0015 to .0024	-.4356 to -.0189	-.1462 to .0242	.26	2
All	.0043	.0015 to .0221	-.4356 to 1.7626	-.8651 to .1022	.55	7
Assumptions Not Met						
N = 50	.1481	.0032 to .0194	.0600 to 1.2161	-.7574 to -.3595	.74	9
N = 200	.1328	.0026 to .0047	-.1075 to .2745	-.3962 to -.1492	.39	5
All	.1426	.0026 to .0194	-.1075 to 1.2161	-.7574 to -.1492	.62	14
All						
N = 50	.0972	.0032 to .0221	-.3022 to 1.7626	-.8651 to .1022	.72	14
N = 200	.0951	.0015 to .0047	-.4356 to .2744	-.3962 to .0242	.36	7
All	.0965	.0015 to .0221	-.4356 to 1.7626	-.8651 to .1022	.60	21

The range of kurtosis values for R is greater when meeting the assumptions than when not. The opposite is true for the range of kurtosis values for C-V. Similarly, the range of skewness values for R is greater when the assumptions are met than when they are not and the range of the skewness values for C-V is smaller when the assumptions are met than when they are not. Finally, the average range of the respective values is smaller for both indices when the assumptions are met.

Sample size has a marked effect on the results of the simulation for any parameter set. Noted above was the effect that sample size had on the range of the variances. Also the mean of the estimated values is closer to the true value for both indices when the sample size is 200. However, the increase in sample size for both indices has a greater effect on the mean of the estimated values when the assumptions are met than when the assumptions are not met. Of course, the increase in sample size also decreases the range of estimated values for both indices. For R, this average range is reduced by 61 percent for the parameter sets meeting the assumptions, but only by 47 percent for those not meeting the assumptions. For C-V, the reduction in the average range is 47 percent and 46 percent respectively. The increase in sample size narrows the range of estimated values considerably.

The ranges of skewness and kurtosis values are much narrower for the parameter sets where N = 200 than for the parameter sets where N = 50 for the R index. For C-V, the ranges are closer, although generally smaller for N = 200. There is one exception; the

range of the kurtosis values when the C-V assumptions are met is greater for N = 200 than for N = 50. However, there were only two parameter sets included for these categories, so the statistics must be interpreted cautiously.

Two factors have been considered above; one, whether or not the assumptions of a particular index were met and two, sample size, i.e. what happened to the distributions of estimated values when the sample size was increased from 50 to 200. The analysis of the data with respect to these two factors seems to indicate that the C-V method provides a better estimate when the assumptions are met; although, the technique is still good under other assumptions. The R method seems to be unstable. The descriptive statistics indicate that the R method does not provide good estimates even under the best of circumstances. An increase in sample size helps the R method. The C-V technique, although a better technique with a larger sample size, remains stable with smaller sample sizes.

Now consider the individual parameter sets. Consider only the mean of the estimates, the variances, and the ranges for each index for each parameter set. Table 5.9 indicates for each parameter set whether the absolute deviation from the true value is smaller for R or C-V, the variance is smaller for R or C-V, and the range is smaller for R or C-V. For each column the letter R or C-V indicates that the statistic is smaller for that technique.

In 29 percent of the parameter sets the R technique estimates the true value better than the C-V technique estimates the true value. In less than 10 percent of the cases the variance for R is

Table 5.9
Comparison of R and C-V by Parameter Set[1]

Parameter Set #	Absolute Deviation From True Value	Variance	Range
1.	C-V	C-V	C-V
2.	R	C-V	C-V
3.	C-V	C-V	C-V
4.	R	C-V	C-V
5.	C-V	C-V	C-V
6.	C-V	C-V	C-V
7.	R	C-V	C-V
8.	R	R	R
9.	R	R	R
10.	C-V	C-V	C-V
11.	C-V	C-V	C-V
12.	C-V	C-V	C-V
13.	C-V	C-V	C-V
14.	C-V	C-V	C-V
15.	R	C-V	C-V
16.	C-V	C-V	R
17.	C-V	C-V	C-V
18.	C-V	C-V	C-V
19.	C-V	C-V	C-V
20.	C-V	C-V	C-V
21.	C-V	C-V	C-V

[1]For each column the letter R or C-V indicates that the statistic is smaller for that technique.

smaller, and in 14 percent of the cases the range is smaller.

Consider the two parameter sets where the R technique appears to be the better technique (#8 and #9). In these parameter sets, it

was assumed that 80 percent of the individuals would not know the answer at pretest but would know it at posttest ($\pi_2 = .80$). (See Table 4.5 in Chapter IV.) In parameter set #8, it was also assumed that instruction would improve the chance of guessing ($q_{21} = .25$, $q_{21}' = .50$), and that for the pretest there would be some chance that an individual knowing the answer would fail the item ($q_{12} = .10$). Parameter set #9 assumed only that there was the same chance of guessing for both pretest and posttest ($q_{21} = q_{21}' = .25$). This parameter set meets the R assumptions. It is interesting to note that for the parameter sets where an R occurs in any column the assumptions for the R index are met in six of these seven cases.

Other than the two factors, sample size and meeting the correct R assumptions, there seems to be no pattern for the estimates being better for one parameter set than for another. It does appear, however, that the more assumptions of the R technique that are not met, the less accurate the estimates. The C-V technique seems to provide reasonable estimates regardless of sample size or meeting assumptions.

Summary

Three questions were considered in the designing of the simulation. These were:

1. Do the C-V and R techniques adequately estimate the true values?
2. Does one technique estimate the true values better than the other?

3. Do the C-V and R techniques estimate some true values better than others?

The answers to these questions were discussed in this chapter. First, the adequacy of a technique (R or C-V) was determined by consideration of a number of descriptive statistics. It was found that for R, when the assumptions are met, the technique provides a more stable and accurate estimate. It was also found that when the sample size is increased from 50 to 200, the stability and accuracy increases greatly. A correlation coefficient of .759 between the true R value and the mean R value for 1000 estimates implies that the procedure provides a reasonable estimate of the true R.

The C-V technique seems to provide a reasonably accurate and stable estimate regardless of whether the assumptions are met. The estimates, however, are more stable with larger sample sizes, e.g. average range is .54 for N = 50 and .29 for N = 200. A correlation coefficient of .80 between the true C-V and the mean C-V value for 1000 estimates implies that the procedure provides a reasonable estimate of the true C-V.

Second, the C-V technique seems to estimate the C-V true value better than the R technique estimates the R true value. The average absolute deviation from the respective true values is smaller for C-V than for R (.0891 and .0965, respectively). In addition the range of variances is considerably smaller for the C-V estimates than for the R estimates (.0013 to .0094 for C-V and .0015 to .0221 for R) and the average range of estimated values is smaller (.46 for C-V and .60 for R).

The third question was primarily answered in considering the question of adequacy and stability. For both techniques the estimates are better when the sample size is larger (N = 50 versus N = 200). In addition, the R approach is better when the assumptions are met. This is not true for the C-V approach. The C-V approach seems to provide a good estimate under almost any assumptions.

The next chapter describes the results of the comparison of the R and C-V approaches using actual data on 128 items. In addition a third approach, the B-S method, is also used on 64 of the 128 items and the results compared to the R and C-V values.

CHAPTER VI

RESULTS OF THE COMPARISON OF THE THREE INDICES WITH ACTUAL DATA

The purpose of this part of the study was to determine the comparability of the three item analysis procedures, C-V, R and B-S. For this comparison, data were obtained from the Michigan Middle Cities Project. Sixteen objectives were chosen from two subject areas, Mathematics and Reading. In addition two levels, Middle and Upper, were considered in the selection of the objectives. These levels refer to grades three and four, and five and six, respectively. Each objective was tested by four items on a pretest, four different items on a posttest and all eight items on a retention test. The retention test was given approximately 40 days after the posttest. There were also two treatment groups considered in the selection of the objectives. In treatment A, teachers were assigned objectives. In treatment B, teachers selected the objectives. Diagram 6.1 shows the complete design.

The major question considered was "Do the C-V, R and B-S item analysis procedures provide comparable results?" Three other questions were also considered:

1. Are the three procedures more comparable for items in Mathematics than for items in Reading?;

Subject	Level	Treatment	Objective Number	N	Items
Reading	Middle	A	142	31	1-8
			116	59	9-16
		B	112	21	17-24
			120	20	25-32
	Upper	A	145	66	33-40
			199	57	41-48
		B	182	30	49-56
			166	18	57-64
Mathematics	Middle	A	108	52	65-72
			111	43	73-80
		B	107	42	81-88
			109	37	89-96
	Upper	A	198	22	97-104
			176	46	105-112
		B	187	16	113-120
			167	17	121-128

Diagram 6.1
Design of Administration of Items

2. Does the comparability of the three procedures depend on the grade level?; and,
3. Are the three procedures equally comparable for items given in treatment A as for items given in treatment B?

These last three questions are part of the major question and will be treated as such in the discussion of the results.

Comparability

C-V and R

Consider the testing procedure for each objective. Four items were given on the pretest, four different items were used on the posttest, and all eight items were included on the retention test. For computation of a C-V index or an R index it is necessary to have data on a given item at two times, preferably a pretest and a posttest. In this situation, it was necessary to compute the C-V and R indices from pretest-retention test data and from posttest-retention test data. The indices can be computed from pretest-posttest data on parallel items, but the usefulness and meaningfulness of these data for item selection and revision is questionable. There are 64 items using pretest-retention test data for which C-V and R can be computed and 64 different items using posttest-retention test data for which C-V and R can also be computed. These two sets of data were considered separately in the analyses.

The results of the correlations of C-V and R are presented in Table 6.1.

Table 6.1
Correlations of C-V and R

	N	r(C-V, R) (Pretest-Retention)	N#	r(C-V, R) (Posttest-Retention)
All	64	.80**	55	.76**
Math	32	.88**	27	.81**
Reading	32	.87**	28	.67**
Upper	32	.81**	28	.62**
Middle	32	.80**	27	.82**
Treatment A	32	.79**	30	.69**
Treatment B	32	.82**	25	.80**

**Significant at $p < .01$

#Some values of R did not exist because there were no individuals in the combined categories of f_1 (fail-fail) and f_2 (fail-pass). The computation of R involves $f_1 + f_2$ in the denominator and if this is zero, R does not exist.

The correlations between C-V and R for the indices computed on pretest and retention test data range from .79 to .88. All these correlations are significantly different from zero ($p < .01$). Using Fisher's Z-transformation, pairwise comparisons of the correlations between Mathematics and Reading, Upper and Middle, and Treatment A and Treatment B showed no significant differences.

The correlations between C-V and R for the indices computed on posttest and retention test data range from .62 to .82. Again all these correlations are significantly different from zero ($p < .01$). Using Fisher's Z-transformation, pairwise comparisons of the correlations between Mathematics and Reading, Upper and Middle, and Treatment A and Treatment B were made. There were no significant differences.

For both sets of data (items given on the pretest and retention test and items given on the posttest and retention test), the analyses indicate that:

1. The C-V and R values are significantly related and the procedures would result in similar item selection;
2. The C-V and R values are not more related for Mathematics than for Reading;
3. The relationship between the C-V and R procedures does not depend on grade level; and,
4. There is no difference in the relationship between the C-V and R procedures when treatments are considered.

B-S and C-V

The B-S procedure requires that an item be given on a pretest and on a posttest. In this situation, it was necessary to apply the B-S rules to pretest-retention test data only. There are 64 items for which a decision about item revision, using the B-S procedure, can be made. Using the rules on posttest-retention test data is not meaningful.

There is also one additional restriction. To apply some of the decision rules, it is necessary to select cut-off values. The analyses of the items using the B-S approach were based on an arbitrary cut-off value of .50 for the error rates. If the error rate was below .50 the error rate was considered low; if above .50, the error rate was considered high. The original intent was to select multiple cut-off values for the error rates. But the data indicated

that choosing different cut-off values would not change the decision about the revision of the items. Only 18 items met the criterion of no significant positive difference between the theoretical error rates (TER) and the pretest error rates (BER). These items all had values of BER greater than .50. TER, since the items were three-option multiple choice items, is always .67. It would be meaningless to lower the cut-off for the error rates since the same 18 items would be chosen. To raise the cut-off would exclude more items but since the stronger criterion of no significant positive difference between TER and BER is met for these 18 items the increase in the cut-off does not seem particularly reasonable.

First the individual statistics, TER, BER, PER, BDI and PDI were computed for the 64 items. Then the appropriate rules were applied and a decision was made about the quality of the item; that is, does the item need to be revised? Each item was assigned a "0" or a "1" depending on the outcome of the application of the rules. A "0" indicates revision is required, and a "1" indicates no revision is required. See Appendix IV for the statistics on the 64 items and the resulting application of the rules. Application of the rules resulted in 18 items needing no revision.

Point-biserial correlations were computed between the resulting values from the B-S procedure and the C-V values. The correlations are presented in Table 6.2.

The correlations between C-V and B-S values range from .45 to .84. All these correlations are significantly different from zero

Table 6.2
Correlations Between B-S and C-V

	N	$r_{p\text{-}bis}$
All	64	.70**
Mathematics	32	.69**
Reading	32	.50**
Upper	32	.70**
Middle	32	.68**
Treatment A	32	.45**
Treatment B	32	.84**

**Significant at the .01 level.

($p < .01$). Pairwise comparisons reveal the largest difference is between the point-biserials for treatment A and treatment B.

These analyses indicate that application of the B-S or C-V procedure results in selection of many of the same items. In addition, the B-S and C-V procedures are slightly more comparable for Mathematics than for Reading; the relationship between the procedures does not depend on grade level; and the B-S and C-V procedures are considerably more comparable for treatment B than for treatment A.

B-S and R

The same restrictions apply to the comparisons of the B-S and R indices as to the comparisons of the B-S and C-V indices. Point-biserial correlations were computed between the B-S values and the R values. The correlations are presented in Table 6.3.

Table 6.3
Correlations Between B-S and R

	N	$r_{p\text{-}bis}$
All	64	.36**
Mathematics	32	.39*
Reading	32	.24
Upper	32	.37*
Middle	32	.37*
Treatment A	32	.21
Treatment B	32	.52**

*Significant at the .05 level.

**Significant at the .01 level.

The correlations between R and B-S values range from .21 to .52, considerably smaller than the correlations between the C-V and B-S values. Only correlations between all the R and B-S values and the R and B-S values for treatment B are significant at the .01 level. The correlations for Mathematics, Upper and Middle are significant at the .05 level. The correlations are not significantly different from zero for Reading and treatment A. Pairwise comparisons show that the largest difference is between the correlations for treatment A and treatment B.

These analyses indicate that the relationship between the results of the B-S and R procedures is not very strong, but many of the same items would be selected with either procedure. In addition the B-S and R procedures are considerably more comparable for Mathematics than for Reading and for treatment B than for treatment A.

The relationship does not appear to depend on grade level.

B-S and C-V and R

The correlations between the indices, R, C-V and B-S for the 64 pretest-retention test items are all significantly different than zero ($p < .01$). The relationship between the R and B-S values is markedly different than the other two relationships. Table 6.4 summarizes the three correlations.

Table 6.4
Correlations for All Items

	(N=64)	
	R	B-S
C-V	.80**	.70**
B-S	.36**	

**Significant at the .01 level.

These significant correlations indicate that the three indices are related. In particular the R and C-V indices are the most similar. The R index, however, does not appear to give results as similar to the B-S procedure as does the C-V index.

Consider the correlations between the indices for each subject area, grade level and treatment for the 64 items (pretest-retention test). These correlations are reported in Tables 6.5 A and B, 6.6 A and B and 6.7 A and B.

Table 6.5 A
Correlations--Mathematics

	R	B-S
C-V	.88**	.69**
B-S	.39*	

Table 6.5 B
Correlations--Reading

	R	B-S
C-V	.87**	.50**
B-S	.24	

Table 6.6 A
Correlations--Middle

	R	B-S
C-V	.80**	.68**
B-S	.37*	

Table 6.6 B
Correlations--Upper

	R	B-S
C-V	.81**	.70**
B-S	.37*	

Table 6.7 A
Correlations--Treatment A

	R	B-S
C-V	.79**	.45**
B-S	.21	

Table 6.7 B
Correlations--Treatment B

	R	B-S
C-V	.82**	.84**
B-S	.52**	

*Significant at the .05 level.

**Significant at the .01 level.

Based on these correlations it appears that the three procedures are more comparable for items in Mathematics than for items in Reading, and for items given in treatment B than for items given in treatment A.

Although all the procedures are significantly related for Mathematics the relationship between the B-S procedure and the R index is markedly different than the R--C-V and C-V--B-S relationships. This same difference in the size of the relationships appears in all of the other comparisons, i.e. Reading, Middle, Upper, treatment A, and treatment B. The difference is less for treatment B correlations than for the other comparisons.

An alternate method of analyzing the comparability of the three approaches would be to consider the agreement among the three methods. If a cut-off value for the C-V and R index is chosen as .50, i.e. those items with an R or C-V value equal or above .50 are considered to be good items, then 32 items out of 64 items would be selected based on the R values and 11 items would be selected based on the C-V values. Of the 32 items selected based on the R values, 13 were also selected using the B-S procedure. All 11 items selected based on the C-V values were selected using the B-S procedure. Similarly all 11 items selected based on the C-V values were selected using the R procedure (see Table 6.8).

Table 6.9 represents the agreement among the three indices. There is complete agreement for 39 of the 64 items or 61 percent. Of the items where there is 100 percent agreement, 21 of the 39 were Reading items (54 percent); 20 of the 39 were given in the Middle grades (51 percent) and 15 of the 39 items were used in treatment A (38 percent). The disagreement among procedures is more noticeable between treatments. Of the 64 items there is agreement between the C-V and B-S procedures for 57 items or 89 percent. There is considerably less agreement between the C-V and R and R and B-S, the percentage agreement being 69 percent and 64 percent respectively.

Summary

The purpose of this part of the study was to determine the comparability of three item analysis procedures; C-V, R and B-S. Sixteen objectives were chosen from two subject areas, Mathematics

Table 6.8
B-S, R and C-V Values for Items Given
on the Pretest and Retention Test

Item Identification*	B-S	R	C-V
R116G1 RAM	0	.1	.03
R116G2 RAM	0	.48	.25
R116G3 RAM	0	.11	.02
R116G4 RAM	0	0	0
R120S1 RBM	0	0	0
R120S2 RBM	0	0	0
R120S3 RBM	0	-1.0	-.25
R120S4 RBM	0	-1.0	-.15
R142G1 RAM	0	1.0	.06
R142G2 RAM	1	1.0	.52
R142G3 RAM	0	.86	.19
R142G4 RAM	1	.88	.48
R112S1 RBM	0	.71	.24
R112S2 RBM	0	.75	.14
R112S3 RBM	1	.09	.05
R112S4 RBM	0	.33	.05
M109S1 MBM	0	.33	.03
M109S2 MBM	0	.27	.11
M109S3 MBM	0	.5	.19
M109S4 MBM	0	.21	.08
M108G1 MAM	0	.71	.38
M108G2 MAM	1	.61	.38
M108G3 MAM	1	.90	.73
M108G4 MAM	1	.83	.58
M107S1 MBM	0	.06	.024
M107S2 MBM	0	0	0
M107S3 MBM	0	.18	.071
M107S4 MBM	0	.06	.024
M111G1 MAM	0	.75	.28
M111G2 MAM	0	.65	.26
M111G3 MAM	0	.47	.19
M111G4 MAM	0	.53	.21
M187S1 MBU	1	.90	.56
M187S2 MBU	1	.91	.63
M187S3 MBU	1	1.0	.63
M187S4 MBU	1	1.0	.75
R182S1 RBU	0	.67	.2
R182S2 RBU	0	0	0
R182S3 RBU	0	.33	.07
R182S4 RBU	0	-.33	-.03
M167S1 MBU	1	.75	.53
M167S2 MBU	1	.73	.65
M167S3 MBU	1	.8	.71
M167S4 MBU	1	.71	.59

Table 6.8--Continued

Item Identification*	B-S	R	C-V
R166S1 RBU	0	1.0	.33
R166S2 RBU	0	.2	.06
R166S3 RBU	0	.67	.11
R166S4 RBU	0	.5	.17
M198G1 MAU	0	.5	.14
M198G2 MAU	0	.67	.18
M198G3 MAU	0	.7	.32
M198G4 MAU	0	.5	.14
M176G1 MAU	1	.23	.15
M176G2 MAU	1	.26	.26
M176G3 MAU	1	.04	.04
M176G4 MAU	1	.11	.11
R145G1 RAU	0	.27	.09
R145G2 RAU	0	.5	.20
R145G3 RAU	0	.3	.15
R145G4 RAU	0	.3	.11
R199G1 RAU	0	-.4	-.14
R199G2 RAU	0	-.5	-.11
R199G3 RAU	0	-1.78	-.28
R199G4 RAU	0	-.125	-.05

*The last three letters of the Item Identification refer to subject area (M = Mathematics, R = Reading); treatment (A or B); and grade level (M = Middle, U = Upper).

Table 6.9
Agreement of the Three Item Indices
100% Agreement

	Items Unacceptable		Items Acceptable		Total
All	28	(44%)	11	(17%)	39 (61%)
Mathematics	8	(20%)	10	(26%)	18 (46%)
Reading	20	(51%)	1	(3%)	21 (54%)
Middle	17	(43%)	3	(8%)	20 (51%)
Upper	11	(28%)	8	(21%)	19 (49%)
Treatment A	12	(30%)	3	(8%)	15 (38%)
Treatment B	16	(41%)	8	(21%)	24 (62%)

Table 6.9 A
Agreement of the Three Item Indices
67% Agreement

	Items Unacceptable		Items Acceptable		Total
All	23	(36%)	2	(3%)	25 (39%)
Mathematics	13	(52%)	1	(4%)	14 (56%)
Reading	10	(40%)	1	(4%)	11 (44%)
Middle	10	(40%)	2	(8%)	12 (48%)
Upper	13	(52%)	0	(0%)	13 (52%)
Treatment A	15	(60%)	2	(8%)	17 (68%)
Treatment B	8	(32%)	0	(0%)	8 (32%)

Table 6.9 B
Agreement of the Three Item Indices
67% Agreement

	Items Unacceptable			Items Acceptable			Total		
	R, B-S	C-V, B-S	R, C-V	R, B-S	C-V, B-S	R, C-V	R, B-S	C-V, B-S	R, C-V
All	0 (0%)	18 (28%)	5 (8%)	2 (3%)	0 (0%)	0 (0%)	2 (3%)	18 (28%)	5 (8%)
Mathematics	0 (0%)	9 (50%)	4 (80%)	1 (50%)	0 (0%)	0 (0%)	1 (50%)	9 (50%)	4 (80%)
Reading	0 (0%)	9 (50%)	1 (20%)	1 (50%)	0 (0%)	0 (0%)	1 (50%)	9 (50%)	1 (20%)
Middle	0 (0%)	9 (50%)	1 (20%)	2(100%)	0 (0%)	0 (0%)	2(100%)	9 (50%)	1 (20%)
Upper	0 (0%)	9 (50%)	4 (80%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9 (50%)	4 (80%)
Treatment A	0 (0%)	11 (61%)	4 (80%)	2(100%)	0 (0%)	0 (0%)	2(100%)	11 (61%)	4 (80%)
Treatment B	0 (0%)	7 (39%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	7 (39%)	1 (20%)

and Reading, two grade levels, Middle and Upper, and two treatments, assigned objectives (treatment A) and selected objectives (treatment B). A total of 64 items were analyzed using each of the three item analysis procedures. An additional 64 items were analyzed using only the C-V and R procedures.

The major question to be answered was do the C-V, R and B-S item analysis procedures provide comparable results? Three additional questions also were considered:

1. Are the three procedures more comparable for items in Mathematics than for items in Reading?;
2. Does the comparability of the three procedures depend on the grade level?; and,
3. Are the three procedures more comparable for items given in treatment A than for items given in treatment B?

Correlation coefficients were computed between the indices for the 64 items given on a pretest and a retention test. The Pearson product moment correlation coefficient between the R and C-V indices was significantly different than zero ($r = .80$, $p < .01$). The point-biserial correlation coefficients between the B-S procedure and the C-V index and the B-S procedure and the R index were also significantly different than zero ($r = .70$, $p < .01$ and $r = .36$, $p < .01$, respectively). These correlations indicate that the three indices are related and provide reasonably comparable results.

The separate analyses of the indices for each subject area, grade level and treatment indicated that the indices were more comparable for Mathematics than for Reading, with all significant

correlations between the indices for Mathematics [r(R,C-V) = .88, $p < .01$; r(C-V, B-S) = .69, $p < .01$; r(R,B-S) = .39, $p < .05$] and only two out of the three correlations significant for Reading [r(R,C-V) = .87, $p < .01$; r(C-V, B-S) = .50, $p < .01$; r(R,B-S) = .24, not significant]. The indices were also more comparable for treatment B than for treatment A, with all significant correlations for treatment B [r(R,C-V) = .82, $p < .01$; r(C-V, B-S) = .84, $p < .01$; r(R,B-S) = .52, $p < .01$] and only two out of the three correlations significant for treatment A [r(R,C-V) = .79, $p < .01$; r(C-V, B-S) = .45, $p < .01$; r(R, B-S) = .21, not significant]. The correlations between the indices for the grade levels, Middle and Upper, were almost identical, with r(C-V, R) = .80 for Middle and .81 for Upper, r(C-V, B-S) = .68 for Middle and .70 for Upper, r(R, B-S) = .37 for both Middle and Upper. Although all the correlations were significant, the correlations between the R index and the B-S procedure were significant at the .05 level while the other correlations were significant at the .01 level.

The comparison of the R and C-V indices on the 64 items given on a posttest and a retention test provide additional support that the use of either the R index or C-V index would result in selection of many of the same items. Although all the correlations were significant at the .01 level, the correlations between R and C-V for Mathematics (.81), Middle (.82) and treatment B (.80) were larger than the correlations between R and C-V for Reading (.67), Upper (.62) and treatment A (.69).

It is interesting to note that the predictions that the indices would be more related for Mathematics than for Reading and more related for treatment B than for treatment A were supported by the results. In addition, pairwise comparisons of the indices by grade level did not reveal any differences. This was also predicted.

An analysis of the agreement among the three item analysis procedures showed that when a cut-off of .50 for the R and C-V indices was used for selection of items, there was complete agreement for 39 of the 64 items (61 percent) given on the pretest and retention test.

In the final chapter the results presented in Chapter V and Chapter VI are reviewed. The implications of these results for test development are also discussed in Chapter VII.

CHAPTER VII

SUMMARY AND CONCLUSIONS

Summary

The first purpose of this study was to propose a theoretical conception of criterion-referenced testing and to explain two basic item analysis techniques (Cox and Vargas, C-V and Roudabush, R) theoretically with respect to this general model. The second purpose was to determine the adequacy of the C-V and R procedures using the theoretical model. The final purpose was to compare three item analysis techniques, the C-V, R and the Brennan and Stolurow (B-S), using real data.

Previous research indicated that the C-V index, defined as the difference between the proportion of individuals answering the item correctly at posttest and the proportion of individuals answering the item correctly at pretest, was an appropriate item analysis technique for criterion-referenced tests. Most of the comparative studies did compare the C-V index to other indices, but in general these other indices were traditional indices rather than other proposed item analysis techniques for criterion-referenced tests.

The other two indices included in the study (R and B-S) had not been previously researched. The R index is a refinement of the C-V technique and the B-S procedure combines traditional methods

with a set of decision rules to provide a guide for selecting items which are sensitive to instruction.

A theoretical model for criterion-referenced testing was proposed. The model includes 12 parameters that completely describe the pretest-posttest situation. The model is represented in matrix notation as:

$$\underline{P} = Q \boxtimes Q' \underline{\pi}$$

where $\boxtimes$ symbolizes the Kronecker product, and

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}, \quad Q' = \begin{pmatrix} q_{11}' & q_{12}' \\ q_{21}' & q_{22}' \end{pmatrix}, \quad \underline{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix},$$

and

$$\underline{P} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

and q_{ij}, q_{ij}', π_k and p_k are defined by the Tables 7.1, 7.2, 7.3 and 7.4 below.

Table 7.1
Pretest--Actual

		Does Not Know	Knows
OBSERVED	Fail	q_{11}	q_{12}
	Pass	q_{21}	q_{22}

Table 7.2
Posttest--Actual

		Does Not Know	Knows
OBSERVED	Fail	q_{11}'	q_{12}'
	Pass	q_{21}'	q_{22}'

Table 7.3
Categories for a Given Item--True Proportions

		Posttest	
		Does Not Know	Knows
PRETEST	Does Not Know	π_1	π_2
	Knows	π_3	π_4

Table 7.4
Categories for a Given Item--Observed Proportions

		Posttest	
		Fail	Pass
PRETEST	Fail	p_1	p_2
	Pass	p_3	p_4

The q_{ij}'s and q_{ij}''s are conditional probabilities, with $q_{11} + q_{21} = 1$, $q_{12} + q_{22} = 1$, $q_{11}' + q_{21}' = 1$, $q_{12}' + q_{22}' = 1$, the π_k's are true proportions and the p_k's are observed proportions with $\sum_{k=1}^{4} \pi_k = 1$ and $\sum_{k=1}^{4} p_k = 1$.

The R and C-V indices can be explained in terms of this general model by making certain assumptions. The theoretical framework can be simplified by assuming that the pretest and posttest conditional probabilities are equal, $Q = Q'$. Additional assumptions needed for the model to fit the R procedure are that $\pi_3 = 0$, i.e. there is no forgetting, and $q_{22} = q_{22}' = 1.0$, i.e. someone who knows the answer to an item can not fail the item. So the general model can be reduced to

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} q_{11} & 0 \\ q_{21} & 1 \end{pmatrix} \cdot \begin{pmatrix} q_{11}' & 0 \\ q_{21}' & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_0 \\ \pi_4 \end{pmatrix}$$

which reduces to the four equations used by Roudabush in the development of the R index.

The C-V index is a further simplification of the general model where $q_{21} = 0$, i.e. there is no guessing. However, π_3 is not assumed to be zero. So for the C-V index, the model is

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} \quad \text{or } \underline{P} = \underline{\pi}.$$

There were two parts to this study. The first part attempted to determine if the C-V and R indices adequately estimated the true values, if one technique estimated the true values better than the other and if the C-V and R indices were better estimators of the true values for some parameter sets. These questions were investigated by simulating data for 21 different sets of parameter values using the model described in Chapter III and briefly described above, as the theoretical framework. For each set of parameter values 1000 samples of size 50 or 200 were generated. The R and C-V indices were computed for each sample. Descriptive statistics, means, variances, kurtosis and skewness values were computed based on the R and C-V values for the 1000 samples for each parameter set.

The adequacy of R and C-V was determined by consideration of a number of descriptive statistics, including means, variances, kurtosis, skewness, and range. It was found that for R, when the assumptions were met, the technique provided a more stable and accurate estimate than when the assumptions were not met. It was also found that when the sample size was increased from 50 to 200, the stability and accuracy increased greatly. A correlation coefficient of .759 between the true R values and the mean R values for 1000 estimates over the 21 parameter sets indicated that the technique does produce a reasonable estimate of the true R.

The C-V technique seemed to provide a reasonably accurate and stable estimate regardless of whether the assumptions were met. The estimates were more stable with larger sample sizes. A correlation coefficient of .80 between the true C-V values and the mean C-V

values for 1000 estimates over the 21 parameter sets indicated that the technique does produce a reasonable estimate of the true C-V.

The C-V technique seemed to estimate the C-V true value better than the R technique estimated the R true value. The average absolute deviation from the respective true values was smaller for C-V than for R (.0891 and .0965, respectively). In addition the range of variances was considerably smaller for the C-V estimates than for the R estimates (.0013 to .0094 for C-V and .0015 to .0221 for R) and the average range of estimated values was smaller (.46 for C-V and .60 for R).

The third questions, i.e. were the C-V and R techniques better estimators of the true values for some parameter sets, was primarily answered by considering the question of adequacy and stability. For both techniques the estimates were better when the sample size was larger. In addition the R approach was better when the assumptions were met. This was not true for the C-V approach which seemed to provide a good estimate under almost any assumptions.

The second part of the study was designed to determine the comparability of the three item analysis procedures, R, C-V and B-S. C-V and R values were computed for 128 items and B-S values were computed for 64 items. These items were testing 16 objectives from two subject areas, Mathematics and Reading, two grade levels, Middle and Upper, and two treatments, assigned objectives (treatment A) and selected objectives (treatment B).

The major question to be answered was whether the C-V, R and B-S item analysis procedures provide comparable results. Three additional questions were also considered:

1. Are the three procedures more comparable for items in Mathematics than for items in Reading?;
2. Does the comparability of the three procedures depend on the grade level?; and,
3. Are the three procedures more comparable for items given in treatment A than for items given in treatment B?

Correlation coefficients were computed between the indices for the 64 items given on a pretest and a retention test. The Pearson product moment correlation coefficient between the R and C-V indices was significantly different than zero ($r = .80$, $p < .01$). The point-biserial correlation coefficients between the B-S procedure and the C-V index and the B-S procedure and the R index were also significantly different than zero ($r = .70$, $p < .01$ and $r = .36$, $p < .01$, respectively). These correlations indicated that the three indices were related and similar results in item selection would be obtained using any of the three approaches.

The separate analyses of the indices for each subject area, grade level and treatment indicated that the indices were more comparable for Mathematics than for Reading, with all significant correlations between the indices for Mathematics [$r(R, C\text{-}V) = .88$, $p < .01$; $r(C\text{-}V, B\text{-}S) = .69$, $p < .01$; $r(R, B\text{-}S) = .39$, $p < .05$] and only two out of the three correlations significant for Reading [$r(R, C\text{-}V) = .87$, $p < .01$; $r(C\text{-}V, B\text{-}S) = .50$, $p < .01$; $r(R, B\text{-}S) = .24$, not

significant]. The indices were also more comparable for treatment B than for treatment A, with all significant correlations for treatment B [r(R,C-V = .82, $p < .01$; r(C-V, B-S) = .84, $p < .01$; r(R, B-S) = .52, $p < .01$] and only two out of the three correlations significant for treatment A [r(R, C-V) = .79, $p < .01$; r(C-V, B-S) = .45, $p < .01$; r(R, B-S) = .21, not significant]. The correlations between the indices for the grade levels, Middle and Upper, were almost identical, with r (C-V, R) = .80 for Middle and .81 for Upper, r(C-V, B-S) = .68 for Middle and .70 for Upper, r(R, B-S) = .37 for both Middle and Upper. Although all the correlations were significant, the correlations between the R index and the B-S procedure were significant at the .05 level while the other correlations were significant at the .01 level.

The comparison of the R and C-V indices on the 64 items given on a posttest and a retention test provided additional support that the use of either the R index or the C-V index would identify many of the same items as good or bad. Although all the correlations were significant at the .01 level, the correlations between R and C-V for Mathematics (.81), Middle (.82) and treatment B (.80) were larger than the correlations between R and C-V for Reading (.67), Upper (.62) and treatment A (.69).

An analysis of the agreement among the three item analysis procedures showed that when a cut-off of .50 for the R and C-V indices was used for selection of items, there was complete agreement for 39 of the 64 items (61 percent) given on the pretest and retention test.

Conclusions

There does exist a general model that explains the pretest-posttest situation. This general model can be used in the development and explanation of item analysis techniques. The R and C-V techniques fit the general model with some additional assumptions.

The results of the simulation indicate that the C-V and R techniques adequately estimate the true values of the item parameters. However, the C-V procedure provides better estimates of the true values when there are deviations from appropriate assumptions. In general the C-V technique seems to estimate the C-V true value better than the R technique estimates the R true value. Therefore the C-V item analysis technique probably should be used for analyzing items from pretest-posttest situations. Both techniques improve with an increase in sample size. One perhaps can infer from this that in developing tests it would be best to use a sample size larger than 50.

The major question to be answered in the study with actual data was, "Do the C-V, R and B-S item analysis procedures provide comparable results?" It was found that the three indices were related and did provide results that were reasonably comparable. The comparability was stronger for items in Mathematics and treatment B (selected objectives). The R and C-V procedures were more comparable than the B-S and R techniques but the correlations between the C-V and B-S procedures were close to the correlations between the R and C-V techniques.

In summary:

1. A theoretical conception or a general model of criterion-referenced testing can be defined.
   a. The R technique fits the general model given certain assumptions.
   b. The C-V technique fits the general model given certain assumptions.
2. The C-V and R techniques provide reasonable estimates of the true values of the respective indices.
   a. The C-V technique estimates the true C-V values better than the R technique estimates the true R values.
   b. The R technique estimates the true R values better when the assumptions are met and when the sample size is larger.

      The C-V technique while reasonably accurate under any assumptions does become more stable as the sample size increases.
3. The C-V, R and B-S item analysis procedures are related and similar results would be obtained using any of the three procedures.
   a. The three procedures are more comparable for items in Mathematics than for items in Reading.
   b. The comparability of the three procedures does not depend on the grade level.

Discussion

This study was intended to determine if an accurate and easy-to-use item analysis technique existed among the three techniques. The results of the simulation provided evidence that the C-V index is a reasonably accurate procedure for the estimation of the true C-V values even when the assumptions are not met. The R index, on the other hand, is less accurate and less stable.

The result of the comparisons with actual data leads to the conclusion that the C-V index provides reasonably close approximations to the other two methods and is the easiest to compute. The B-S procedure, while providing a generous amount of information, is tedious and time-consuming. The R procedure, while not necessarily more difficult to compute than the C-V index, is perhaps less understandable to the everyday practitioner and provides no more information.

Not only does this study provide information as to which item analysis technique is most accurate, most stable and easiest to compute of the three approaches considered, it also suggests a theoretical framework. The theoretical development explains two item analysis approaches and demonstrates how the pretest-posttest situation can be conceptualized. Other studies have failed to explain, at least so explicitly, the underlying framework of pretest-posttest situations and the reasoning behind the suggested methods of analyses of the items.

While this study has suggested one method which most likely is the best method for analyzing items and has presented a theoretical

framework, there is one limitation to be considered. The analyses of actual data were limited to the extent that the sample sizes for each item were relatively small. The results of the analyses, however, were in agreement with the results of the simulation, i.e. the C-V index is a reasonable method of item analysis for criterion-referenced tests. This agreement between the two parts of the research tends to lessen the impact of the small sample sizes.

As was pointed out in Chapter II, there is a need to be alert to the negative implications of selecting items sensitive to instruction. If items are selected which are sensitive to instruction one might argue that the items, over a number of administrations and revisions, could become very easy or perhaps require only recall of simple facts. Care must be taken to include items that measure all aspects of the domain and to ensure that these items are not only sensitive to instruction but sensitive to the domain. In addition, items after a number of administrations and revisions should probably be piloted in a group consisting of individuals with and without previous instruction. The quality of the items should be checked using a number of statistical procedures including traditional statistics. The individuals included in this pilot who have received instruction should probably have not just received instruction.

Implications for Future Research

This study, along with other research on item analysis procedures for criterion-referenced tests, points to a practical, easy-

to-use item analysis index, the C-V index. The study does not predict that tests developed using this index would be the most valid and reliable (by whatever definition) tests. However, this study does indicate that the C-V index provides a reasonable estimate of the sensitivity of the item to instruction. In addition, the study did show that the index is reasonably comparable to two other more complicated (or refined) procedures.

It is interesting to note that the R index which is a more reasonable one, i.e. less restrictive and fewer assumptions are needed, does not prove to be the better technique. In fact, the R procedure provides poor estimates of the true R value and is very unstable.

Additional research should probably be concerned with the theoretical conceptualization of criterion-referenced measurement (pretest-posttest situations) that was proposed in this study. The theoretical framework could provide a basis for future research in several areas. One of these areas is the estimation of some of the parameters. Unfortunately, the model contains 12 parameters and with only three pieces of information, p_1, p_2 and p_3, available it is not possible to estimate the 12 parameters. However, some restrictive assumptions could be applied and perhaps, then some of the underlying parameters could be estimated.

Also within the estimation process there is the possibility of determining the probability of an individual who knows the item will actually pass the item. This type of information could be valuable in being confident of an individuals' attainment of some

given mastery level. Information such as that suggested above, that can be obtained from further investigation of the general model may prove valuable to the improvement of criterion-referenced measurement.

APPENDICES

APPENDIX I

Roudabush's Technique

Roudabush's Technique

For this model, consider the folowing 2 x 2 table:

Table I.1
Categories for a Given Item

		Posttest	
		Failed	Passed
PRETEST	Failed	f_1	f_2
	Passed	f_3	f_4

where f_1 equals the number of students who failed the item at both pretest and posttest; f_2 equals the number of students who failed the item at pretest and passed it at posttest; f_3 equals the number of students who passed the pretest and failed the posttest; and f_4 equals the number of students who passed the item at pretest and posttest.

Now assume there is some fixed non-zero probability, p, that a student who does not know the answer to the item will guess the correct answer. This p-value is determined by the item only and does not vary from student to student nor from occasion to occasion for the same student. This fixed p-value suggests that there is no partial knowledge on the part of the student, and that the student's responses are independent at pretest and posttest when he does not know the correct answer and fails to learn it.

Assume, also, that the only possible result of exposure between pretest and posttest is that the student learn the correct response to an item. Assume that the non-zero frequency of f_3 is solely due to guessing, further implying that there is no forgetting and implying that the "true" value of f_3 is zero.

With these assumptions Roudabush derives a number which serves as an index of the degree to which examinees select the correct response to the item as a function of the instruction received between pretest and posttest. This number is called a sensitivity index (s) by Roudabush.

In order to clarify this procedure, it is necessary to sketch briefly the derivation of the sensitivity index.

Since we have already assumed that the "true" value of f_3 is zero ($\hat{f}_3 = 0$) then we can say that f_1 is the probability of guessing wrong twice times the number of students in the sample who do not learn the answer. We can state this symbolically as

$$f_1 = (1 - p)^2 \hat{f}_1 \qquad (1)$$

where $\hat{f}_1$ is the "true" number of students who do not learn.

Similarly, f_2 enumerates those students who learned the answer after instruction and did not guess correctly at the pretest, and those who did not learn but were able to guess the correct response at the posttest but not at the pretest. This would then say that

$$f_2 = p(1 - p) \hat{f}_1 + (1 - p) \hat{f}_2 \qquad (2)$$

where $\hat{f}_2$ is the "true" number of students in the sample who did not know at the pretest but have learned by the posttest.

Next f_3, defined as the number of students who passed the item at pretest but failed it at posttest, enumerates those students who correctly guessed the item at pretest, but did not learn the correct answer via instruction and were not able to guess the correct answer at posttest. Therefore,

$$f_3 = p(1 - p)\ \hat{f}_1. \tag{3}$$

Finally, f_4, defined as the number of students who passed the item at both pretest and posttest, enumerates students from three different categories. The first category comprises all the students who do know the correct answer at pretest and posttest. The second category consists of those students who learned the correct answer for the posttest and guessed correctly the answer at pretest. The third category represents the students who did not know nor learned the answer, but were able to guess correctly at both pretest and posttest. We can represent this as

$$f_4 = \hat{f}_4 + p\hat{f}_2 + p^2\hat{f}_1 \tag{4}$$

where $\hat{f}_4$ if the "true" number of students in the sample who know the correct answer at both pretest and posttest.

Using equations (1) and (3) we can solve for p:

$$f_1 = (1 - p)^2 \hat{f}_1$$

$$f_3 = p(1 - p) \hat{f}_1$$

$$\frac{f_1}{f_3} = \frac{(1 - p)}{p}$$

$$f_1 p = f_3 (1 - p)$$

$$f_1 p + f_3 p = f_3$$

$$p(f_1 + f_3) = f_3$$

$$p = \frac{f_3}{f_1 + f_3} \tag{5}$$

Now we need to find solutions for $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_4$. (Recall we have already assumed $\hat{f}_3 = 0$.)

Since $f_1 = (1 - p)^2 \hat{f}_1$ and $p = \frac{f_3}{f_1 + f_3}$, then

$$f_1 = \left(1 - \frac{f_3}{f_1 + f_3}\right)^2 \hat{f}_1$$

$$f_1 = \left(\frac{f_1 + f_3 - f_3}{f_1 + f_3}\right)^2 \hat{f}_1$$

$$f_1 = \left(\frac{f_1}{f_1 + f_3}\right)^2 \hat{f}_1$$

$$\hat{f}_1 = f_1 \left(\frac{f_1 + f_3}{f_1}\right)^2 = \frac{(f_1 + f_3)^2}{f_1} \tag{6}$$

Also since $f_2 = p(1 - p)\hat{f}_1 + (1 - p)\hat{f}_2$ and $p = \frac{f_3}{f_1 + f_3}$ and $\hat{f}_1$ is equal to the above (6), we have

$$f_2 = \left(\frac{f_3}{f_1 + f_3}\right)\left(\frac{f_1}{f_1 + f_3}\right)\frac{(f_1 + f_3)^2}{f_1} + \left(\frac{f_1}{f_1 + f_3}\right)\hat{f}_2$$

$$f_2 = f_3 + \left(\frac{f_1}{f_1 + f_3}\right)\hat{f}_2$$

$$f_2 - f_3 = \left(\frac{f_1}{f_1 + f_3}\right)\hat{f}_2$$

$$\hat{f}_2 = \frac{(f_2 - f_3)(f_1 + f_3)}{f_1} \qquad (7)$$

And finally, since $f_4 = \hat{f}_4 + p\hat{f}_2 + p^2\hat{f}_1$ and using (5), (6) and (7):

$$f_4 = \hat{f}_4 + \left(\frac{f_3}{f_1 + f_3}\right)\frac{(f_2 - f_3)(f_1 + f_3)}{f_1} + \left(\frac{f_3}{f_1 + f_3}\right)\frac{(f_1 + f_3)^2}{f_1}$$

$$f_4 = \hat{f}_4 + \frac{f_3(f_2 - f_3)}{f_1} + \frac{f_3^2}{f_1}$$

$$\hat{f}_4 = f_4 + \frac{f_3^2 - f_3 f_2}{f_1} - \frac{f_3^2}{f_1}$$

$$\hat{f}_4 = \frac{f_1 f_4 - f_3 f_2 + f_3^2 - f_3^2}{f_1}$$

$$\hat{f}_4 = \frac{f_1 f_4 - f_3 f_2}{f_1} \text{ or } \hat{f}_4 = f_4 - \frac{f_3 f_2}{f_1}.$$

The sensitivity index is then defined as

$$R = \frac{\hat{f}_2}{\hat{f}_1 + \hat{f}_2}.$$

Substituting the observed frequencies for the true frequencies we have:

$$R = \frac{\dfrac{(f_2 - f_3)(f_1 + f_3)}{f_1}}{\dfrac{(f_1 + f_3)^2}{f_1} + \dfrac{(f_2 - f_3)(f_1 + f_3)}{f_1}}$$

$$R = \frac{(f_2 - f_3)(f_1 + f_3)}{(f_1 + f_3)^2 + (f_2 - f_3)(f_1 + f_3)}$$

$$R = \frac{f_2 - f_3}{f_1 + f_3 + f_2 - f_3}$$

$$R = \frac{f_2 - f_3}{f_1 + f_2} .$$

APPENDIX II

Brennan's and Stolurow's Procedure

Brennan's and Stolurow's Procedure

This procedure was suggested by Brennan and Stolurow (AERA, February 1971). Using traditional item analysis techniques, they combine four "error rates" and two discrimination indices with a set of rules.

The Theoretical Error Rate (TER) is one suggested error rate. This is the expected proportion of students answering a pretest item incorrectly simply on the basis of random guessing. If J is the number of possible answers to an item, then TER = (J - 1)/J. A second error rate is called the Base Error Rate (BER). This is the observed proportion of students answering a pretest item incorrectly. The third error rate, Instructional Error Rate (IER) is the error rate on a terminal test item for a given objective obtained by students who have been exposed to instruction. In addition a Posttest Error Rate (PER) is used. PER is the observed proportion of students answering a posttest item incorrectly. IER is only used in the decision rules related to instruction so it will not be included in further discussions.

The two discrimination indices used are the Base Discrimination Index (BDI) and the Posttest Discrimination Index (PDI). Discrimination indices are computed using the total score on the appropriate test as the criterion. For BDI, the criterion will be the pretest total score. For PDI, the criterion will be the posttest total score.

The error rates are classified as high (H) or low (L) with the evaluator predetermining an appropriate cut-off point between high and low error rate. In addition the discrimination indices can be classified as positive, negative or non-discriminating. By positive and negative indices, it is meant that the indices discriminate significantly (at some α - level) in the positive and negative directions, respectively. Brennan and Stolurow recommend the phi-coefficient and the B index (Brennan, 1972) for criterion-referenced tests.

An abbreviated list of the rules that Brennan and Stolurow suggest are presented in the following table.

Table II.1
Rules for Decision-Making

Rule No.	TER	BER	BDI	PER	PDI	Item Decision[a]
1.	H	H				NR
	L	L				NR
2.	L	H				NR
3.	H	L				R
4.			-			?
6.				L	0	NR
7.				L	+	?
				L	-	?
8.				H	-	R
9.				H	+	?
				H	0	?
11.			-		-	R

Table II.1--Continued

Rule No.	TER	BER	BDI	PER	PDI	Item Decision[a]
16.	DER*[b]					R
	DER(NS)[c]					NR

[a]"NR" means no revision required.

"R" means revision is required.

"?" means the data are not sufficient to make a sound judgment about whether or not revision is required.

[b]DER is significantly greater than zero at the .05 level for a one-tailed test of significance.

[c]DER is not significantly greater than zero at the .05 level for a one-tailed test of significance.

*DER is defined as TER minus BER and stands for "Difference Error Rate."

The significance of a positive difference between BER and TER can be tested by computing:

$$Z = \frac{DER - \frac{1}{2N}}{\sqrt{TER(1 - TER)/N}}$$

where N is the total number of students in the sample.

According to Brennan and Stolurow, this computed Z value is then compared with the normal curve standard score at an appropriate level of significance for a one-tailed test.

APPENDIX III

Further Analyses of C-V and R

Further Analyses of C-V and R

The basic model $\underline{P} = Q \blacksquare Q' \underline{\pi}$ provides an expected proportion of individuals for each category. For any observed frequency (f_i) produced by the simulation the f_i are distributed multinomially with parameters N, p_1, p_2 and p_3. (Of course, p_4 is understood.) In other words, $f_i \sim MN(N, p_1, p_2, p_3)$. Based on this information a theoretical estimate of the mean and variance of each index can be determined.

C-V

Let's consider the C-V index because it is simplest to understand.

The expected value of the C-V estimates for each parameter set can be determined by noting, as above, that $f_i \sim MN(N, p_1, p_2, p_3)$. The expected value of f_i/N is simply by definition $E(f_i/N) = p_i$. Therefore, the $E(f_2/N - f_3/N) = p_2 - p_3$. The C-V index is defined as $\frac{f_2 - f_3}{N}$. So the mean (or expected value) of the C-V index is $p_2 - p_3$.

The variance can also be theoretically estimated by

$$V(C\text{-}V) = \frac{p_2q_2}{N} + \frac{p_3q_3}{N} + 2\frac{p_2p_3}{N}$$

since the variance of a difference is the sum of the variances minus twice the covariance. The covariance in the case of a multinomial distribution is Cov f_i, $f_j = -N\, p_ip_j$.

Additionally we can estimate the standard error of the mean (expected value) as the standard deviation divided by the square root of n, which is 1000 in our case. Also the standard error of the variance can be approximated by the square root of

$$\frac{(\hat{\sigma}^2)^2}{n} \text{ (Kurtosis + 3) (Kurtosis + 2),}$$ [1]

where n = 1000, $\hat{\sigma}^2$ is the sample variance, and kurtosis is the sample kurtosis.

If we consider parameter set #1 from the simulation, (see Table 4.5 for the parameter values), $p_1 = .1125$, $p_2 = .5075$, $p_3 = .0375$ and $p_4 = .3425$. The expected value of the C-V index is $p_2 - p_3$ or .4700. The reported mean C-V value is .4722 (Table 5.1). If we also compute the standard error of the mean, $s_d/\sqrt{n}$, where n = 1000, we find that the standard error is .0023. The reported mean C-V of .4722 is within one standard error of the expected value.

Also consider this same parameter set with respect to the estimate of the variance. For parameter set #1, the expected variance is .006482. The reported variance is .0070. The standard error of the variance is .0052. So the reported variance is within one (approximated) standard error of the expected variance.

R

A similar analysis of the expected value (mean) of R can be done. However, this analysis is considerably more complicated. The

---

[1] The approximation for the standard error of the variance is based on calculations and formulae from Sampling Techniques, William C. Cochran, 2nd Edition, 1963, John Wiley & Sons, Inc. and Statistics in Biology, Vol. 1, C.I. Bliss, 1967, McGraw-Hill.

expected value can be approximated by $\frac{p_2 - p_3}{p_1 + p_2}$ which is simply saying that the $E(x_1/x_2)$ is approximately equal to the $E(x_1)/E(x_2)$ where $E(x_1) = p_2 - p_3$ and $E(x_2) = p_1 + p_2$.

There are problems with the R index and any theoretical derivations since theoretically and practically the denominator can be zero.

However, for our purposes consider the approximation of the mean R $\frac{(p_2 - p_3)}{p_1 + p_2}$ for parameter #1.

Again $p_1 = .1125$, $p_2 = .5075$, and $p_3 = .0375$, then the expected value (mean) equals .7580. Note the closeness of the reported value of .7553. If we compute the standard error of the mean ($s_d/\sqrt{n}$, where n = 1000), we find that the standard error = .0027, and the reported value is within one standard error of the approximated expected value.

The derivations of the expected variance of the R index, in principle, should follow from the derivations of the expected variance of the C-V index but will not be attempted here because of the complexity of the derivations.

APPENDIX IV

B-S Statistics

Application of the B-S Decision Rules

Table IV.1
B-S Statistics

Objective #	Item #	TER[@1]	BER[@]	BDI	PER[@]	PDI	DER	N
M187S	1	.67H	.63H	.87**/.91	.0625L	1.0**/1.0	.04	16
	2		.69H	1.0**/1.0	.0625L	1.0**/1.0	-.02	
	3		.63H	.87**/.91	0.0 L	und/0.0	.04	
	4		.75H	.86**/.80	0.0 L	und/0.0	-.08	
R182S	1		.30L	.84**/.84	.10L	.67**/.5	.37**	30
	2		.17L	.68**/.56	.17L	.45*/.42	.50**	
	3		.20L	.40*/.35	.13L	.54**/.46	.47**	
	4		.10L	.27/.17	.13L	.78**/.67	.57**	
M167S	1		.71H	.57*/.8	.18L	.6*/.6	-.04	17
	2		.88H	1.0**/1.0	.24L	.83*/.93	-.21	
	3		.88H	1.0**/1.0	.18L	1.0**/1.0	-.21	
	4		.82H	.31/.37	.24L	.83**/.93	-.15	
R166S	1		.33L	.35/.37	0.0L	und/0.0	.34**	18
	2		.28L	1.0**/1.0	.22L	.84**/.93	.39**	
	3		.17L	.72**/.6	.06L	.82**/.78	.50**	
	4		.33L	.88**/.92	.17L	1.0**/1.0	.34**	
M198G	1		.27L	1.0**/1.0	.14L	1.0**/1.0	.40**	22
	2		.27L	1.0**/1.0	.09L	.80**/.67	.40**	
	3		.45L	.67**/.75	.14L	1.0**/1.0	.22*	
	4		.27L	1.0**/1.0	.14L	.61**/.61	.40**	
M176G	1		.67H	und/und	.52H	.22/.55	0.0	46
	2		1.0H	und/und	.74H	.36*/.77	-.33	
	3		1.0H	und/und	.96H	-.05/-.05	-.33	
	4		1.0H	und/und	.89H	.61**/.93	-.33	
R145G	1		.33L	.64**/.61	.24L	.68**/.62	.34**	66
	2		.39L	.62**/.61	.20L	.48**/.41	.28**	
	3		.5H	.73**/.73	.35L	.52**/.54	.17**	
	4		.35L	.54**/.52	.24L	.6**/.55	.32**	
R199G	1		.35L	.77**/.79	.49L	.68**/.68	.32**	57
	2		.21L	.76**/.67	.32L	.67**/.62	.46**	
	3		.16L	.64**/.5	.44L	.66**/.65	.51**	
	4		.42L	.64**/.68	.47L	.79**/.79	.25**	
R116G	1		.34L	.43**/.41	.30L	.47**/.48	.33**	59
	2		.53H	.44**/.45	.27L	.37**/.36	.14*	
	3		.15L	.42**/.30	.14L	.62**/.47	.52**	
	4		.29L	.69**/.64	.29L	.59**/.59	.38**	
R120S	1		.25L	.73**/.73	.25L	.47*/.42	.42**	20
	2		.25L	.2/.2	.25L	.36/.33	.42**	
	3		.25L	.73**/.73	.5H	.41/.42	.42**	
	4		.15L	.73**/.6	.3L	.58**/.58	.52**	
R142G	1		.06L	.27/.13	0.0L	und/0.0	.61**	31
	2		.52H	.81**/.81	0.0L	und/0.0	.15	
	3		.23L	.40*/.34	.03L	1.0**/1.0	.44**	
	4		.55H	.75**/.75	.06L	.70**/.97	.12	
R112S	1		.33L	.79**/.79	.10L	.45*/.45	.34**	21
	2		.19L	.69**/.57	.05L	.69**/.50	.48**	
	3		.52H	.67**/.71	.48L	.34/.58	.15	
	4		.14L	.58**/.43	.10L	1.0**/1.0	.53**	
M109S	1		.08L	.38*/.21	.05L	-.14/-.07	.59**	37
	2		.41L	.94**/.96	.3L	.87**/.93	.26**	
	3		.38L	1.0**/1.0	.19L	.85**/.78	.29**	
	4		.38L	1.0**/1.0	.3L	.87**/.93	.29**	
M108G	1		.54H	.48**/.54	.15L	.68**/.77	.13*	52
	2		.63H	.71**/.77	.25L	.63**/.85	.04	
	3		.81H	.8**/.76	.08L	.43**/.43	-.14	
	4		.69H	.82**/.85	.12L	.82**/.81	-.02	
M107S	1		.38L	1.0**/1.0	.36L	1.0**/1.0	.29**	42
	2		.38L	.9**/.9	.38L	.95**/.96	.29**	
	3		.40L	.9**/.9	.33L	.96**/.93	.27**	
	4		.38L	1.0**/1.0	.35L	.9**/.89	.29**	
M111G	1		.37L	.7**/.7	.09L	.73**/.57	.30**	43
	2		.40L	.95**/.96	.14L	.91**/.86	.27**	
	3		.40L	.95**/.96	.21L	.86**/.94	.27**	
	4		.40L	.95**/.96	.19L	.76**/.80	.27**	

[1]TER = .67 for all items.

[@]An error rate is designated High (H) if it equals or is greater than .50; otherwise, it is designated as Low (L).

*Significance at the .05 level.

**Significance at the .01 level.

For the discrimination indices, BDI and PDI, the first number is the phi-coefficient, the second number is the B index. The notation "und" implies that the computation was not possible because there was a zero in the denominator. Therefore, the phi-coefficient or the B index in these cases is undefined.

Table IV.2
Application of the B-S Decision Rules

Objective #	Item #	Rule										Result
		1	2	3	4	6	7	8	9	11	16	
R116G	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NR	NA	NA	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
R120S	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NR	NA	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	NA	NA	?	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
R142G	1	NA	NA	R	NA	NR	NA	NA	NA	NA	R	0
	2	NR	NA	NA	NA	NR	NA	NA	NA	NA	NR	1
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
R112S	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NR	NA	NA	NA	NR	NA	NA	NA	NA	NR	1
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M109S	1	NA	NA	R	NA	NR	NA	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M108G	1	NR	NA	NA	NA	NA	?	NA	NA	NA	R	0
	2	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	3	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	4	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
M107S	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M111G	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M187S	1	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	2	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	3	NR	NA	NA	NA	NR	NA	NA	NA	NA	NR	1
	4	NR	NA	NA	NA	NR	NA	NA	NA	NA	NR	1
R182S	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M167S	1	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	2	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	3	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
	4	NR	NA	NA	NA	NA	?	NA	NA	NA	NR	1
R166S	1	NA	NA	R	NA	NR	NA	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M198G	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
M176G	1	NR	NA	NA	?	NA	NA	NA	?	?	NR	1
	2	NR	NA	NA	?	NA	NA	NA	?	?	NR	1
	3	NR	NA	NA	?	NA	NA	NA	?	?	NR	1
	4	NR	NA	NA	?	NA	NA	NA	?	?	NR	1
R145G	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NR	NA	NA	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
R199G	1	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	2	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	3	NA	NA	R	NA	NA	?	NA	NA	NA	R	0
	4	NA	NA	R	NA	NA	?	NA	NA	NA	R	0

NR no revision.

R revision.

NA not applicable.

? data insufficient to make judgment.

APPENDIX V

Reliability Estimates of Tests

Table V.1
Reliability Estimates of Tests

Objective #	Pretest	Posttest	Retention		All
			Pretest	Posttest	
R111G	$\bar{X}$=2.7119 V=.93 KR_{20}=.22	$\bar{X}$=2.8644 V=1.03 KR_{20}=.58	$\bar{X}$=3.00 V=1.01 KR_{20}=.39	$\bar{X}$=2.8983 V=.97 KR_{20}=.41	$\bar{X}$=5.8983 V=3.21 KR_{20}=.62
R120S	$\bar{X}$=3.1 V=1.19 KR_{20}=.56	$\bar{X}$=2.4 V=1.44 KR_{20}=.46	$\bar{X}$=2.7 V=1.21 KR_{20}=.41	$\bar{X}$=2.55 V=1.1475 KR_{20}=.63	$\bar{X}$=5.25 V=3.3875 KR_{20}=.77
R142G	$\bar{X}$=2.6452 V=1.26 KR_{20}=.56	$\bar{X}$=3.9032 V=.09 KR_{20}=0.0	$\bar{X}$=3.9032 V=.15 KR_{20}=.53	$\bar{X}$=3.8387 V=.14 KR_{20}=0.0	$\bar{X}$=7.7419 V=.51 KR_{20}=.63
R112S	$\bar{X}$=2.8095 V=1.77 KR_{20}=.78	$\bar{X}$=3.2381 V=1.90 KR_{20}=.74	$\bar{X}$=3.2857 V=.78 KR_{20}=.51	$\bar{X}$=3.4762 V=.63 KR_{20}=.49	$\bar{X}$=6.7619 V=2.18 KR_{20}=.68
M109S	$\bar{X}$=2.7568 V=2.45 KR_{20}=.91	$\bar{X}$=3.973 V=.03 KR_{20}=0.0	$\bar{X}$=3.1622 V=1.38 KR_{20}=.73	$\bar{X}$=3.8378 V=.51 KR_{20}=.86	$\bar{X}$=7.00 V=2.27 KR_{20}=.74
M108G	$\bar{X}$=1.3462 V=2.13 KR_{20}=.79	$\bar{X}$=3.0192 V=1.71 KR_{20}=.76	$\bar{X}$=3.3654 V=1.08 KR_{20}=.69	$\bar{X}$=3.2115 V=1.04 KR_{20}=.54	$\bar{X}$=6.5769 V=3.55 KR_{20}=.78
M107S	$\bar{X}$=2.45 V=3.58 KR_{20}=.99	$\bar{X}$=3.55 V=1.15 KR_{20}=.87	$\bar{X}$=2.57 V=3.44 KR_{20}=.98	$\bar{X}$=2.45 V=3.49 KR_{20}=.98	$\bar{X}$=5.02 V=13.26 KR_{20}=.98
M111G	$\bar{X}$=2.44 V=3.18 KR_{20}=.94	$\bar{X}$=2.12 V=3.36 KR_{20}=.94	$\bar{X}$=3.37 V=1.58 KR_{20}=.90	$\bar{X}$=3.09 V=2.27 KR_{20}=.92	$\bar{X}$=6.47 V=6.81 KR_{20}=.94
M187S	$\bar{X}$=1.31 V=3.09 KR_{20}=.96	$\bar{X}$=3.0 V=0.0 KR_{20}=*	$\bar{X}$=3.88 V=.23 KR_{20}=.67	$\bar{X}$=3.44 V=.25 KR_{20}=0.0	$\bar{X}$=7.31 V=.34 KR_{20}=-.08
R182S	$\bar{X}$=3.23 V=.91 KR_{20}=.46	$\bar{X}$=3.77 V=.25 KR_{20}=.20	$\bar{X}$=3.47 V=.78 KR_{20}=.88	$\bar{X}$=3.80 V=.16 KR_{20}=-.17	$\bar{X}$=7.27 V=.86 KR_{20}=.28

Table V.1--Continued

Objective #	Pretest	Posttest	Retention Pretest	Retention Posttest	All
M167S	$\overline{X}$=.71 V=1.27 KR_{20}=.74	$\overline{X}$=3.35 V=1.99 KR_{20}=.97	$\overline{X}$=3.18 V=1.91 KR_{20}=.88	$\overline{X}$=3.47 V=1.66 KR_{20}=.97	$\overline{X}$=6.65 V=6.46 KR_{20}=.95
R166S	$\overline{X}$=2.89 V=1.99 KR_{20}=.81	$\overline{X}$=3.78 V=.40 KR_{20}=.66	$\overline{X}$=3.56 V=.80 KR_{20}=.73	$\overline{X}$=3.72 V=.65 KR_{20}=.82	$\overline{X}$=7.28 V=2.65 KR_{20}=.88
M189G	$\overline{X}$=2.73 V=2.93 KR_{20}=.94	$\overline{X}$=3.64 V=.32 KR_{20}=.08	$\overline{X}$=3.50 V=1.43 KR_{20}=.92	$\overline{X}$=3.14 V=1.30 KR_{20}=.91	$\overline{X}$=6.64 V=5.05 KR_{20}=.78
M176G	$\overline{X}$=.33 V=.22 KR_{20}=0.0	$\overline{X}$=.98 V=.76 KR_{20}=.26	$\overline{X}$=.89 V=.75 KR_{20}=.32	$\overline{X}$=1.17 V=1.71 KR_{20}=.80	$\overline{X}$=2.07 V=3.63 KR_{20}=.75
R145G	$\overline{X}$=2.42 V=1.85 KR_{20}=.66	$\overline{X}$=2.58 V=1.36 KR_{20}=.44	$\overline{X}$=2.97 V=1.33 KR_{20}=.58	$\overline{X}$=2.64 V=1.11 KR_{20}=.28	$\overline{X}$=5.61 V=4.36 KR_{20}=.76
R199G	$\overline{X}$=2.86 V=1.88 KR_{20}=.78	$\overline{X}$=3.33 V=1.20 KR_{20}=.74	$\overline{X}$=2.28 V=2.38 KR_{20}=.79	$\overline{X}$=3.12 V=1.65 KR_{20}=.78	$\overline{X}$=5.40 V=3.92 KR_{20}=.66

APPENDIX VI

Sample Tests and Objectives

NAME_____ DATE_____

SCHOOL_____ GRADE_____

Objective #116. Given a sentence with a word underlined that is either unfamiliar or familiar and used in a new way, the learner will write a definition using context clues to get the meaning.

Circle the correct answer in each.

1. Sam's team score 10 runs. The other team scored 5 runs. Sam's team score twice as many runs as the other team.

   TWICE means:

   a. twenty times as much

   b. two times as much

   c. four times as much

2. The rest of you may leave, but I would like John to stay.

   REST means:

   a. to sleep

   b. to be awake

   c. everyone else

3. The farmer waited until the ground was warm before he planted the seeds.

   GROUND means:

   a. to make something fine

   b. to make meal for bread

   c. to have a place to grow seeds

4. The ball broke the bridge of his nose.

   BRIDGE means:

   a. something to drive on to go over to the other side of the river

   b. a card game played by a group

   c. the bony part of the nose

NAME _____ DATE _____

SCHOOL _____ GRADE _____

Objective #116. Given a sentence with a word underlined that is either unfamiliar or familiar and used in a new way, the learner will write a definition using context clues to get the meaning.

Circle the correct answer in each.

1. The tire was found against the tree.

   TIRE means:

   a. you need some rest
   b. you don't want to read any longer
   c. you find it on a car

2. The corn was picked before the silk on the ear turned dark.

   EAR means:

   a. something to hear with
   b. what a piece of corn is called
   c. earrings are what girls sometimes wear on them

3. Please pass over the boards with care.

   PASS means:

   a. to move or walk with your feet
   b. to move with your hands
   c. to move in your car

4. We heard a soft foot step in the hall.

   STEP means:

   a. a ladder
   b. to go over
   c. to walk

#116 Answer Key

Pretest

1. B
2. C
3. C
4. C.

Post-test

1. C
2. B
3. A
4. C

Reading--Primary Level

1. Given oral directions, the learner will run, walk, march, tiptoe, hop, jum, slide, skate, and skip.

2. Given aural stimuli (variety of sounds), the learner will describe sounds as loud or soft, high or low, fast or slow, short or long, single or repeated.

3. Given 3 identical stimuli (shapes, pictures, designs or letters, etc.) and one clearly different, the learner will identify the one that is different.

4. Given a simple story read orally, the learner will demonstrate his understanding of the main idea by retelling, drawing, or describing main events and characters.

5. Given a picture and a sentence begun by the teacher, the learner will complete the sentence with a reasonable action or concept.

6. When directed to close his eyes and listen to three words spoken by the teacher, the learner reproduces the words in the sequence in which they are spoken.

7. Given a letter and a series of 3 letters the learner will mark the letter in the series which is the same as the initially given letter.

8. Given the 9 basic colors, the learner will state the name of the colors.

9. Given a specific environmental sound, the learner will associate the sound with its source or reproduce sound, based on child's knowledge of what specific sources are possible.

10. Given small individual pictures of children, animals and toys, the learner will categorize the pictures by sorting them into separate groups, and explain to the teacher the reason he grouped the pictures together.

11. Given pictures or objects, grouped categorically, the learner will supply categories, names or labels to adequately describe each group.

12. Given 4 to 6 pictures which illustrate a known story or activity, the child will arrange them in proper left to right sequence to follow the story or action.

13. Given an action picture, the learner will provide a reasonable and logical description of what might happen next or predict an event that led up to the situation depicted in the picture.

15. Given a circle, square, triangle and rectangle, the learner will contrast 2 shapes by describing different characteristics of the shapes (a square has 4 corners, a triangle has 3 corners, etc.)

16. Given a verbal direction only once describing body movement involving three actions, the learner will act as directed.

17. Given a story read orally, the learner will demonstrate his ability to identify the selection as fact or fantasy, (real or unreal).

18. Given 2 words orally, (i.e. blue and glue, or soap and soup, top and Tom, or blue and blue) child will distinguish when words are same or different.

19. Given a printed upper or lower case letter and a series of 3 letters, the learner will mark the letter in the series which is the same as the initially given letter but of the opposite case.

20. Given 4 sets of pictures with 3 similar pictures in each set and 4 additional pictures one for each set in a separate place, the learner will identify which should go in each row.

21. Given an oral selection and a series of 3 text-related pictures the learner will select the event which occurred first, next or last.

22. Given a simple story, orally, the learner will retell its event in sequence so that the story makes sense.

23. Given an action picture, the learner will respond to a question about the picture by using a complete sentence.

24. Given 3 overlapping figures, the learner will use different colored crayons to trace over figure.

25. Given scissors and prepared 8 x 11 paper (2 shapes drawn on paper), the learner will cut geometric and abstract shapes with no greater than a ¼ inch deviation.

26. Given a task of naming a group of objects arranged in 4 rows, 4 to a row, the learner will point to and name pictures in left to right progression beginning with row one, then 2, 3 and 4.

27. Given a group of 3 or 4 displayed objects which are viewed and then covered while one object is removed, the learner will identify missing objects when shown the changed group.

28. Given a printed letter and a printed series of three words, the learner will mark the word that begins with that letter.

29. Given directions to print his name, the learner will do so correctly.

30. Given oral directions, including words such as color, draw, circle, and underline, the learner will follow the directions.

31. Given two related words, read to him by the teacher and asked how they are alike or related, the learner will state the manner in which the two objects or concepts are related, accepting as correct any justifiable and adequate responses.

Lower Level

33. Given a letter name orally and a series of three letters in print, the learner will mark the letter name spoken.

34. Given a printed word and a printed series of three words, one of which has the same initial consonant as the first word, the learner will mark the word that begins with the consonant.

35. Given an oral story which expresses a mood the learner will mark from a choice of three pictures, the picture which identifies the mood in the story.

36. Given three words, orally, two of which have the same beginning consonant sound, the learner will name the two words which have the same beginning consonant sound.

37. Given a word orally and a series of three letters in print, the learner will mark the letter which represents the beginning sound of that word.

38. Given one word verbally and a series of three printed words, the learner will mark the word that has the same initial consonant as the verbally given word.

39. Given written directions including words such as color, make, draw, circle and underline, the learner will follow these directions.

40. Given a picture and a series of three words, the learner will mark the word that rhymes with the picture name.

41. Given a word orally and a series of three words orally and in print, one of which rhymes with the orally given word, the learner will mark the correct word.

42. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the first letter of that word (m, d, l, s, h), the learner will say a word that fits the context and begins with that letter sound.

43. Given an oral sentence with one word missing and cused for the missing word with a card having printed on it the first letter of the missing word (t, b, p, w, r, f, g, k, j, and n), the learner will say a word beginning with the letter that fits the context of the sentence.

44. Given a word orally, the learner will print the letter which is the initial consonant for that word.

45. Given a printed consonant, the learner will supply orally a word beginning with the consonant sound.

46. Given a set of the 8 basic colors names, i.e., red, blue, yellow, green, black, purple, brown and orange, the learner will indicate the corresponding color.

47. Given a sentence read orally, the learner will circle the period or question mark to indicate the punctuation to be used at the end of the sentence.

48. Given a list of Basic Dolch words for pre-primer level, the learner will read it.

49. Given a word orally, the learner will supply a word which has the same ending consonant sound as the spoken word.

50. Given pictures, the learner will write the final consonant for each picture.

51. Given a word orally, the learner will write the final consonant sound.

52. Given a series of nouns with at least one being plural, the learner will identify the plural nouns.

53. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the first letter of that word (v, y, or z), the learner will say a word beginning with the letter that fits the context of the sentence.

54. Given pictures of a one syllable long vowel word, and the word in print, without the vowel, the learner will supply the long vowel.

55. Given a list of Basic Dolch words for primer level, the learner will read it.

56. Given two lists of words, the learner will identify the words with the same graphemic base.

57. Given pictures of a one syllable short vowel word, and the word in print without the vowel, the learner will supply the proper short vowel.

58. Given a sentence with a word omitted and a series of three words, the learner will use the context of the sentence to mark the correct word.

59. Given a written sentence, the learner will identify material read as being real or fantasy.

60. Given a picture of a familiar activity and three sentences, the learner will select the sentence which best describes the picture.

61. Given a list of Basic Dolch words for first level, the learner will read it.

62. Given two words which could be changed to a contraction, and a series of three contractions, the learner will mark the correct contraction.

63. Given the beginning part of a word, a picture and a list of endings, the learner will match the beginning and ending parts to name the picture.

64. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the blend with which the word begins (br, cr, dr, fr, gr, tr, or pr), the learner says a word beginning with the letter blend that fits the context of the sentence.

65. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the blend with which the word begins (bl, cl, fl, gl, or sl), the learner says a word beginning with the letter blend that fits the context of the sentence.

66. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the blend with which the word begins (sk, sw, sm, sn, sp, or st), the learner says a word beginning with the letter blend that fits the context of the sentence.

67. Given an oral sentence with one word missing and cued for the missing word with a card with the sh or th digraph printed on it, the learner says a word beginning with the digraph that fits the context of the sentence.

68. Given a verb with an "ing" ending and a series of three verbs the learner will mark the base word.

69. Given an oral sentence with one word missing and cued for the missing word with a card having printed on it the digraph with which the word begins (ch, wh), the learner will say a word beginning with the digraph that fits the context of the sentence.

70. Given a list of words, the learner will write the words in alphabetical order by the first letter only.

71. Given a compound word, the learner will divide them into the two root words.

72. Given three printed sentences, the learner will select the two that have similar ideas.

73. Given four paragraphs and four sentences, the learner will select the sentence that implies what the paragraph says.

74. Given a list of Basic Dolch words for second level, the learner will read it.

75. Given a printed short story, the learner will identify the setting.

76. Given an oral sentence containing a word unknown in meaning and a direct definition clue to the word's meaning, the learner states the meaning of the unknown word.

77. Given a reading selection, the learner will arrange a series of randomly placed details into chronological order.

78. Given a reading selection, the learner will determine the main idea of the selection.

79. Given a short story read orally to him by the teacher, the learner provides details about the story which were implied but not stated.

80. Given the title of a possible story and a series of possible details, the learner selects the details which would be appropriate for the title.

81. Given a list of words, the learner will be able to categorize words as (contractions or compound words).

82. Given a sentence orally, the learner will determine if the sentence answers the questions, how, where, when, who or what.

83. Given a sample table of contents, the learner will demonstrate his ability to interpret a table of contents by selecting from a set of printed choices, the page number where certain information may be found.

84. Given a word orally, the learner will determine the number of syllables in the word.

85. Given a sentence in which an affix has been omitted from one of the words and a choice of three affixes, the learner will select the affix appropriate to the sentence.

86. Given two lists, the learner will match equivalent forms (contracted and uncontracted forms, possessives, and parphrases of same).

87. Given a list of words in which part of each word is underlined the learner will form two new words by substituting new letters for the underlined parts.

Middle Level

88. Given a list of words, the learner will classify words according to their structural endings (plural or singular, past or present tense).

89. Given a list of words, the learner will locate letters in different words that stand for the same sound (vowel or consonant) including multiple spelling variations.

90. Given a list of words the learner will add the given suffix, ing, ed, or s.

91. Given two lists of words the learner will match the antonyms.

92. Given a list of words, each beginning with the same letter, the learner will write the words in alphabetical order using the second letter.

93. Given a list of Basic Dolch words for third level, the learner will read it.

94. Given a written selection, the student will compose a title suitable to the material.

95. Given a reading selection, the learner will list characters included in the selection.

96. Given a review word, the learner will write a sentence that defines the word.

97. Given a paragraph, the learner will select from a list of three statements the one which most closely describes the main idea of the paragraph.

98. Given a table of contents, the learner will correctly find the chapter headings or titles and the page number.

99. Given sentences, the learner will identify which ones describe past time and which ones describe present time.

100. Given words and paraphrases of some forms, the learner will be able to match the equivalent forms.

101. Given a reading selection, the learner will identify aspects of literature by classifying the selection as humorous or not, with happy or sad endings.

102. Given sets containing three words, the learner will choose between words to fit context or match definitions or answering questions to indicate identification of words and the sounds they contain.

103. Given paris of synonyms, homonyms and antonyms, the learner will circle the pairs of synonyms.

104. Given a list of words and a list of definitions, the learner will match words with their most appropriate meaning.

105. Given a group of words containing a specific variety of suffixes (er, est, ly, ful, ness, y) the learner will find the suffix in each word.

106. Given a selection to read and a list of sentences, the learner will locate untrue statements about the selections.

107. Given a reading selection, the learner will identify elements of content by indicating objects to fit descriptions or answer riddles.

108. Given a topic, the learner will write a story of at least four sentences.

109. Given a poem to read, the learner will identify patterns of rhyme or repetition.

110. Given a short story, the learner will state the main idea.

111. Given a story beginning, the learner will write an ending in such a way that the relationship between the beginning and ending is logical.

112. Given a dictionary and a group of words the student will identify in which quarter of the dictionary each word is located.

113. Given a list of words, with the first two letters identical in each word, the learner will alphabetize them.

115. Given a new word, the learner will use a dictionary to locate its definition.

116. Given a sentence with a word underlined that is either unfamiliar or familiar and used in a new way, the learner will write a definition using context clues to get the meaning.

117. Given a list of topics and a table of contents, the learner will find the unit titles, chapter headings, sub-titles and page numbers related to the topic and record them.

119. Given a pictorial graph (e.g., weather, population, attendance, scores, etc.) the learner will interpret the data orally or in writing.

120. Given a sentence containing an underlined, incomplete root or base word and a list of prefixes, the learner will select that prefix from the list which completes the root word according to the sentence context (dix, in, de, com, en, sub, mis, re, un).

121. Given a list of known words the learner will add the following suffixes where appropriate: er, est, ly, ful, ness, y.

122. Given a reading selection, the learner will demonstrate ability to read with understanding by answering questions about details.

123. Given a list of words, the learner will be able to categorize structural components of words as affixed or possessives.

124. Given a reading selection, the learner will demonstrate his ability to read with understanding by answering questions about sequence of events.

125. Given a short story to read, the learner will demonstrate ability to read with understanding by answering questions about main ideas.

126. Given a sentence with a missing multi-meaning word, the learner will use the context of the sentence to supply the missing word.

127. Given a set of homonyms, synonyms and/or antonyms, the learner will define each set as directed.

128. Given a table of contents, the learner will locate specific information.

129. Given a reading selection, the learner will skim the selection to locate specific information.

130. Given a topic and several book titles, the learner will identify the one(s) whose contents would cover the topic.

131. Given a list of personal pronouns, the learner will write the correct possessive form and use each in a sentence.

132. Given information found within a telephone directory, the learner will tell where the information is found.

133. Given a map, the learner will utilize scales and symbols in answering questions about a given map.

134. Given a sentence with a verb (sit, set, lay, learn and teach) omitted, the learner will write in the correct tense of the verb located in parentheses next to the blank.

135. Given a paragraph describing a character in a particular situation, the learner will identify emotions he imagines were experienced by that character, consequent to the situation described.

136. Given incomplete sentences and a present tense verb for each, complete each sentence with the correct past tense form of the verb given.

137. Given a set of guide words, the learner will identify from a list, those words which would be found on a dictionary page having guide words.

138. Given a map, the learner will interpret information to answer questions.

139. Given an encyclopedia and a dictionary, the learner will identify two similarities and two differences between them.

140. Given the lists of reference sources, the learner will select the appropriate reference sources to obtain specific information.

141. Given a sample dictionary page, the learner will be able to discriminate between guide words, entry words, pronunciation key, and definitions.

142. Given a textbook and a list of words within its glossary, the learner will locate the glossary and list the definition it gives for each word.

143. Given a paragraph, the learner will locate the topic sentence.

Upper Level

145. Given a problem or question, the learner will identify the key words he would look up in an index to find information related to the problem.

146. Given an index of an encyclopedia, the learner will locate the volume and page number of a given topic, illustration or map.

147. Given derived words, the learner will be able to use a dictionary to locate the base words.

148. Given an article, the learner will outline, in topic form, its main points.

149. Given a list of words whose first three letters are exactly the same, the learner will be able to arrange the words alphabetically.

150. Given a scrambled set of words or sentences, the learner will arrange them into a logical order.

151. Given the names of the days of the week, months, etc., the learner will identify abbreviations of these given words, by matching.

152. Given a list of pronouns the learner will identify or write the possessive form of a given pronoun in a given situation.

153. Given a list of words, the learner will be able to demonstrate his knowledge of verb suffixes by matching each of the following suffixes to the appropriate root word in a given list: ize, fy (or - ify), -ate, en.

154. Given a list of words, the learner will be able to demonstrate his ability to divide words according to the rules of syllabication, by drawing a line between each syllable.

155. Given a reading selection, the learner will identify key words, phrases or passages important to the meaning of the selection.

156. Given sets of sentences, each containing the same word but with variations in its meaning, the learner will use the context of the sentences and the dictionary to identify the meaning of the word in each sentence.

158. Given a specific word or list of words, the learner will use a dictionary to find the syllables, parts of speech, meaning and synonyms for a given word.

159. Given a section of a dictionary and a list of words, the student will locate each word and identify what its grammatical abbreviation represents.

160. Given a list of unfamiliar words of three, four or five syllables, some of which have been extended by the addition of prefixes and suffixes, the learner will say each word and sound out the syllables.

161. Given a paragraph to read, the learner will be able to identify traits of a character, found in the paragraph.

162. Given a reading selection, the student will select the event which creates the major conflict or problem in the story.

163. Given four categories of mood and a list of words, the learner classifies each word according to the mood it fits.

164. Given several phrases, the student will be able to identify which of the five senses (sight, smell, tase, touch and hearing) each phrase appeals to.

165. Given an example for an author card in the card catalogue, the learner will find the title and call number of the book by the author.

166. Given an example of a title card, the learner will find its author and call number in the card catalogue.

167. Given an example of a subject card, the learner will locate in the card catalogue the title, author and call number of one or more books on that topic.

169. Given sentences, each an example of figurative language, and given possible interpretations of meaning of each, the learner will select the meaning.

170. Given a random group of factual and opinionated statements the student will classify each one according to those categories.

171. Given a list of words beginning with prefixes, the student will identify the prefix of each word and state the meaning of the prefix.

172. Given a selection of cause and effect relationships, the student will identify these relationships, by matching each cause statement with its corresponding effect statement.

173. Given a paragraph to read, the learner will identify the authors purpose in writing a selection, (e.g. entertainment, instruction, or persuasion).

176. Given a section heading from a textbook, the learner will briefly explain what that section might be about.

177. Given a list, the learner will identify the use of chapter overviews.

179. Given an index of a given book, the learner will find pages where information is found.

180. Given characteristics of Myths and tall tales, the student will identify them.

181. Given a reading selection the learner will choose from a series of sentences the one which best describes motive for some action or activity.

182. Given a paragraph to read, the learner will answer a related question whose answer is implied, but not directly stated within its content.

183. Given a selection whose content infers a moral or value, the student will interpret the content by writing an explanation of its meaning.

184. Given specific selections, the learner will be able to evaluate selections read as to type of literature, such as fable, legend, mystery, poem or biography.

185. Given a list of derived words that are not entry words, the learner will use the dictionary to locate the base (root) word, and define the derived word.

186. Given a root word, the learner will add a suffix or prefix, making appropriate spelling changes in the root when necessary to form new words as directed.

187. Given a printed selection and an opinion, the learner will skim to locate information which supports the opinion.

189. Given a form or application, the learner will correctly follow instructions to complete the form.

190. Given a set of sentences containing two omissions and a choice of two words, one possessive pronoun and one contraction, the learner will identify the correct word for ea ch omisssion (it --it's).

191. Given the comparative and superlative forms of adjectives, including the irregular forms of good, bad, many and little, the learner will write sentences, using the correct form.

192. Given a set of words denoting business or organizational terms, the learner will supply the abbreivations for each one.

193. Given statements from reading selections, the learner will identify examples of similies, metaphors and alliteration.

194. Given a list of words which have changed in meaning, the learner will identify both their original and their current meanings.

195. Given a paragraph and a list of generalizations about the paragraph, the learner will recognize those generalizations that are true.

196. Given a list of phrases, the learner will identify descriptive phrases as describing action, painting visual pictures and/or denoting sound.

197. Given a literary selection (essay, poem or biography) the learner will write five or more phrases that the author has used to describe how people feel and/or write his own interpretation of the selection.

198. Given the following parts of a book: title page, copyright, dedication, and table of contents, the learner will define them.

199. Given a new word in context and its etymology the learner will identify the word's meaning as used in the context.

200. Given a sentence expressing a definite mood, the learner will choose a word describing the mood and match another sentence expressing the same mood with the first sentence.

201. Given an article from an encyclopedia or a given selection, the learner will chart materials in outline form (I, A, B, C, D).

204. Given a particular situation, the learner will explain why a person or group of persons often give very different accounts of the same events.

207. Given a reading selection, the learner will identify propaganda techniques, such as persuasion, unstated assumptions, and emotionally charged statements.

208. Given a reading selection, the learner will state whether it is relatively biased or unbiased.

Math--Primary Level

1. Given a set of geometric shapes, the learner will name a circle, square, triangle and a rectange.

2. Given a pair of equivalent sets, with zero to five members, the learner will be able to achieve a one to one matching between members of the sets.

3. Given a set having one to five objects as members, the learner will pick out sets that have the same number of objects.

4. Given a set with less than ten objects, the learner will make an equivalent set by using actual objects.

5. Given a set of one to six objects, the learner will form another set that has exactly one more object and tell how many are in the new set.

5.5 Given a set of ten objects, the learner will count the objects in the set using the numeral name.

5.6 Given picture cards showing sets with one to nine members (one set per card), the learner will arrange cards in sequential order.

5.7 Given cards showing numberals one to nin, (one numeral per card), the learner will arrange cards in sequential order.

5.8 Given a numeral from one to nine, the learner will tell the name of the numeral.

5.9 Given a number line one to nine, the learner will tell the name of the numeral that comes just before or just after a given numeral.

6. Given two sequentially ordered sets of objects, one of which has one more than the other, the learner will form a third set that comes next in order.

6.5 Given a set of ten objects varying in attributes, the learner will sort the objects into two sets according to attributes (color, shape, size, texture).

7. Given two sets of objects from one to five, the learner will combine the sets and tell how many members are in the new set.

8. Given a set of two to six objects, the learner will separate into two subsets and will tell how many members are in each subset.

9. Given a set of one to six objects, the learner will take one away and tell how many are in the new set.

10. Given real or pictured sets, the learner will indicate which set identified the following quantitative descriptors: full, empty, greater, less, same, least, most.

11a. Given a penny, nickel, dime and quarter, the learner will give the name of each.

12. Given a set of objects, the learner will compare them, identify and name the heaviest and the lightest.

14. Given pictures showing fractional divisions, the learner will color or circle an area to demonstrate the concept of ½.

15. Given a calendar, the learner will demonstrate its use by identifying a day, week and month, when asked to do so.

17a. Given a thermometer, the learner will demonstrate its use by stating what it's used for in various situations.

18a. Given non-standard units of measure, the learner will use these units to measure objects in the classroom.

19a. Given a meter stick, the learner will demonstrate its use by showing how you would use it to measure.

20. Given real or pictured settings, using three-dimensional objects, the learner will identify other objects that are in the following position relationships to the given object. Above, below, under, on, in, top, bottom, in front of, between, in back of, inside, and outside.

21. Given the direction to rote count from 1-25, the learner will do so.

22. Given no model, the learner will write numerals 0-9 in correct form so that they are recognizable (reversals are acceptable).

23. Given the number words from 1-9, the learner will write corresponding numerals.

24. Given a row and a column of objects, not exceeding five, the learner will name the ordinal name of objects in a row and objects in a column.

24.5 Given a pattern using objects of two or more colors, the learner will duplicate the pattern selecting from a set of similar objects (red-blue-red-blue-red-blue).

24.6 Given a pattern using objects of two or more shapes, the learner will duplicate the pattern, selecting from a set of similar objects (X-O-X-O-X-O).

25. Given a two part pattern, the learner will continue the pattern.

26a. Given the direction to recite chronologically the days of the week, the learner will do so.

Lower Level

27a. Given a blank clock face, the learner will writer numerals to 12 in correct place on the clock face.

28. Given a number line, the learner will demonstrate the relationship between adding numbers and adding objects, by counting steps on the number line.

29. Given pictures of six sets, some of which are empty, the learner will identify the empty sets.

30. Given any addition combination, in mixed horizontal or vertical form, (0+0, to 5+5) the learner will write the sum with the aid of a manipulative device.

31. Given a set of nine or fewer small objects, the learner will separate the given set into at least two pairs of subsets, then write the number for each subset.

32. Given a number line, the learner will demonstrate the relationship between subtracting numbers and subtracting objects by counting steps on the number line.

33. Given two written numerals less than ten, the learner will indicate which is greater or lesser in value.

34. Given a set of pictured objects, not exceeding 25, the learner will count and write in numeral form, the number of objects.

35. Given any addition combination in mixed horizontal or vertical form (0+0 to 0+0) the learner will write the sum with the aid of a manipulative device.

36a. Given a marked clock face, the learner will identify the time to the hour.

37a. Given a meter stick or a 20 cm ruler, the learner will measure objects or lines to the nearest meter or centimeter.

39a. Given the direction to recite chronologically the months of the year, the learner will do so.

40a. Given pictures of a penny, nickel, dime or quarter, the learner will identify numerical values of each.

42. Given addition problems in both horizontal and vertical forms, through sums of 9, the learner will find the missing addends with aid of a manipulative device.

43. Given two consecutive even or odd numerals, 0-25, the learner will write the numeral that comes between the two given numerals.

44. Given a numeral, 1-25, the learner will write the numerals that come before and after the given numeral.

45. Given up to 90 pictorial objects, (the number must be a multiple of 10) the learner will form sets of 10 and group and label sets of tens, as two tens, ... 9 tens.

46a. Given a set of no more than 90 objects groups by tens, the learner will say and write the numeral.

47. Given subtraction problems with minuends less than 19, subtrahends less than ten, written in both horizontal and vertical form, the learner will find the differences.

48. Given any missing addends sentence, with sums up to 18, the learner will be able to give the missing addend.

49. Given a mathematical statement, the learner will be able to place the correct sign > , < , or = between two numerals in the range 1-50.

50. Given a number sentence with the operation sign (+ or -) missing, the learner will complete them by writing in the correct sign.

51. Given a sequence of numerals, involving skip counting, by two's up to 30, the learner will write the missing numeral.

52. Given a sequence of numerals, involving skip counting, by five's and ten's up to 50, the learner will write the missing numeral.

53. Given an oral numeral, not to exceed 50, the learner will be able to write it.

54. Given an oral word problem requiring addition, with sums less than 18, the learner will find the sum.

55a. Given a marked clock face, the learner will state time to nearest indicated ½ hour.

56. Given pictures of a circle, square or triangle, the learner will identify the shaded portion that corresponds to ½, or ¼.

57. Given three, one-digit numerals, with the sum less than 21, the learner will find the sum.

58. Given number phrases less than 20, the learner will supply the appropriate symbol of equality or inequality, > , = , or < .

59. Given two two digit numerals, requiring no regrouping, the learner will find the sum.

61. Given a problem or the form (two digit - one digit with no regrouping), the learner will find the difference).

62a. Given hours, minutes and days, the learner will indicate the correct relationship between them.

63a. Given line segments and a ruler, the learner will measure the line segments to the nearest half centimeter.

64a. Given a 20 centimeter ruler, the learner will construct a line segment of specified length, designated to the nearest half centimeter.

65. Given an addition problem, of the form 21 + 34 = 34 + __, the learner will give the missing addend.

66. Given any one, two or three digit numeral, the learner will write it in expanded notation.

67. Given two two digit numerals, requiring regrouping, the learner will find the sum.

68. Given a problem of the form, two digit minus one digit, the learner will find the difference, regrouping if necessary.

69a. Given pictures of money or play money, less than or equal to $1.00, the learner will compare the values between coins.

70a. Given pictures of money or play money, less than or equal to $20.00, the learner will write the given money value using the symbols of dollar sign and decimal.

71a. Given a clock face with hands, the learner will write time in time notation to half hour and quarter hour.

72a. Given cup, pint, quart and liter containers, the learner will determine experimentally, the number of cups in a pint, pints in a quart, approximate quart in a liter.

73. Given oral word problems involving addition and subtraction with numbers less than 18, the learner will solve them.

73.5 Given subtraction problems in both horizontal and vertical forms, with minuends not to exceed 18, the learner without regrouping, will find the missing subtrahend.

74. Given a problem of the form (three digit minus one, two or three digit), the learner will find the difference when no regrouping is required.

75. Given a problem of the form (two digit minus two digit), the learner will find the difference, regrouping if necessary.

76. Given a story problem read orally by the teacher, the learner will tell which he must use to solve the problem (addition or subtraction).

77. Given an oral word problem requiring subtraction, requiring regrouping the learner will state and do what operation is necessary to find the difference.

79a. Given several objects divided into (¼'s, 1/3's, ½'s or whole) by comparing to the whole unit.

80. Given number sequences in which some of the numbers are omitted, the learner will complete the number sequences up to 200.

81. Given two, three digit numerals, the learner will apply the appropriate symbol between them (> , < , =).

82a. Given play money, the learner will make change from $1.00 for any amount up to $.99.

83a. Given drawings of lines, the learner will point out which ones are (relatively) horizontal and which are (relatively) vertical.

84. Given two three digit numbers, the learner will find the sum, regrouping, if necessary.

85. Given a three digit minuend and a two or three digit subtrahend, the learner will find the difference, regrouping if necessary.

86. Given a number and the consecutive multiples of ten or 100 between which it falls, the learner will choose the nearer estimate.

87. Given column addition exercises involving three two digit addends, the learner will find the sum, regrouping if necessary.

88. Given a pair of numbers or number phrases less than 1,000, the learner will supply the appropriate symbol > , <, or =.

89. Given two addends less than 10,000, the learner will find the sum, regrouping if necessary.

90a. Given a clock face with hands, the learner will read the time to the nearest minute.

91a. Given a length expressed in centimeters, the learner will express it as a number of centimeters plus a number of millimeters.

Middle Level

92a. Given an object or line segment, the learner will, without the use of a ruler, choose the correct estimate from a set of answers of this form.

2 millimeters, 2 centimeters
2 meters, 2 decimenters

93a. Given the terms, centimeter, meter and decimeter, the learner will state the relationship between them.

94. Given a repeated addition sentence, the learner will represent it as a multiplication sentence with its product.

95. Given multiplication problems using one as a factor, the learner will find the product.

96. Given any multiplication combinations, less than 5 x 5, the learner will write the product.

97. Given multiplication problems using zero as a factor, the learner will find the products.

98. Given sets of not more than 20 elements, the learner will divide them into equivalent sub-sets.

99. Given a mathematical sentence of the form (3 x 4 = 4 x __), the learner will identify the missing factor.

100. Given basic multiplication problems, the learner will find the products, using the distributive property of multiplication over addition.

101. Given any multiplication combination up to 9 x 9, the learner will write the product.

102. Given multiplication problems in which the factors are whole numbers less than ten, and one factor is missing, the learner will record the missing factor.

103. Given the basic division facts through the nines, the learner will find the quotients.

104. Given a multiplication number sentence with two missing factors, the learner will supply any two basic factors to make the given multiplication number sentence true (ex: __ x __ = 16).

105. Given any number as the dividend, and zero as the divisor, the learner will indicate that there is no solution to the problem.

106. Given a word problem requiring multiplication, the learner will write the correct equation to go with the problem.

107. Given a set of multiplication equations in which one factor is a multiple of 10,000 or 1,000, the learner will write related division equations.

108. Given a multiplication problem of the form (one digit number x two digit number) the learner will find the product.

109a. Given a shaded region located on a piece of graph paper or some other grid, the learner will find the area by counting the number of square units.

110a. Given pictures or models of geometric figures; cube, cylinder, sphere, the learner will identify them.

111. Given two factors which are multiples of ten, the learner will find the product.

112a. Given a sentence involving the terms "in the morning, in the afternoon, in the evening," the learner will supply the appropriate AM or PM notation.

113a. Given two times to the nearest half hour, the learner will find the length of the interval between them.

114. Given two or three whole number addends less than 100,000 in horizontal or vertical form, the learner will find the sum, regrouping if necessary.

115. Given subtraction problems with up to four digit minuends and subtrahends, the learner will find the differences, regrouping if necessary.

116a. Given Arabic numerals 1 through 39, the learner will convert them to Roman numerals.

117a. Given Roman numerals I through XXXIX, the learner will rewrite them to Arabic numerals.

118. Given a numeral with up to four digits, the learner will rewrite the given numerals, using expanded notation.

119. Given a completed division problem, the learner will identify the divisor, dividend, quotient and remainder.

120. Given a series of four numbers, the learner will compute the average.

121. Given a multiplication problem with two two digit factors, the learner will find the product.

122. Given a division problem, with a two digit divident and a one digit divisor, the learner will determine the quotient, with or without a remainder.

123. Given multiplication problems involving a multiple of ten times a multiple of 100, the learner will find the products.

124. Given a multiplication problem with a two digit factor and a three digit factor, the learner will find the product.

125a. Given the length (whole numbers less than 10), of the sides of a rectangular region, the learner will find the area.

126a. Given a line segment to measure and a 20 cm ruler with millimeter markings, the learner will express its measure in whole centimeters or millimeters.

127a. Given a sequence of metric pre-fixes, the learner will arrange them in order from smallest to largest.

129. Given a fraction orally, the learner will write the fraction.

130. Given a proper fraction, the learner will identify the numerator and the denominator of the fraction.

131. Given a denominator, the learner will supply the correct numerator to make the value of the fraction equal to one, without the use of aids.

132. Given a proper fraction with a denominator less than nine, the learner will explain the meaning of each fraction by making a drawing or by using fractional cut-outs.

133. Given a simple fraction, the learner will give at least two equivalent fractions.

134. Given fractions with like denominators, the learner will add to the sum of less than one.

135. Given any five fractions with like denominators, in random order, the learner will write them in numerical order.

136. Given a division problem with a three digit divident and a one digit divisor, the learner will determine the wuotient, with or without a remainder.

137. Given two factors of up to three digits each, the learner will estimate the product by rounding both factors to the nearest ten and multiplying.

138. Given the decimal fraction of no more than three places, the learner will name the place value of the digit.

139. Given an addition or subtraction of decimal problem in vertical form with no more than five digits and no more than three decimal places, with each problem having the same number of decimal places, the learner will find the sum or difference and correctly place the decimal point.

140. Given an expressed amount of money, the learner will multiple or divide the given amount by a whole number.

141. Given numerals between ten and 5,000, the learner will round off numerals to the nearest 10's, 100's, or 1,000's place.

142a. Given any numeral from 1,000 to 9,999,999, the learner will locate and separate the periods with commas.

143. Given a story problem, with whole numbers and requiring only one operation (addition, subtraction, multiplication or simple division), the learner will choose the correct operation and do the computation.

144a. Given a six digit numeral in oral form, the learner will write the given six digit numeral.

145a. Given two times to the nearest minute, the learner will find the time interval.

146. Given any four digit number, the learner will give the number that is 100 or 1,000 less than it is without using formal addition or subtraction.

147. Given an exercise in multiplication, the learner will multiply a three or four digit factor by a two or three digit factor.

148. Given a division problem with a four digit divident and a one digit divisor, the learner will determine the quotient with or without a remainder.

149. Given division problems with multiples of 100 as dividends and two digit divisors, the learner will estimate the quotient by rounding off the divisors to the nearest ten.

150a. Given pictures or models of prisms, cones and pyramids, the learner will correctly identify them.

151. Given a numeral expressed as a power with an exponent less than five, the learner will express it as an ordinary base ten numeral.

152. Given an addition or subtraction decimal problem in horizontal or vertical form, with no more than three decimal places, the learner will find the sum.

153. Given a number with no more than three decimal places, the learner will round to the nearest whole number, tenth or hundredth as requested.

154. Given a number less than 100, the learner will identify the factors of the given number.

155. Given a number, the learner will identify multiples of the given number.

157. Given division problems with two digit dividends and two digit multiples of ten as divisors, the learner will find the quotients, with or without a remainder.

Upper Level

158. Given division problems with three digit dividends and two digit multiples of ten as divisors, the learner will find the quotients, with or without a remainder.

159a. Given word problems, requiring division, the learner will give the equation and find the quotient.

160. Given the measurement of each side of a polygon, the learner will find the perimeter of the given polygon.

161a. Given a list of familiar objects, the learner will choose the volume measure (cu, cm., cu, dm., cu.m.) that would be nearest in size.

162. Given a division problem with a two digit divisor, a four digit dividend, with or without a remainder, the learner will find the quotient.

163. Given a story problem, with whole numbers and requiring only one operation, (addition, subtraction, multiplication or division) the learner will choose the correct operation and do the computation.

164. Given a list of numbers, the learner will identify prime numbers less than 100 (by circling them).

165. Given a pair of numbers, each less than 60, the learner will identify their greatest common factor.

166. Given a fraction, the learner will reduce it to its simplest form.

167. Given a number line segment (0, 1) with dots indicating division of the segment into equal segments, the learner will identify the fraction correspondong to a particular dot.

168. Given an improper fraction, the learner will write it as a mixed number.

169. Given a mixed number, the learner will write it as an improper fraction.

170. Given a whole number and a mixed number, the learner will find their sum.

171. Given a whole number and a mixed number, the learner will find the difference.

172. Given a decimal fraction, the learner will rename it as a common fraction.

173. Given a common fraction whose decimal equivalent terminates in two places or less, the learner will write its decimal equivalent.

174. Given a whole number and a fraction less than one, the learner will multiply to find the product.

175. Given a multiplication problem with fractions less than one as factors, the learner will find the product in simplest form.

176. Given a fractional number and a mixed number, the learner will find the product.

177. Given multiplication problems having two mixed numerals, the learner will find the product.

178. Given a common fraction whose decimal equivalent terminates in three places or less, the learner will rename the common fraction as a decimal fraction $\frac{1}{2} = 5/10 = .5$.

179. Given any six digit numeral, the learner will rewrite it with expanded notation, first by using place value words, and then by using numerals.

180. Given a pair of numbers, each less than 20, the learner will identify their least common multiple.

181. Given two fractional numbers, that may or may not require renaming, the learner will find their sum.

182. Given two mixed numbers that may or may not require renaming, the learning will find their sum.

183. Given subtraction problems involving mixed numerals, the learner will subtract mixed numerals with renaming and find the difference in simplest form.

184. Given two unequal fractions, with denominators of 2, 3, 4, 6 or 8, the learner will tell which is greatest in value.

185. Given a measurement involving two units in the same system, the learner will multiply the measurement by a whole number and regroup as necessary.

186. Given a division problem with a dividend of no more than five digits and divisor with no more than three places, the learner will find the quotient, with or without remainder. The remainders will be written as fractions in simplest form.

187. Given a list of fractional numbers, the learner will write the reciprocal of a number.

187.5 Given two fractional numbers, less than one, the learner will find the quotient.

188. Given a whole number divisor and a fraction, the learner will find the quotient.

189. Given a whole number dividend and a fractional divisor, the learner will find the quotient.

191. Given a fraction and a mixed number, the learner will find the quotient.

192. Given two mixed numbers, the learner will find the quotient.

193. Given a numeral from .001 through hundred millions, the learner will read and identify numerals, expanded numerals, or in word form.

194. Given a list of numerals from .999 to 1,000,000, the learner will round off each numeral to the place value indicated in the heading.

195. Given a multiplication of decimal problems, with no more than five digits and no more than three decimal places, the learner will find the product.

196. Given a decimal division problem in which the divisor and dividend have no more than five digits and no more than three decimal places, the learner will find the quotient.

197a. Given a set of equations, the learner will label, identify or compute as indicat-d, using the properties of addition and multiplication (dist., assoc., 1's and 0's).

198a. Given a measurement such as 1.463 meters, the learner will express it as one meter + four decimeters + six centimeters + three millimeters.

199a. Given diagrams or models of points, lines and planes, the learner will associate each diagram with one of the words: point, line, plane.

200a. Given drawings of parallel lines and perpendicular lines, the learner will associate each diagram with the correct words (parallel, perpendicular).

201a. Given a circle and its related parts, the learner will identify the center, radius, diameter and circumference.

202a. Given diagrams of segments, lines, rays and angles, the learner will select and name each as requested.

203a. Given a set of pictured angles, the learner will select those which are right angles.

204a. Given the formula for finding the area of a triangle and the measures of the base and height of the triangle, the learner will find the area.

205a. Given an English or a metric table of equivalent measurements, the learner will convert from one to another within the same system.

206a. Given a circle with its radius or diameter, the learner will find its circumference.

207a. Given the formula for finding the area of a circle and the measurement of the radius or diameter of a circle, the learner will find its area.

208a. Given a protractor, the learner will read the measure of any given angle from 0° to 180°--within two degrees.

209a. Given a drawing of a rectangular solid with its dimension (small whole numbers), the learner will compute the volume.

210a. Given a coordinates, the learner will locate the points on the grid.

211a. Given three pairs of coordinates and a grid, the learner will construct a line graph.

212. Given a square subdivided into an area of 10 x 10 unit squares, some of which are shaded, the learner will state the indicated ratio and percent represented by the shaded area.

213. Given a list of ratios, the learner will express an equivalent ratio of the given ratios.

214. Given a list of one or two digit decimal numerals less than one, the learner will express them as percents.

215. Given a ratio and the numerator or denominator of an equivalent ratio, the learner will write the missing numerator or denominator of the equivalent ratio.

216. Given a set of percents and two digit decimal numerals, the learner will write them as fractions in simplified form.

217. Given a set of proportion problems, where the given terms and the answer are each whole numbers less than 100, the learner will find the solution.

218. Given a percent problem, the learner will write the appropriate proportion needed to solve the problem in the form,

$$\frac{2}{8} = \frac{n}{100}, \frac{n}{8} = \frac{25}{100}, \text{ or } \frac{2}{n} = \frac{25}{100}$$

219. Given a set of problems, involving the three types of percent, the learner will write the appropriate proportion and solve the problem.

APPENDIX VII

Computer Program for the Simulation

```
      DIMENSION A(1000),B(1000),ND(31),NS(31)
      ISEED=4913
  110 READ 100,Q11,Q12,Q11PR,Q12PR,PI1,PI2,PI3
  100 FORMAT(7F6.0)
      Q21=1.-Q11
      Q22=1.-Q12
      Q21PR=1.-Q11PR
      Q22PR=1.-Q12PR
      PI4=1.-PI1-PI2-PI3
      P1=Q11*Q11PR*PI1+Q11*Q12PR*PI2+Q12*Q11PR*PI3+Q12*Q12PR*PI4
      P2=Q11*Q21PR*PI1+Q11*Q22PR*PI2+Q12*Q21PR*PI3+Q12*Q22PR*PI4
      P3=Q21*Q11PR*PI1+Q21*Q12PR*PI2+Q22*Q11PR*PI3+Q22*Q12PR*PI4
      P4=Q21*Q21PR*PI1+Q21*Q22PR*PI2+Q22*Q21PR*PI3+Q22*Q22PR*PI4
      STRUE=PI2/(PI1+PI2)
      DTRUE=PI2-PI3
      SP2=P1+P2
      SP3=SP2+P3
      READ 120,NIT,NOS
  120 FORMAT(I4,1X,I5)
C NIT EQUALS NUMBER IN SAMPLE,NOS EQUALS NUMBER OF SAMPLES
      PRINT 190,Q11,Q21,Q12,Q22,Q11PR,Q21PR,Q12PR,Q22PR,PI1,PI2,PI3,PI4,
     1P1,P2,P3,P4,STRUE,DTRUE,NIT,NOS
  190 FORMAT(1X,4(F4.2,2X),4(F4.2,2X),4(F3.2,1X),4(F5.4,1X),2(F5.4,3X),
     12(I4,1X))
      SUMS=0.0
      SUMD=0.0
      DO 140 J=1,NOS
      NF1=0
      NF2=0
      NF3=0
      NF4=0
      SEST=0.0
      DPEST=0.0
      DO 130 I=1,NIT
      CALL GGU3(ISEED,1,Y)
      IF(Y.LT.P1) GO TO 70
      IF(Y.LT.SP2) GO TO 80
      IF(Y.LT.SP3) GO TO 90
      NF4=NF4+1
      GO TO 130
   70 NF1=NF1+1
      GO TO 130
   80 NF2=NF2+1
      GO TO 130
   90 NF3=NF3+1
  130 CONTINUE
      SEST=FLOAT(NF2-NF3)/FLOAT(NF1+NF2)
      DPEST=FLOAT(NF2-NF3)/FLOAT(NIT)
```

```
  170   A(J)=SEST
        B(J)=DPPEST
        SUMS=SUMS+SEST
        SUMD=SUMD+DPPEST
  140   CONTINUE
        BIGS=A(1)
        BIGD=B(1)
        SMALLS=A(1)
        SMALLD=B(1)
        DO 200 JJ=2,NOS
        IF(B(JJ).LT.SMALLD)SMALLD=B(JJ)
        IF(A(JJ).LT.SMALLS)SMALLS=A(JJ)
        IF(B(JJ).GT.BIGD)BIGD=B(JJ)
        IF(A(JJ).GT.BIGS)BIGS=A(JJ)
  200   CONTINUE
        DO 803 KL=1,31
        ND(KL)=0
  803   NS(KL)=0
        DO 375 KK=1,NOS
        IHCP=-10
        IHCPY=-9
        HCP=FLOAT(IHCP)/20.
        HCPY=FLOAT(IHCPY)/20.
        DO 301 KL=1,31
        IF(A(KK).LT.HCP.OR.A(KK).GE.HCPY)GO TO 351
  350   NS(KL)=NS(KL)+1
        GO TO 300
  351   HCP=HCPY
        HCPY=FLOAT(IHCPY+1)/20.
  301   IHCPY=IHCPY+1
  300   IHCP=-10
        IHCPY=-9
        HCP=FLOAT(IHCP)/20.
        HCPY=FLOAT(IHCPY)/20.
        DO 360 KL=1,31
        IF(B(KK).LT.HCP.OR.B(KK).GE.HCPY)GO TO 353
  352   ND(KL)=ND(KL)+1
        GO TO 375
  353   HCP=HCPY
        HCPY=FLOAT(IHCPY+1)/20.
  360   IHCPY=IHCPY+1
  375   CONTINUE
        SMEAN=SUMS/FLOAT(NOS)
        DMEAN=SUMD/FLOAT(NOS)
        SUMSSQ=0.0
        SUMDSQ=0.0
        DO 250 JK=1,NOS
        SQA=A(JK)*A(JK)
```

```
      SUMSSQ=SUMSSQ+SQA
      SQB=B(JK)*B(JK)
      SUMDSQ=SUMDSQ+SQB
  250 CONTINUE
      VARS=(FLOAT(NOS)*SUMSSQ-(SUMS*SUMS))/FLOAT(NOS*NOS)
      VARD=(FLOAT(NOS)*SUMDSQ-(SUMD*SUMD))/FLOAT(NOS*NOS)
      SDS=SQRT(VARS)
      SDD=SQRT(VARD)
      SM3=0.0
      SM4=0.0
      DM3=0.0
      DM4=0.0
      DO 400 JJ=1,NOS
      STEMP=A(JJ)-SMEAN
      STEMP1=STEMP*STEMP*STEMP
      SM3=SM3+STEMP1
      SM4=SM4+STEMP1*STEMP
      DTEMP=B(JJ)-DMEAN
      DTEMP1=DTEMP*DTEMP*DTEMP
      DM3=DM3+DTEMP1
      DM4=DM4+DTEMP1*DTEMP
  400 CONTINUE
      SKS=SM3/(VARS*SDS*FLOAT(NOS))
      SKD=DM3/(VARD*SDD*FLOAT(NOS))
      RKUS=SM4/(VARS*VARS*FLOAT(NOS))-3.0
      RKUD=DM4/(VARD*VARD*FLOAT(NOS))-3.0
      PRINT 410,SKS,SKD,RKUS,RKUD
  410 FORMAT(' SKS=',F8.4,'SKD=',F8.4,'KUS=',F8.4,
     1' KUD=',F8.4)
      PRINT 510,NS
  510 FORMAT(1X,3II4)
      PRINT 530,ND
  530 FORMAT(1X,3II4)
      PRINT 270,BIGS,SMALLS,SUMS,SUMSSQ,SMEAN,VARS,BIGD,SMALLD,SUMD,
     1SUMDSQ,DMEAN,VARD,SDS,SDD
  270 FORMAT(1X,F6.4,1X,F7.4,2X,2(F9.3,1X),F6.4,2X,2(F6.4,1X),F6.4,2X,
     12(F9.3,1X),F6.4,2X,F6.4,1X,F6.4,1X,F6.4)
      READ 280,NSTOP
  280 FORMAT(I1)
      IF(NSTOP.EQ.0) GO TO 290
      GO TO 110
  290 CONTINUE
      END
```

BIBLIOGRAPHY

BIBLIOGRAPHY

Bernknopf, Stanley & Bashaw, W.L. An investigation of criterion-referenced tests under different conditions of sample variability and item homogeneity. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Brennan, R.L. A generalized upper-lower item discrimination index. Educational & psychological measurement, 1972, 32, 289-303.

Brennan, R.L. & Stolurow, Lawrence M. An elementary decision process for the formative evaluation of an instructional system. A paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Cox, R.C. & Vargas, Julie S. A comparison of item selection techniques for norm-referenced & criterion-referenced tests. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Crehan, Kevin D. Item analysis for teacher-made mastery tests. Journal of educational measurement, winter 1974, 11, no. 4, 255-262.

Davis, Frederick B. and Diamond, James J. The preparation of criterion-referenced tests. Problems in criterion-referenced measurement, CSE monograph series in evaluation, #3. Edited by Chester W. Harris, Marvin C. Alkin & James W. Popham, 1974.

Ebel, Robert L. Evaluation & educational objectives. Journal of educational measurement, winter 1973, 10, no. 4, 273-279.

Edmonston, Leon P.; Randall, Robert S.; & Oakland, Thomas D. A model for estimating the reliability and validity of criterion-referenced measures. A paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Haladyna, T.M. Effects of different samples on item & test characteristics of criterion-referenced tests. Journal of educational measurement, summer 1974, 11, no. 2, 93-99.

_____. The paradox of criterion-referenced measurement. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Haladyna, T.M. and Roid, G.H. The quality of domain-referenced test items. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Hambleton, Ronald K. & Gorth, William P. Criterion-referenced testing: issues & applications. Technical report no. 13; University of Massachusetts, September 1971; ERIC ED 060 025.

Hambleton, Ronald K.; Swaminathan, Hariharan; Algina, James; & Coulson, Douglas. Criterion-referenced testing & measurement: A review of technical issues & developments. Symposium presented at the annual meeting of the American Educational Research Association, April 1975, Washington, D.C.

Harris, Chester W. Some technical characteristics of mastery tests. Problems in criterion-referenced measurement, CSE monograph series in evaluation, #3. Edited by Chester W. Harris, Marvin C. Alkin & James W. Popham, 1974.

Helmstadter, G.C. A comparison of bayesian & traditional indexes of test item effectiveness. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Hsu, Tse-Chi. Empirical data on C-R tests. A paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Ivens, Stephen H. A pragmatic approach to criterion-referenced measures. A paper presented at the annual meeting of the American Educational Research Association & National Council on Measurement in Education, Chicago, April 1972.

_____. An investigation of item analysis, reliability & validity in relation to criterion-referenced tests. Unpublished doctoral dissertation. The Florida State University, 1970.

Kifer, Edward & Bramble, William. The calibration of a criterion-referenced test. A paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Kosecoff, Jacqueline B. & Klein, Stephen P. Instructional sensitivity statistics appropriate for objectives-based test items. CSE report no. 91, Center for the Study of Evaluation, UCLA, April 1974.

Livingston, Samuel A. Criterion-referenced applications of classical test theory. Journal of educational measurement, summer 1972, 9, no. 1, 13-26.

Marks, Edmund & Noll, Gary A. Procedures & criteria for evaluating reading & listening comprehension tests. Educational & psychological measurement, 1967, 27, 335-348.

Nitko, Anthony J. A model for criterion-referenced tests based on use. A paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Oakland, Thomas. An evaluation of available models for estimating the reliability & validity of criterion-referenced measures. A paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Ozenne, Dan Gilbert. Toward an evaluative methodology for criterion-referenced measures: test sensitivity. CSE report no. 72, Center for the Study of Evaluation, UCLA, October 1971.

Popham, James W. & Husek, T.R. Implications of criterion-referenced measurement. Journal of educational measurement, spring 1969, 6, no. 1, 1-9.

Roudabush, G.E. Item selection for criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Saupe, J.L. Selecting items to measure change. Journal of educational measurement, fall 1966, 3, no. 3, 223-228.

Schooley, Daniel E.; Schultz, Daniel W.; Donovan, David L.; & Lehmann, Irvin J. Quality control for evaluation systems based on objective-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Skager, Rodney W. Generating criterion-referenced tests from objective-based assessment systems: unsolved problems in test development, assembly & interpretation. Problems in criterion-referenced measure, CSE monograph series in evaluation, #3. Edited by Chester W. Harris, Marvin C. Alkin & James W. Popham, 1974.

Woodson, M.I. Chas. E. The issue of item & test variance for criterion-referenced tests. Journal of educational measurement, spring 1974, 11, no. 1, 63-64.